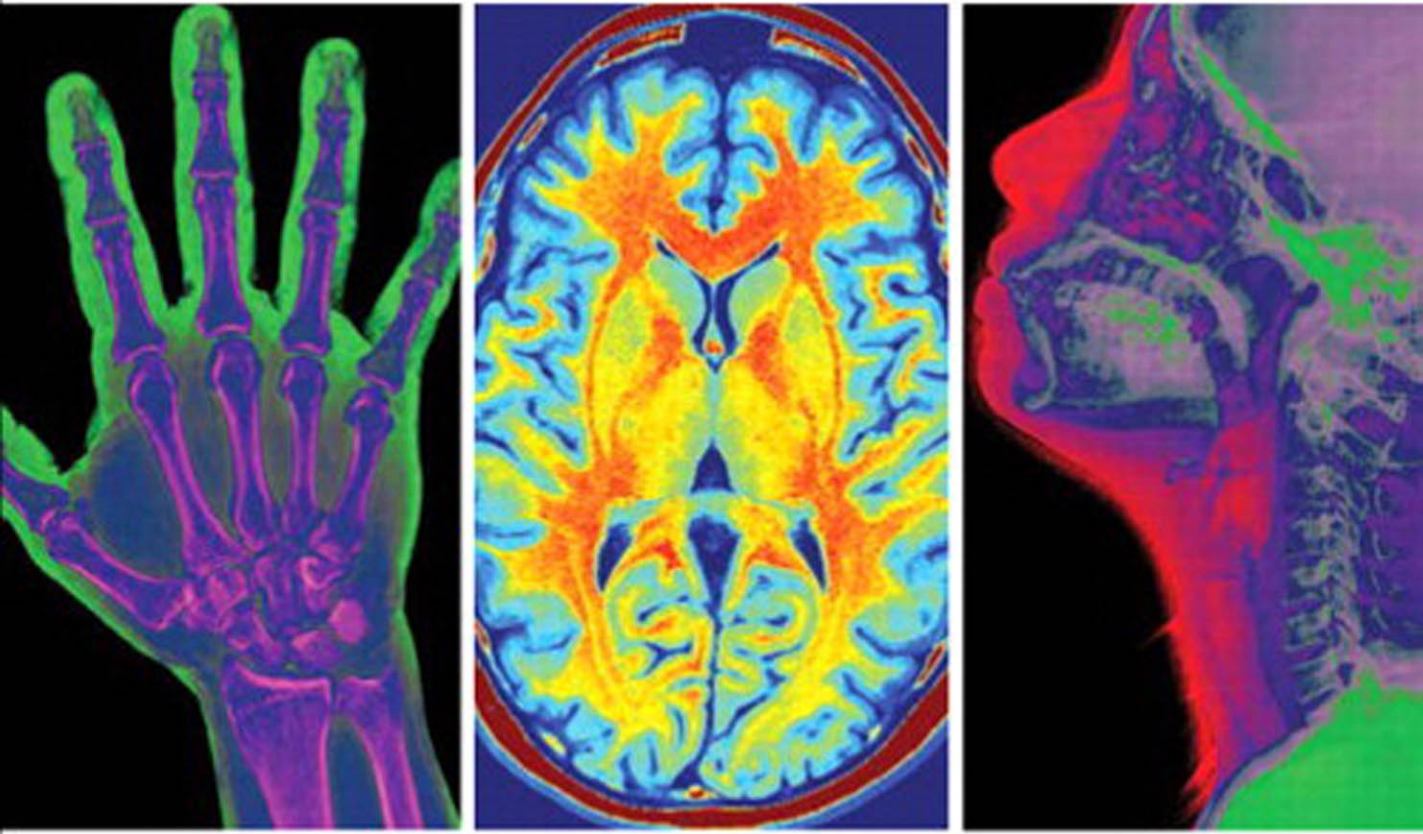
MEDICAL IMAGING

SIGNALS AND SYSTEMS

SECOND EDITION



JERRY L. PRINCE
JONATHAN M. LINKS

Medical Imaging Signals and Systems

This page intentionally left blank

Medical Imaging Signals and Systems

Jerry L. Prince
Electrical and Computer Engineering
Whiting School of Engineering
Johns Hopkins University

Jonathan M. Links
Environmental Health Sciences
Bloomberg School of Public Health
Johns Hopkins University

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Vice President and Editorial
Director, ECS: *Marcia J. Horton*
Executive Editor: *Holly Stark*
Marketing Manager: *Tim Galligan*
Marketing Assistant: *Jon Bryant*
Senior Managing Editor: *Scott Disanno*
Project Manager: *Irwin Zucker*
Art Director: *Jayne Conte*
Cover Designer: *Bruce Kenselaar*
Cover Image: From left to right: X-ray image of the hand, MRI cross section of the brain, X-ray image of the head and neck. Photos Courtesy of GE Healthcare.
Full-Service Project Management/
Composition: *SPI Global*
Printer/Binder: *Courier Kendallville*
Cover Printer: *Courier Kendallville*

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text.

Copyright © 2015, 2006 by Pearson Education, Inc., publishing as Prentice Hall, 1 Lake Street, Upper Saddle River, NJ 07458.

All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. To obtain permission(s) to use material from this work, please submit a written request to Pearson Education, Inc., Permissions Department, imprint permissions address.

Many of the designations by manufacturers and seller to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed in initial caps or all caps.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Library of Congress Cataloging-in-Publication Data

Prince, Jerry L., author.
Medical imaging signals and systems / Jerry L. Prince, Jonathan M. Links. –2.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-13-214518-3 (alk. paper)
I. Links, Jonathan M., author. II. Title.
[DNLM: 1. Diagnostic Imaging. 2. Signal Processing, Computer-Assisted. WN 180]
RC78.7.D53
616.07'54–dc23
2014000639

10 9 8 7 6 5 4 3 2 1

PEARSON

ISBN 10: 0-13-214518-9
ISBN 13: 978-0-13-214518-3

To our families
Carol, Emily, Ben, Mark, and David
Laura, Annie, and Beth
who help us see what's important and what's not.

This page intentionally left blank

Contents

Preface xv

Part I Basic Imaging Principles 1

1

Introduction 5

1.1 History of Medical Imaging 5
1.2 Physical Signals 6
1.3 Imaging Modalities 7
1.4 Projection Radiography 7
1.5 Computed Tomography 9
1.6 Nuclear Medicine 10
1.7 Ultrasound Imaging 11
1.8 Magnetic Resonance Imaging 12
1.9 Multimodality Imaging 13
1.10 Summary and Key Concepts 13

2

Signals and Systems 15

2.1 Introduction 15
2.2 Signals 16
  2.2.1 Point Impulse 16
  2.2.2 Line Impulse 19
  2.2.3 Comb and Sampling Functions 19
  2.2.4 Rect and Sinc Functions 20
  2.2.5 Exponential and Sinusoidal Signals 22
  2.2.6 Separable Signals 23
  2.2.7 Periodic Signals 23
2.3 Systems 24
  2.3.1 Linear Systems 24
  2.3.2 Impulse Response 25
  2.3.3 Shift Invariance 25
  2.3.4 Connections of LSI Systems 28

  - **2.3.5** Separable Systems 30
  - **2.3.6** Stable Systems 31
- **2.4** The Fourier Transform 31
- **2.5** Properties of the Fourier Transform 36
  - **2.5.1** Linearity 36
  - **2.5.2** Translation 37
  - **2.5.3** Conjugation and Conjugate Symmetry 37
  - **2.5.4** Scaling 37
  - **2.5.5** Rotation 38
  - **2.5.6** Convolution 38
  - **2.5.7** Product 39
  - **2.5.8** Separable Product 40
  - **2.5.9** Parseval's Theorem 40
  - **2.5.10** Separability 40
- **2.6** Transfer Function 41
- **2.7** Circular Symmetry and the Hankel Transform 43
- **2.8** Summary and Key Concepts 47

3

Image Quality 54

- **3.1** Introduction 54
- **3.2** Contrast 55
  - **3.2.1** Modulation 56
  - **3.2.2** Modulation Transfer Function 56
  - **3.2.3** Local Contrast 60
- **3.3** Resolution 61
  - **3.3.1** Line Spread Function 61
  - **3.3.2** Full Width at Half Maximum 62
  - **3.3.3** Resolution and Modulation Transfer Function 63
  - **3.3.4** Subsystem Cascade 65
  - **3.3.5** Resolution Tool 68
  - **3.3.6** Temporal and Spectral Resolution 68
- **3.4** Noise 69
  - **3.4.1** Random Variables 70
  - **3.4.2** Continuous Random Variables 70
  - **3.4.3** Discrete Random Variables 72
  - **3.4.4** Independent Random Variables 75
- **3.5** Signal-to-Noise Ratio 76
  - **3.5.1** Amplitude SNR 77
  - **3.5.2** Power SNR 77
  - **3.5.3** Differential SNR 79
  - **3.5.4** Decibels 80
- **3.6** Sampling 80
  - **3.6.1** Signal Model for Sampling 81
  - **3.6.2** Nyquist Sampling Theorem 83
  - **3.6.3** Anti-Aliasing Filters 85

3.7 Other Effects 86
  3.7.1 Artifacts 86
  3.7.2 Distortion 88
3.8 Accuracy 88
  3.8.1 Quantitative Accuracy 89
  3.8.2 Diagnostic Accuracy 89
3.9 Summary and Key Concepts 92

Part II Radiographic Imaging 101

4
Physics of Radiography 106
4.1 Introduction 106
4.2 Ionization 107
  4.2.1 Atomic Structure 107
  4.2.2 Electron Binding Energy 109
  4.2.3 Ionization and Excitation 109
4.3 Forms of Ionizing Radiation 110
  4.3.1 Particulate Radiation 110
  4.3.2 Electromagnetic Radiation 112
4.4 Nature and Properties of Ionizing Radiation 113
  4.4.1 Primary Energetic Electron Interactions 114
  4.4.2 Primary Electromagnetic Radiation Interactions 116
4.5 Attenuation of Electromagnetic Radiation 120
  4.5.1 Measures of X-Ray Beam Strength 121
  4.5.2 Narrow Beam, Monoenergetic Photons 123
  4.5.3 Narrow Beam, Polyenergetic Photons 125
  4.5.4 Broad Beam Case 127
4.6 Radiation Dosimetry 127
  4.6.1 Exposure 127
  4.6.2 Dose and Kerma 128
  4.6.3 Linear Energy Transfer (LET) 128
  4.6.4 The *f*-Factor 128
  4.6.5 Dose Equivalent 129
  4.6.6 Effective Dose 130
4.7 Summary and Key Concepts 131

5
Projection Radiography 135
5.1 Introduction 135
5.2 Instrumentation 136
  5.2.1 X-Ray Tubes 136
  5.2.2 Filtration and Restriction 139

- **5.2.3** Compensation Filters and Contrast Agents 141
- **5.2.4** Grids, Airgaps, and Scanning Slits 143
- **5.2.5** Film-Screen Detectors 146
- **5.2.6** X-Ray Image Intensifiers 148
- **5.2.7** Digital Radiography 149
- **5.2.8** Mammography 154

5.3 Image Formation 154
- **5.3.1** Basic Imaging Equation 154
- **5.3.2** Geometric Effects 155
- **5.3.3** Blurring Effects 162
- **5.3.4** Film Characteristics 166

5.4 Noise and Scattering 169
- **5.4.1** Signal-to-Noise Ratio 169
- **5.4.2** Quantum Efficiency and Detective Quantum Efficiency 171
- **5.4.3** Compton Scattering 173

5.5 Summary and Key Concepts 175

6

Computed Tomography 186

6.1 Introduction 186

6.2 CT Instrumentation 188
- **6.2.1** CT Generations 188
- **6.2.2** X-Ray Source and Collimation 194
- **6.2.3** Dual-Energy CT 194
- **6.2.4** CT Detectors 195
- **6.2.5** Gantry, Slip Ring, and Patient Table 196

6.3 Image Formation 197
- **6.3.1** Line Integrals 197
- **6.3.2** CT Numbers 198
- **6.3.3** Parallel-Ray Reconstruction 198
- **6.3.4** Fan-Beam Reconstruction 208
- **6.3.5** Helical CT Reconstruction 212
- **6.3.6** Cone Beam CT 213
- **6.3.7** Iterative Reconstruction 213

6.4 Image Quality in CT 213
- **6.4.1** Resolution 214
- **6.4.2** Noise 216
- **6.4.3** Artifacts 221

6.5 Summary and Key Points 223

Part III Nuclear Medicine Imaging 235

7

The Physics of Nuclear Medicine 239

7.1 Introduction 239
7.2 Nomenclature 240
7.3 Radioactive Decay 240
  7.3.1 Mass Defect and Binding Energy 240
  7.3.2 Line of Stability 242
  7.3.3 Radioactivity 243
  7.3.4 Radioactive Decay Law 243
7.4 Modes of Decay 245
  7.4.1 Positron Decay and Electron Capture 245
  7.4.2 Isomeric Transition 246
7.5 Statistics of Decay 247
7.6 Radiotracers 249
7.7 Summary and Key Concepts 251

8

Planar Scintigraphy 255

8.1 Introduction 255
8.2 Instrumentation 255
  8.2.1 Collimators 256
  8.2.2 Scintillation Crystal 258
  8.2.3 Photomultiplier Tubes 258
  8.2.4 Positioning Logic 260
  8.2.5 Pulse Height Analyzer 260
  8.2.6 Gating Circuit 262
  8.2.7 Image Capture 263
  8.2.8 Solid State and Other New Cameras 264
8.3 Image Formation 264
  8.3.1 Event Position Estimation 264
  8.3.2 Acquisition Modes 266
  8.3.3 Anger Camera Imaging Equation 269
8.4 Image Quality 272
  8.4.1 Resolution 273
  8.4.2 Sensitivity 276
  8.4.3 Uniformity 278
  8.4.4 Energy Resolution 279

- **8.4.5** Noise 280
- **8.4.6** Factors Affecting Count Rate 281

8.5 Summary and Key Concepts 282

9

Emission Computed Tomography 293

9.1 Instrumentation 294
- **9.1.1** SPECT Instrumentation 294
- **9.1.2** PET Instrumentation 298

9.2 Image Formation 304
- **9.2.1** SPECT Image Formation 304
- **9.2.2** PET Image Formation 309
- **9.2.3** Iterative Reconstruction 313

9.3 Image Quality in SPECT and PET 317
- **9.3.1** Spatial Resolution 318
- **9.3.2** Attenuation and Scatter 319
- **9.3.3** Random Coincidences 320
- **9.3.4** Contrast 320
- **9.3.5** Noise and Signal-to-Noise Ratio 321

9.4 Summary and Key Concepts 321

Part IV Ultrasound Imaging 331

10

The Physics of Ultrasound 335

10.1 Introduction 335

10.2 The Wave Equation 336
- **10.2.1** Three-Dimensional Acoustic Waves 336
- **10.2.2** Plane Waves 338
- **10.2.3** Spherical Waves 340

10.3 Wave Propagation 341
- **10.3.1** Acoustic Energy and Intensity 341
- **10.3.2** Reflection and Refraction at Plane Interfaces 342
- **10.3.3** Transmission and Reflection Coefficients at Plane Interfaces 343
- **10.3.4** Attenuation 344
- **10.3.5** Scattering 347
- **10.3.6** Nonlinear Wave Propagation 347

10.4 Doppler Effect 349

10.5 Beam Pattern Formation and Focusing 353
  10.5.1 Simple Field Pattern Model 354
  10.5.2 Diffraction Formulation 355
  10.5.3 Focusing 361
10.6 Summary and Key Concepts 362

11

Ultrasound Imaging Systems 367
11.1 Introduction 367
11.2 Instrumentation 367
  11.2.1 Ultrasound Transducer 367
  11.2.2 Ultrasound Probes 372
11.3 Pulse-Echo Imaging 374
  11.3.1 The Pulse-Echo Equation 374
11.4 Transducer Motion 377
11.5 Ultrasound Imaging Modes 380
  11.5.1 A-Mode Scan 380
  11.5.2 M-Mode Scan 381
  11.5.3 B-Mode Scan 381
11.6 Steering and Focusing 386
  11.6.1 Transmit Steering and Focusing 386
  11.6.2 Beamforming and Dynamic Focusing 388
11.7 Three-Dimensional Ultrasound Imaging 391
11.8 Image Quality 392
  11.8.1 Resolution 392
  11.8.2 Noise and Speckle 395
11.9 Summary and Key Concepts 396

Part V Magnetic Resonance Imaging 407

12

Physics of Magnetic Resonance 410
12.1 Introduction 410
12.2 Microscopic Magnetization 410
12.3 Macroscopic Magnetization 412
12.4 Precession and Larmor Frequency 414
12.5 Transverse and Longitudinal Magnetization 416
  12.5.1 NMR Signals 417
  12.5.2 Rotating Frame 419

12.6 RF Excitation 419
12.7 Relaxation 422
12.8 The Bloch Equations 425
12.9 Spin Echoes 426
12.10 Basic Contrast Mechanisms 429
12.11 Summary and Key Concepts 433

13

Magnetic Resonance Imaging 439

13.1 Instrumentation 439
  13.1.1 System Components 439
  13.1.2 Magnet 441
  13.1.3 Gradient Coils 442
  13.1.4 Radio Frequency Coils 445
  13.1.5 Scanning Console and Computer 446
13.2 MRI Data Acquisition 447
  13.2.1 Encoding Spatial Position 447
  13.2.2 Slice Selection 449
  13.2.3 Frequency Encoding 455
  13.2.4 Polar Scanning 460
  13.2.5 Gradient Echoes 461
  13.2.6 Phase Encoding 462
  13.2.7 Spin Echoes 465
  13.2.8 Pulse Repetition Interval 467
  13.2.9 Realistic Pulse Sequences 467
13.3 Image Reconstruction 469
  13.3.1 Rectilinear Data 470
  13.3.2 Polar Data 471
  13.3.3 Imaging Equations 472
13.4 Image Quality 475
  13.4.1 Sampling 475
  13.4.2 Resolution 477
  13.4.3 Noise 479
  13.4.4 Signal-to-Noise Ratio 481
  13.4.5 Artifacts 482
13.5 Advanced Contrast Mechanisms 483
13.6 Summary and Key Concepts 487

Index 497

Preface

Although the underlying principles of medical imaging have not changed in the nine years since the first edition of this book was published, the instrumentation and practices have continued to evolve and improve. This second edition maintains the *signals and systems* focus of the first edition, with up-to-date descriptions of instrumentation. We still cover the most important *imaging modalities* in radiology: projection radiography, x-ray computed tomography, nuclear medicine scintigraphy and emission tomography, ultrasound imaging, and magnetic resonance imaging. But we now provide additional material on digital radiography, multi-row detector CT systems, 3D ultrasound, both functional and diffusion-weighted magnetic resonance imaging, and much more. As before, we expect the reader to be familiar with signals and systems, which are usually covered in the sophomore year of most engineering curricula, and with elementary probability. Freshman courses in physics, chemistry, and calculus are also assumed.

As with the first edition, the book is organized into parts emphasizing key overall conceptual divisions. Part I introduces basic imaging principles, including an introduction to medical imaging systems in Chapter 1, a review of signal processing (with emphasis on two-dimensional signals) in Chapter 2, and a discussion of image quality in Chapter 3. Our presentation of the theory of medical imaging systems is strongly based on continuous signals; however, a development of discrete signals is included to permit discussions on sampling and implementation. Issues of image quality, including resolution, noise, contrast, geometric distortion, and artifacts are described in a general context here, and are revisited within each modality in subsequent chapters.

Part II describes key modalities in radiographic imaging. It begins in Chapter 4 with a brief presentation of the physics of radiography, including the generation and detection of ionizing radiation and its effect on the human body. Chapter 5 describes projection radiography systems, including chest x-ray, fluoroscopy, and mammography systems. As in all subsequent chapters, coverage focuses on signals, including only enough physics and biology to motivate the modality and provide a model for the analysis. Chapter 5 also presents the mathematics of *projection imaging*, a very fundamental idea in medical imaging. Chapter 6 covers x-ray computed tomography, expanding on the instrumentation and mathematics of projection imaging and introducing the concept of image reconstruction in medical imaging. Computed tomography produces true *tomograms* (images of cross sections of the body) rather than projections of the body.

Part III presents the physics and modalities of nuclear medicine imaging. Chapter 7 describes the physics of nuclear medicine, focusing primarily on the concept of radioactivity. The major modalities in nuclear medicine imaging are described in Chapter 8, which covers planar scintigraphy, and Chapter 9, which covers emission computed tomography.

Part IV covers ultrasound imaging. It begins in Chapter 10 with a brief presentation of the physics of sound, and continues in Chapter 11 with the various imaging modes offered within this rich modality. Part V covers magnetic resonance imaging. Chapter 12 presents the physics of nuclear magnetic resonance, and Chapter 13 continues with a presentation of various magnetic resonance imaging techniques.

We have used the first edition of this book for a one-semester upper-level/graduate course on medical imaging systems. In order to cover the material in one semester, we routinely skip some material in the book, and we move at a very brisk pace. Although it was very tempting to add more depth in modern instrumentation, reconstruction methods, and diagnostic uses of medical imaging, we feel that this breadth of material could not be covered in one semester with sufficient depth, and would be inconsistent with our primary goal of providing a unified view of medical imaging from a signals and systems point of view. On the other hand, we feel that this book could be used as the basis for a two-semester course, perhaps by covering Parts I–III in the first semester and Parts IV–V in the second semester. A two-semester approach would allow instructors to use supplementary materials for additional depth in the physics and instrumentation of medical imaging, or to present current research topics.

Medical imaging is very visual—just ask any radiologist. Although the formalism of signals and systems is mathematical, we understand the advantages offered through visualization. Therefore, the book contains many images and diagrams. Some are strictly pedagogical, offered in conjunction with the exposition or an example problem. Others are motivational, revealing interesting features for discussion or study. Special emphasis is made to provide biologically relevant examples, so that the important context of medical imaging can be appreciated by students. Many images have been added or replaced in this edition, in order to provide better coverage of current use and to provide reference images to help explain features and qualities of the various modalilties.

New to This Edition

The second edition of this book arose primarily from the need to provide updates to the technology and methods in medical imaging systems, which have undergone substantial development since the first edition. At the same time, we were able to incorporate changes to the organization of the book and to improve certain aspects of pedagogy. Instructors and students alike now have more modern material from the core medical imaging modalities while still maintaining the signal processing perspective in a unified treatment of medical imaging signal and systems.

The most significant changes to this new edition include:

- Completely rewritten overview sections including many new images to better motivate and explain the core modalities that use x-rays, radioactivity, ultrasound, and nuclear magnetic resonance.
- New sections on digital radiography systems and mammography in projection radiography.
- A new section on multi-row detectors in computed tomography.

- A new section on iterative reconstruction in emission tomography in nuclear medicine.
- New sections on nonlinear wave propagation and harmonic imaging in ultrasound imaging.
- New development and presentation of imaging equations in planar scintigraphy, single photon emission computed tomography, and positron emission tomography.
- New sections on three-dimensional imaging, noise, and speckle in ultrasound imaging.
- New sections on susceptibility weighted imaging, functional magnetic resonance imaging, and diffusion magnetic resonance imaging in magnetic resonance imaging.
- Reorganization of the chapters on signals and systems and image quality to encourage a better pedagogical flow.
- Many new problems, added primarily to the chapters having relatively fewer problems in the first edition. There are a total of 261 problems in this second edition.

Acknowledgments

Many students, friends, colleagues, and teaching assistants contributed to this book through discussions and critiques. We wish to thank Elliot R. McVeigh and John I. Goutsias, who co-taught our course at Johns Hopkins University during the early years and helped draft the first edition of the book. We are grateful to Drs. Avneesh Chhabra, Harvey Ziessman, Peter Calabresi, and David Yousem for providing several clinical images (and their descriptions) and to Vince Blasko and Beatrice Mudge for providing several images depicting artifacts that can arise in clinical imaging. Xiao Han, Xiaodong Tao, Li Pan, Vijay Parthasarathy, Tara Johnson, Minnan Xu, Abd El-Monem El-Sharkawy, Kahaled Abd-Elmoniem, Lotta Ellingsen, Jing Wan, Snehashis Roy, Harsh Agarwal, Issel Lim, Xian Fan, Nan Li, Sahar Soleimanifard, Nathanel Kuo, Jeffrey Pompe, Min Chen, Chuyang Ye, and Zhen Yang contributed problems and solutions, Dhruv Lamba helped find new images, and Aaron Carass fixed many LaTeX problems. We are grateful for the images provided to us by GE Healthcare, Philips Healthcare, and Osirix. We thank J. Webster Stayman for reviewing the digital radiography section. We would also like to thank six anonymous reviewers who provided comments on their experiences with the first edition, which provided a basis for many of the changes we have made to this edition, and five anonymous reviewers who reviewed a draft of this edition, which helped us to better balance material in several sections. Finally, we would like to thank Dr. William R. Brody, who inspired the creation of the course out of which this book emerged.

Jerry L. Prince
Jonathan M. Links

This page intentionally left blank

Basic Imaging Principles

PART I

Overview

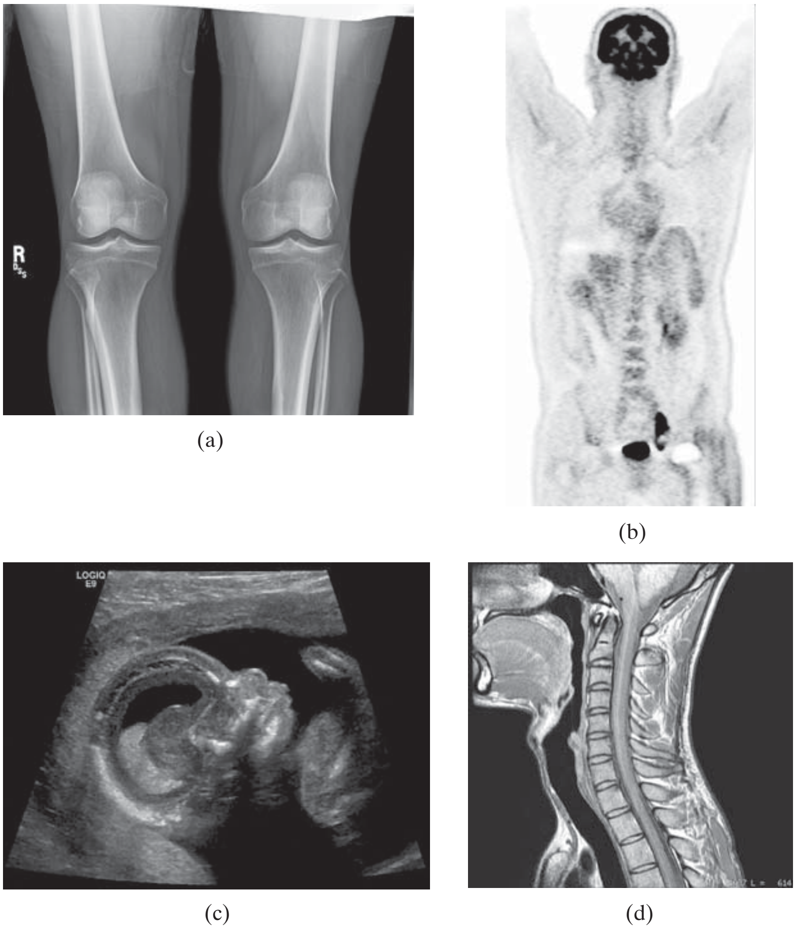
What does the human body look like *on the inside*? The smart answer: It depends on how you look at it. The most direct way to look inside the human body is to cut it open, for example, through surgery. A refinement of this procedure might be to use an endoscope, essentially a light tube that is "threaded" through the body, which conveys an image to a display device. Both methods offer direct optical viewing, but also involve cutting the body, putting something in it, or both. These are *invasive techniques*, which cause (potential) damage or trauma to the body.

The beauty of medical imaging is that we can see inside the human body in ways that are less invasive than surgery or endoscopy. In some cases—for example, magnetic resonance imaging (MRI) and ultrasound imaging—the methods are completely *noninvasive* and risk-free so far as we know. In other cases—for example, projection radiography, x-ray computed tomography (CT), and nuclear medicine—there is some risk associated with the radiation exposure, even though these methods are considered noninvasive as well.

Fundamentally, these medical imaging techniques mean that we do not need to cut the body or put a physical device into it in order to "see inside." Of perhaps even greater importance, these techniques allow us to see things that are not visible to the naked eye in the first place. For example, functional magnetic resonance imaging (fMRI) allows us to obtain images of organ perfusion or blood flow, and positron emission tomography (PET) allows us to obtain images of metabolism or receptor binding. In other words, the various imaging techniques allow us to see inside the body in different ways—the "signal" is different in each case and can reveal information which the other methods cannot. Each of these different methods is a different *imaging modality*, and the "signals" that arise are intrinsically different. This hearkens back to the opening question: What does the human body look like on the inside? The answer: It depends on the measured signal of interest.

In this book, we use a *signals and systems* approach to explain and analyze the most common imaging methods in radiology today. We want to answer the question: What do the images look like and why? We will discover that medical imaging physics allows us to image certain parameters of the body's tissues, such as reflectivity in ultrasound imaging, linear attenuation coefficient in computed tomography, and hydrogen proton density in magnetic resonance imaging. These physical parameters, which one can think of as "signals" within the body, represent the input signal into an imaging system. In medical imaging, the "object" or "signal" arising from the patient depends on the physical processes governing a given imaging modality. Thus, a given patient represents an ensemble of different objects or signals. In considering a given medical image, it is thus important to start with the physics that underlie the creation of signals from the patient for that modality. Accordingly, each part of this book is organized such that the first chapter describes the relevant physics, and subsequent chapters describe those modalities based on the specific physical processes of that part.

The first output of any medical imaging system is based on physical measurements, which might be returning echoes in an ultrasound system, x-ray



(a)

(b)

(c)

(d)

Figure I.1
The four main medical imaging signals discussed in this book: (a) x-ray transmission through the body, (b) gamma ray emission from within the body, (c) ultrasound echoes, and (d) nuclear magnetic resonance induction. The corresponding medical imaging modalities are projection radiography, planar scintigraphy, ultrasound imaging, and magnetic resonance imaging. All images courtesy of GE Healthcare.

intensities in a CT system, or radio frequency waves in an MRI system. The final output in this system is created through *image reconstruction*, the process of creating an image from measurements of signals. The overall quality of a medical image is determined by how well the image portrays the true spatial distribution of the physical parameter(s) of interest within the body. Resolution, noise, contrast, geometric distortion, and artifacts are important considerations in our study of image quality. Ultimately, the clinical utility of a medical image involves both the image's quality and the medical information contained in the parameters themselves.

Figure I.1 shows the four main medical imaging signals discussed in this book: (1) x-ray transmission through the body, (2) gamma ray emission from within the body, (3) ultrasound echoes, and (4) nuclear magnetic resonance induction. Part II covers modalities that use x-ray transmission signals, Part III covers modalities that use gamma ray emission, Part IV covers modalities that use ultrasound signals, and Part V covers magnetic resonance imaging, which uses signals that arise from nuclear magnetic resonance. The specific medical imaging modalities depicted in Figure I.1 are (1) projection radiography, (2) positron emission tomography, (3) ultrasound imaging, and (4) magnetic resonance imaging.

In Figure I.1, parts (a) and (b) represent two-dimensional *projection* images of the three-dimensional human body. A projection is created as a two-dimensional "shadow" of the body, a process that is illustrated in Figure I.2. Figures I.1(c) and (d) are slices within the body. Figure I.3 depicts the three

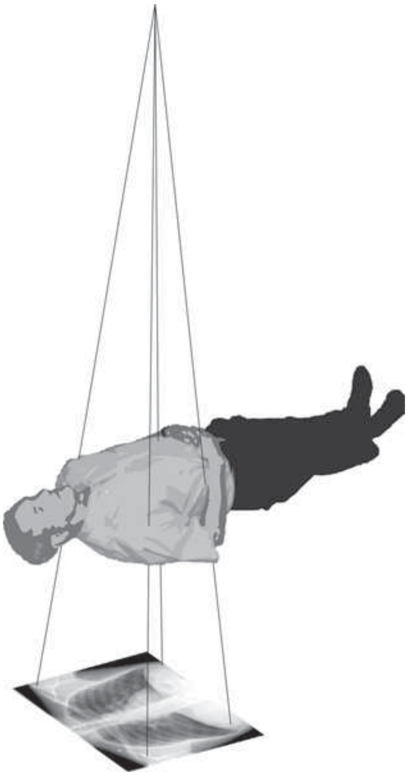


Figure I.2
The creation of a two-dimensional projection through the body. In this case, x-rays are transmitted through a patient creating a radiograph.

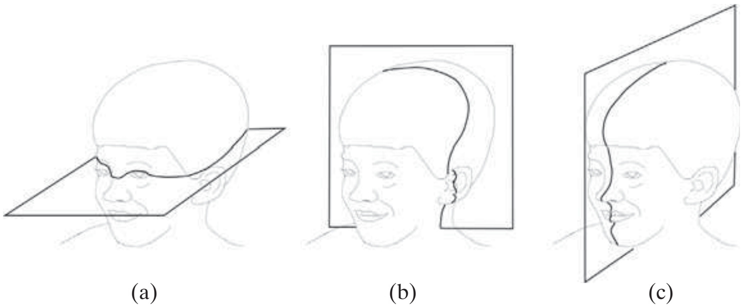


Figure I.3
The three standard orthogonal tomographic or slice or section views: (a) axial or transaxial or transverse, (b) coronal or frontal, and (c) sagittal.

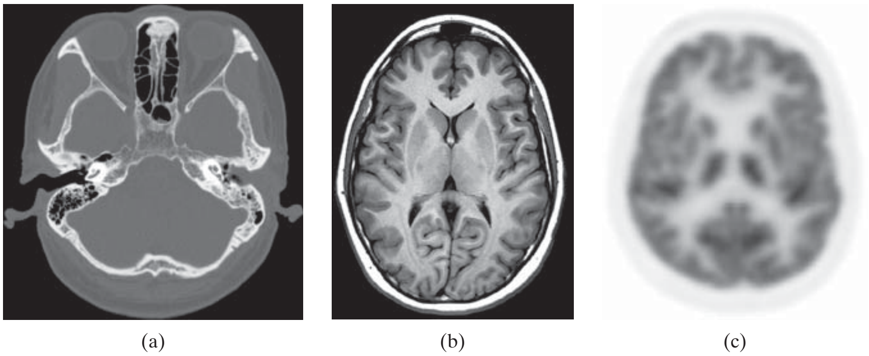


Figure I.4
Representative transverse slice through the brain from three different imaging modalities: (a) computed tomography, (b) magnetic resonance imaging, and (c) positron emission tomography.

standard orientations of slice (or *tomographic*) images, *axial*, *coronal*, and *sagittal*. Figure I.1(d) is a sagittal slice, while Figure I.1(c) is an *oblique* slice, that is, an orientation not corresponding to one of the standard slice orientations.

Figure I.4 also shows slice images. In this case, each image is a transverse slice, oriented perpendicular to the head and body axis through the brain. Each image is obtained from a different imaging modality: (a) computed tomography, (b) magnetic resonance imaging, and (c) positron emission tomography. Even though each image depicts (a slice through) the brain, the images are strikingly different, because the signals giving rise to each image are themselves strikingly different. In this part of the book, we study the common signal processing concepts that relate to all imaging modalities, setting the groundwork for adding the physical differences that account for the different appearances of the imaging modalities, and hence their different uses in medicine.

CHAPTER 1

Introduction

In this book, we take a signals and systems approach to the characterization of medical imaging. As discussed in the Overview, there are a variety of signals in which we are interested; ultimately, this interest stems from the biological and medical significance of these signals in patients with various diseases. In practice, these signals are transformed into images via medical imaging modalities. In this chapter, we begin to consider these modalities and their characteristics.

1.1 History of Medical Imaging

The first published medical image was a radiograph of the hand of Wilhelm Conrad Roentgen's wife in December 1895. Roentgen had been experimenting with a Crooke's tube (the forerunner of today's x-ray tube) and noticed that "a new kind of rays" (hence, *x-rays*) were emitted that could expose a photographic plate even when optically shielded. It was immediately obvious to Roentgen that his discovery could have a profound impact in medicine. Indeed, the first clinical use of x-rays occurred only two months later, in February 1896. The use of x-rays became widespread, and both static and dynamic (*fluoroscopic*) techniques were developed. Here, a *static* technique refers to an image taken at a single point in time, whereas a *dynamic* technique refers to a series of images acquired over time.

For many decades, these *planar* (i.e., two-dimensional projection) radiographs were the only medical images being produced. Ultimately, radiography was extended into transmission computed tomography, or *cross-sectional* imaging. Godfrey Hounsfield produced the first true computed tomography (CT) scanner in 1972 at EMI in England. He used mathematical methods for image reconstruction developed a decade earlier by Allan Cormack of the United States. Hounsfield and Cormack shared the Nobel Prize in Medicine in 1979. Many radiologists consider CT scanning to be the most important development in medical imaging since Roentgen's original discovery.

As radiography arose from the discovery of x-rays, nuclear medicine arose from the discovery of radioactivity by Antoine Henri Becquerrel in 1896. Initially, radionuclides were used in cancer therapy rather than in medical imaging. The concept of using radioactive *tracers* to study physiology was introduced by George de Hevesy in 1923; de Hevesy is considered the father of nuclear medicine. A radiotracer is a radioactively labeled drug that mimics a biological compound of interest; the distribution of the radioactivity implies the distribution of the drug. Early studies with radiotracers used conventional nonimaging radiation detectors to roughly determine amounts of radioactivity in various body regions. In 1949, Benedict Cassen at UCLA started the development of the first imaging system in nuclear medicine, the *rectilinear scanner*. The modern *Anger scintillation camera* was developed by Hal Anger at UC Berkeley in 1952. The element of the most commonly used radionuclide in nuclear medicine, technetium-99m, was discovered in 1937 by Carlo Perrier and Emilio Segre; its first use in medicine was in 1961.

The interaction of acoustic waves with media was first described by Lord John Rayleigh over one hundred years ago in the context of the propagation of sound in air. Modern ultrasound imaging had its roots in World War II Navy sonar technology, and initial medical applications focused on the brain. Ultrasound technology progressed through the 1960s from A-mode, B-mode, and M-mode scans to today's two-dimensional (2-D) Doppler, three-dimensional (3-D), and nonlinear imaging systems.

The phenomenon of *nuclear magnetic resonance*, from which magnetic resonance imaging (MRI) arises, was first described by Felix Bloch and Edward Purcell; they shared the 1952 Nobel Prize in Physics. This work was extended by Richard Ernst, who received the Nobel Prize in Chemistry in 1991. In 1971, Raymond Damadian published a paper suggesting the use of magnetic resonance (MR) in medical imaging; in 1973, a paper by Paul Lauterbur followed. Lauterbur received the Nobel Prize in Medicine in 2003, along with Peter Mansfield, who developed key methods in MRI.

1.2 Physical Signals

In this book, we consider the detection of different physical signals arising from the patient and their transformation into medical images. In practice, these signals arise from four processes:

- Transmission of x-rays through the body (in projection radiography and CT)
- Emission of gamma rays from radiotracers in the body (in nuclear medicine)
- Reflection of ultrasonic waves within the body (in ultrasound imaging)
- Precession of spin systems in a large magnetic field (in MRI)

Radiography, CT scanning, and nuclear medicine all make use of electromagnetic energy. Electromagnetic energy or waves consist of electric and magnetic waves traveling together at right angles. Wavelength and frequency are inversely related; frequency and energy are directly related. The electromagnetic spectrum spans the frequency range from zero to that of cosmic rays; only a relatively small portion of this spectrum is useful in medical imaging. At long wavelengths—for example, longer than 1 angstrom—most electromagnetic energy is highly attenuated by the body, prohibiting its exit and external detection. At wavelengths shorter than about 10^{-2} angstroms, the corresponding energy is too high to be readily detected.

In this book, we express energy in units of *electron volts* (eV), where 1 eV is the amount of energy an electron gains when accelerated across 1 volt potential. We will concentrate on electromagnetic radiation whose wavelengths correspond to energies of roughly 25–500 keV.

Ultrasound imaging utilizes sound waves, and considerations of attenuation and detection are similar to those above. Image resolution is not adequate when wavelengths longer than a couple of millimeters are used, and attenuation is too high for very short wavelengths. An ideal frequency range for ultrasound in medical imaging is 1–20 MHz, where 1 Hz = 1 cycle/second.

The signal in MRI arises from the precession (like the motion of a child's top or dreidel) of nuclei of the hydrogen atom—that is, protons. When placed in a large magnetic field, collections of protons, termed *spin systems*, can be set into motion by applying radio frequency (RF) currents through wire coils surrounding the patient. Although these spin systems precess at RF frequencies (64 MHz is typical), the primary signal source is not from radio waves, but from the Faraday induction of currents in the same or different wire coils.

1.3 Imaging Modalities

The medical imaging areas we consider in detail in this book are projection radiography, CT, nuclear medicine, ultrasound imaging, and MRI. An *imaging modality* is a particular imaging technique or system within one of these areas. In this section, we give a brief overview of these most common imaging modalities.

Projection radiography, CT, and nuclear medicine all use ionizing radiation. The first two transmit x-rays through the body, then use the fact that the body's tissues selectively *attenuate* (reduce) the x-ray intensities to form an image. These are termed *transmission* imaging modalities because they transmit energy through the body. In nuclear medicine, radioactive compounds are injected into the body. These compounds or *tracers* move selectively to different regions or organs within the body, emitting gamma rays with intensity proportional to the compound's local concentration. Nuclear medicine methods are *emission* imaging modalities because the radioactive sources emit radiation from within the body.

Ultrasound imaging transmits high-frequency sound into the body and receives the echoes returning from structures within the body. This method is often called *reflection* imaging because it relies on acoustic reflections to create images. Finally, MRI requires a combination of a high-strength magnetic field and radio frequency Faraday induction to image properties of the proton nucleus of the hydrogen atom. This technique is called *magnetic resonance imaging* since it exploits the property of nuclear magnetic resonance.

1.4 Projection Radiography

Projection radiography includes the following modalities:

- *Routine diagnostic radiography*, including chest x-rays, fluoroscopy, mammography, and motion tomography (a form of tomography that is not *computed* tomography)

- *Digital radiography*, which includes all the scans in routine radiography, but with images that are recorded digitally instead of on film
- *Angiography*, including universal angiography and angiocardiography, in which the systems are specialized for imaging the body's blood arteries and vessels
- *Neuroradiology*, which includes specialized x-ray systems for precision studies of the skull and cervical spine
- *Mobile x-ray systems*, which are small x-ray units designed for operating rooms or emergency vehicles
- *Mammography*, which includes film-based or digital-based systems optimized for breast imaging

All of these modalities are called "projection" radiography because they all represent the projection of a 3-D object or signal onto a 2-D image.

The common element in all of these systems is the *x-ray tube*. As we will see in Chapter 5, the x-ray tube generates an x-ray pulse in an approximately uniform "cone beam" (shaped like a cone) geometry. This pulse passes through the body and is attenuated by the intervening tissues. The x-ray intensity profile across the beam exiting from the body is no longer uniform—shadows have been created by dense objects (such as bone) in the body. This intensity distribution is revealed using a scintillator, which converts the x-rays to visible light. Finally, the light image on the scintillator is captured either on a large sheet of photographic film, a camera, or solid-state detectors.

The most common modality in projection radiography is the chest x-ray; a typical unit is shown in Figure 1.1(a). Here, the x-ray tube is located on the column projecting down from the ceiling. The scintillator and detector can be located either in the pedestal unit on the right or in the table itself. The radiologic technologist stands at a console not shown, protected by lead, but able to see through a window. A typical chest x-ray is shown in Figure 1.1(b). This image shows the spine, ribs, heart, lungs, and many other features radiologists are trained to identify and interpret. A key feature of this image is that structures located at different depths in the body are overlaid (or superimposed) on the 2-D image. For example, we can see both front and back ribs in the chest x-ray in Figure 1.1(b). This is a property of projection imaging, and it is common to

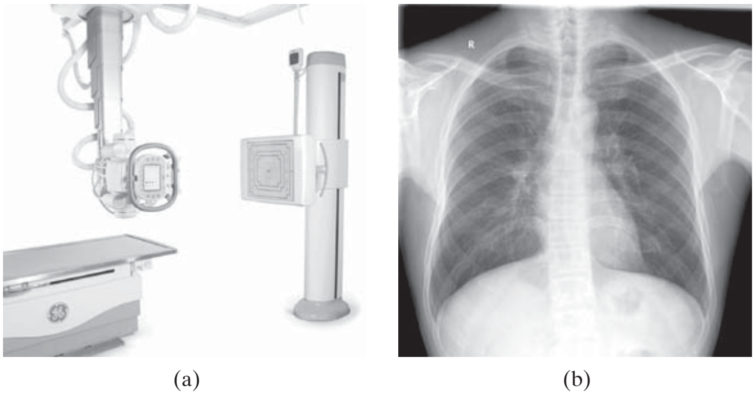

(a) (b)

Figure 1.1
(a) A chest x-ray unit and (b) a chest x-ray image. *Source*: Courtesy of GE Healthcare.

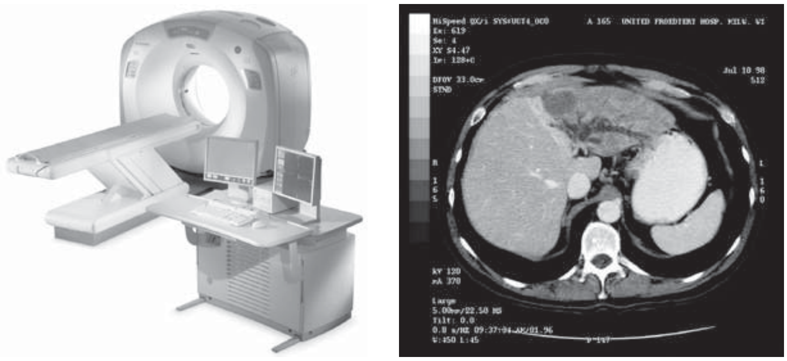
all projection radiographic methods. True *tomography*, the imaging of a 2-D slice of the 3-D body, cannot be directly accomplished using any modality in projection radiography. More details about projection radiography are given in Chapter 5.

1.5 Computed Tomography

As in projection radiography, CT uses x-rays. Unlike projection radiography, however, CT collects multiple projections of the same tissues from different orientations by moving the x-ray source around the body. CT systems have rows of digital detectors whose signals are input directly to a computer, and these signals are used to reconstruct one or more cross sections (slices) of the human body. In this way, although CT systems acquire projections that represent a "shadow" of the body, they generate truly tomographic images after reconstruction.

The important historical phases in CT development are *single-slice CT*, *helical CT*, and *multiple-row detector* CT (MDCT). Single-slice CT systems acquire data within a single plane and reconstruct only one plane per rotation. In helical CT systems, the x-ray tube and detectors continuously rotate around in a large circle, while the patient is moved in a continuous motion through the circle's center. From the patient's perspective, the x-ray tube carves out a helix; hence, the name helical CT. The importance of this technique is in its ability to rapidly acquire 3-D data, such as a whole body scan, in less than a minute. In MDCT systems, there are many rows of detectors used to rapidly gather a *cone* of x-ray data, comprising a 2-D projection of the 3-D patient. When the x-ray source and detectors revolve rapidly around the patient (one to two revolutions per second), very quick (near real-time) 3-D imaging is possible using these CT scanners.

A typical CT scanner is shown in Figure 1.2(a). In the center of the picture, we can see the cylindrical opening in which the patient lies; a patient table is also visible. Around the cylindrical opening is a housing containing both the x-ray tube and the detector array. The gantry holding these components is capable of spinning rapidly around the patient. The computer displays and keyboard in the foreground are used for entering patient data and viewing images. Although CT images can be printed on paper or film, the images are completely digital in nature since they are computed from the measured projections. The CT image



(a) (b)

Figure 1.2
(a) A CT scanner and (b) a CT image of a slice through the liver.
Source: Courtesy of GE Healthcare.

shown in Figure 1.2(b) is a slice through the liver; the data used to reconstruct this image were acquired in just one second. Computed tomography is described in more detail in Chapter 6.

1.6 Nuclear Medicine

Nuclear medicine imaging is distinguished from all other medical imaging modalities by the fact that images can only be made when appropriate radioactive substances are introduced into the body. These substances, which are either injected, ingested, or inhaled into the body, are trace amounts of biochemically active drugs whose molecules are labeled with radionuclides that emit gamma rays. These so-called *radiotracers* move within the body according to the body's natural uptake of the biological carrier molecule. For example, radioactive iodine can be used to study thyroid function. A nuclear medicine image reflects the local concentration of a radiotracer within the body. Since this concentration is tied to the physiological behavior of the carrier molecule within the body, nuclear medicine imaging is an example of a *functional imaging* method, whereas standard CT and MRI are *anatomical* or *structural imaging* methods.

There are three modalities within nuclear medicine: *conventional radionuclide imaging* or *scintigraphy*, *single-photon emission CT* (SPECT), and *positron emission tomography* (PET). Conventional radionuclide imaging and SPECT typically utilize a special 2-D gamma ray scintillation detector called an *Anger camera*. This camera is designed to detect single x-rays or gamma rays, rather than simply detecting the intensity of a collective beam as in projection radiography and CT. In conventional radionuclide imaging, this camera is conceptually analogous to the film or digital detector in projection radiography. A complication is that this procedure combines the effects of emission with the effects of attenuation of the rays by intervening body tissues, producing images that are 2-D projections of the 3-D distribution of radiotracers (which we wish to know) confounded by attenuation.

SPECT and PET produce images of slices within the body. SPECT does this by rotating the Anger camera around the body. Since the Anger camera is a 2-D imager, SPECT is fundamentally a 3-D imaging technique. In conventional radionuclide imaging and SPECT, a radioactive atom's decay produces a single gamma ray, which may intercept the Anger camera. In PET, however, a radionuclide decay produces a positron, which immediately annihilates (with an electron) to produce two gamma rays flying off in opposite directions. The PET scanner looks for coincident detections from opposing detectors in its ring, thus determining the line that passes through the site where the annihilation occurred. Both SPECT and PET use computed tomography reconstruction techniques including iterative reconstruction in order to create diagnostic images.

A SPECT scanner is shown in Figure 1.3(a). The Anger camera at the top (and side) is capable of rotating completely around the patient, who lies on the table. The table can move in coordination with the camera (for a helical tomographic scan), or the table can move with a stationary camera (for a whole-body standard projection scan). The collection of SPECT images shown in Figure 1.3(b) are cardiac scans taken at different spatial positions in the heart depicting blood flow to the heart muscle. Nuclear medicine imaging is described in more detail in Chapters 7 through 9.

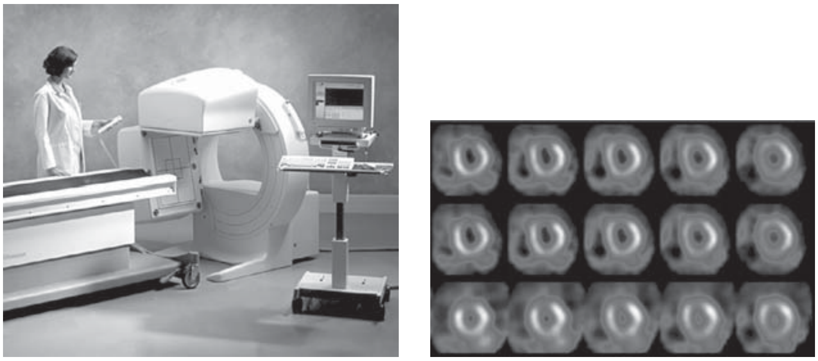

(a) (b)

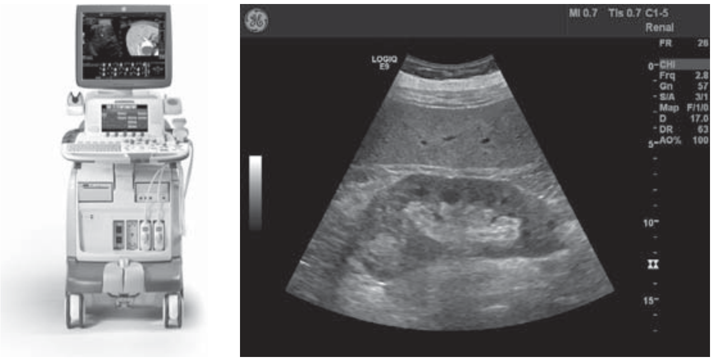
Figure 1.3
(a) A SPECT scanner and (b) a series of SPECT heart scans.
Source: Courtesy of GE Healthcare.

1.7 Ultrasound Imaging

Ultrasound imaging uses electrical-to-acoustical *transducers* to generate repetitive bursts of high-frequency sound. These pulses travel into the soft tissue of the body and reflect back to the transducer. The time-of-return of these pulses gives information about the location (depth) of a reflector, and the intensity of these pulses gives information about the strength of a reflector. By rapidly moving or scanning the transducer or its acoustical beam, real-time images of cross sections of soft tissue can be generated. Ultrasound imaging systems are comparatively inexpensive and completely noninvasive; therefore, these systems are widespread and in common usage. They are designed primarily to image anatomy and, although their image quality in some ways is not as good as other anatomical imaging modalities such as CT and MRI, they are real-time and highly adaptable to a wide variety of imaging goals.

Ultrasound imaging systems offer several imaging modalities:

- *A-mode imaging*, which generates a one-dimensional waveform, and as such does not really comprise an image. This mode, however, can provide very detailed information about rapid or subtle motion (e.g., of a heart valve).
- *B-mode imaging*, which is ordinary cross-sectional anatomical imaging. There are several arrangements of transducers that can give rise to images with different appearances.
- *M-mode imaging*, which generates a succession of A-mode signals, brightness modulated and displayed in real time on a computer display. M-mode generates an image that does not represent a cross section of anatomy but is important for measuring time-varying displacements of, for example, a heart valve.
- *Doppler imaging*, which uses the property of frequency or phase shift caused by moving objects (like a police siren that has a higher frequency when approaching and a lower frequency when departing) to generate images that are color coded by their motion. Doppler is most commonly used, however, in an audio mode. Running the frequency shifts through speakers allows an aural analysis of motion which is not possible with a visual display.
- *Nonlinear imaging*, which permits higher resolution imaging at greater depths and also permits imaging certain tissue properties.



(a) (b)

Figure 1.4
(a) An ultrasound scanner and (b) an ultrasound image of a kidney.
Source: Courtesy of GE Healthcare.

An ultrasound imaging system is shown in Figure 1.4(a). In addition to being inexpensive, ultrasound systems are small; most systems are on wheels, and can be rolled to the bedside or wherever needed. Very compact systems with limited functionality comprising just a transducer and a tablet or a smart phone are currently being marketed. Two transducers are shown in the figure, with their connecting cables; only one is used at any given time. Transducers with differing frequencies and geometries are usually available to serve the examination requirements. The ultrasound image shown in Figure 1.4(b) shows a human kidney. The textured appearance of ultrasound images is called *speckle*, and it is a form of artifact. Ultrasound imaging is described in detail in Chapters 10 and 11.

1.8 Magnetic Resonance Imaging

Magnetic resonance (MR) scanners use the property of *nuclear magnetic resonance* to create images. In a strong magnetic field, the nucleus of the hydrogen atom—a proton—tends to align itself with the field. Given the vast numbers of hydrogen atoms in the body, this tendency results in a net magnetization of the body. It is then possible to selectively excite regions within the body, causing groups of these "little magnets" to tip away from the magnetic field direction. As the protons return back into alignment with the field, they wobble like children's tops. Because protons are charged particles, this *precession* generates an RF electromagnetic signature, which can be sensed with an antenna.

There are many modes in which an MR scanner can operate; these scanners are incredibly flexible imaging devices. The most general categories of operation are the following:

- *Standard MRI*, which includes a whole host of *pulse sequences* (time-series of different excitation pulses)
- *Echo-planar imaging* (EPI), which utilizes specialized methods to generate images in real time
- *Magnetic resonance spectroscopic imaging* (MRS), which records images of other nuclei besides that of the hydrogen atom
- *Functional MRI* (fMRI), which uses oxygenation-sensitive pulse sequences to image blood oxygenation in the brain, which can be related to brain activity

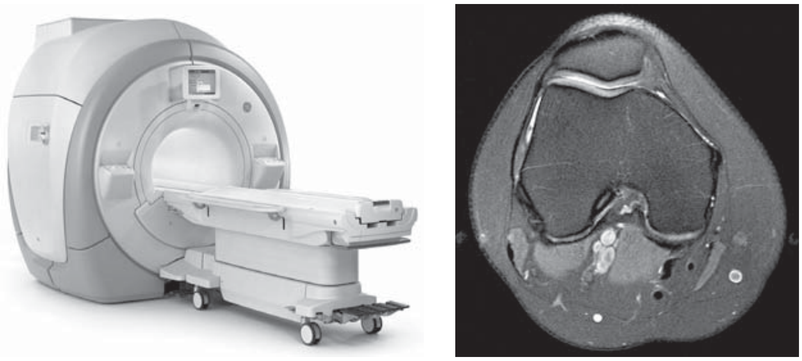

(a) (b)

Figure 1.5
(a) An MR scanner and (b) an MR image of a human knee.
Source: Courtesy of GE Healthcare.

- *Diffusion MRI* (dMRI), which images the degree and orientation of molecular diffusion in tissue

We mainly discuss standard MRI in this book, and brief descriptions of fMRI and dMRI are also provided.

An MR scanner is shown in Figure 1.5(a). Surrounding the tube in the center of the picture, within the housing, is a 3 tesla superconducting magnet. The field is very uniform within the bore, which is required for geometrically accurate imaging. The MR image in Figure 1.5(b) shows a cross section of a human knee. MRI is discussed in detail in Chapters 12 and 13.

1.9 Multimodality Imaging

Different medical imaging modalities reveal different properties of the human body. As a result, it is often useful to obtain diagnostic images of a single patient and medical condition using multiple modalities. For example, one might look at bones using CT and soft tissues using MRI, or one might examine brain structure using MRI and brain function using PET. Combining CT and PET is particularly useful because PET provides functional information that complements the structural information provided by CT and also because the CT data can be directly used to improve the reconstructed PET images. Because of this, virtually all modern PET systems are sold as PET/CT systems that combine the two modalities within a single platform.

1.10 Summary and Key Concepts

In practice, radiologists look for specific patterns in medical images. These patterns depend on both the patient and the imaging modality. It is the job of the engineers and scientists who develop medical imaging systems to produce images that are as accurate and useful as possible; these systems depend on the physics of each modality. In this chapter, we presented the following key concepts that you should now understand:

1. Medical imaging relies on *noninvasive techniques* to image body structures and function.

2. Each technique or method is a different imaging *modality*.
3. The main imaging modalities are *projection radiography*, *computed tomography*, *nuclear medicine*, *ultrasound imaging*, and *magnetic resonance imaging*.
4. The *signal* of interest is defined by the modality and specific imaging parameters.
5. Radiologists are trained to look for *specific patterns*, defined by the modality, specific imaging parameters, and differences in the expected signal in health and disease.

Further Reading

Brown, M.A. and Semelka, R.C. *MRI Basic Principles and Applications*. Hoboken, NJ: Wiley, 2003.

Bushberg, J.T., Seibert, J.A., Leidholdt, E.M., and Boone, J.M. *The Essential Physics of Medical Imaging*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins, 2012.

Bushhong, S.C. *Magnetic Resonance Imaging*. 3rd ed. St. Louis, MO: Mosby, 2003.

Carlton, R.R. and Adler, A.M. *Principles of Radiographic Imaging: An Art and a Science*, 5th ed. Clifton Park, NY: Delmar Cengage Learning, 2012.

Christian, P.E. and Waterstram-Rich, K.M. (eds.) *Nuclear Medicine and PET/CT: Technology and Techniques*. 7th ed. St. Louis, MO: Elsevier/Mosby, 2011.

Cobbold, R.S.C. *Foundations of Biomedical Ultrasound*. New York, NY: Oxford University Press, USA, 2006.

Hsieh, J. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. 2nd ed. Bellingham, WA: SPIE Press, 2009.

Mahesh, M. *MDCT Physics: The Basics*, Philadelphia, PA: Lippincott Williams and Wilkins, 2009.

Signals and Systems

CHAPTER 2

2.1 Introduction

Signals and *systems* are two fundamental concepts for modeling medical imaging systems. Signals are mathematical functions of one or more independent variables, capable of modeling a variety of physical processes. Systems respond to signals by producing new signals. They are useful for modeling how physical processes (signals) change in natural environments and how devices, such as medical imaging instruments, create new signals (i.e., images). This chapter provides an introduction to the theory of signals and systems, with a focus on those tools required for modeling medical imaging systems.

Signals can be classified into three categories: (1) continuous, (2) discrete, and (3) mixed. A *continuous signal* is a function of independent variables that range over a continuum of values. For example, in computed tomography (CT), the distribution of x-ray attenuation in a cross-section within the body can be mathematically modeled using a function $f(x, y)$ of two independent real-valued variables x and y, which represent two spatial dimensions. Physical processes are usually modeled using continuous signals. Some medical images, such as those obtained on x-ray film, are also modeled as continuous signals.

A *discrete signal* is a function of independent variables that range over discrete values. These signals can be used to model physical processes that naturally correspond to discrete values. For example, the times of arrival of photons in a radioactive decay process form a discrete sequence of arrival times—that is, an intrinsically discrete signal. Discrete signals can also be used to represent continuous signals. For example, the x-ray distribution $f(x, y)$ described above might be represented in a computer as a function $f_d(m, n)$ of two independent discrete variables m and n.

Finally, a *mixed signal* is a function of some continuous and some discrete independent variables. Signals of this type are acquired by certain medical imaging instruments. For example, in CT, a signal $g(\ell, \theta_k)$, $k = 1, 2, \ldots,$ is recorded as a result of a parallel or fan beam of x-rays passing through an object. Variable ℓ is often modeled as continuous-valued, representing the distance of a particular ray from the center of the object, whereas θ_k is a discrete-valued variable, representing the beam's angle relative to a reference coordinate system (see Chapter 6).

Based on the previous signal classification, systems can also be organized into categories. For example, we may consider a *continuous-to-continuous system* that responds to a continuous signal by producing a continuous signal, or a *continuous-to-discrete system* that responds to a continuous signal by producing a discrete signal. For example, an analog-to-digital converter takes a continuous signal as input and produces a discrete signal as output through a process called *sampling* (see Chapter 3).

In this book, we consider one-dimensional (1-D), two-dimensional (2-D), and three-dimensional (3-D) signals and systems. In this chapter, we primarily develop 2-D continuous signals and systems; restriction to one dimension or extension to three dimensions is straightforward in most cases. In all cases, you can think of the *input signal* (to the imaging system) as coming from the patient and the *output signal* as coming from the imaging system. The challenge in medical imaging is to make the output signal a faithful representation of the input signal.

2.2 Signals

A *continuous signal* f is defined as a function

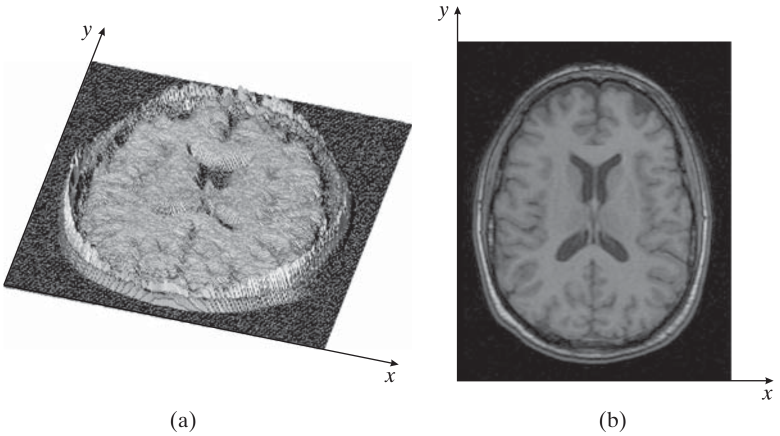
$$f(x,y), \qquad -\infty \le x,\ y \le \infty \tag{2.1}$$

of two independent real-valued variables x and y. This signal can be represented and visualized in two different ways. We may plot the signal as a function of the two independent variables x and y, or we may display it by assigning an intensity or brightness proportional to its value at (x, y). This is illustrated in Figure 2.1. In the first case, we call $f(x, y)$ a *function* or a *signal* and (x, y) a *point*. In the second case, we call f an *image* and (x, y) a *pixel*, a word that is derived from "**pic**ture **el**ement." The 3-D analog of the pixel is the *voxel*, from "**vo**lume **el**ement." This is illustrated in Figure 2.2. The representation in Figure 2.1(a) is useful in mathematical calculations, while that in Figure 2.1(b) is most appropriate for human observers. Accordingly, in this book, we focus primarily on two-dimensional pixel-based representations.

A number of special signals will frequently be used throughout this book. These include the point impulse, the comb and sampling functions, the line impulse, the rect and sinc functions, and exponential and sinusoidal signals. In addition, many signals will have properties that distinguish them from other signals and make their use in analysis easier. Separability and periodicity are two such common properties. We now study these special signals and properties.

2.2.1 Point Impulse

It is very useful in medical imaging to mathematically model the concept of a point source, which is used in the characterization of imaging system resolution. For example, if the source is very small, yet appears to be very large (blurred out)



(a) (b)

Figure 2.1
Two alternative signal visualizations: (a) a functional plot and (b) an image display.

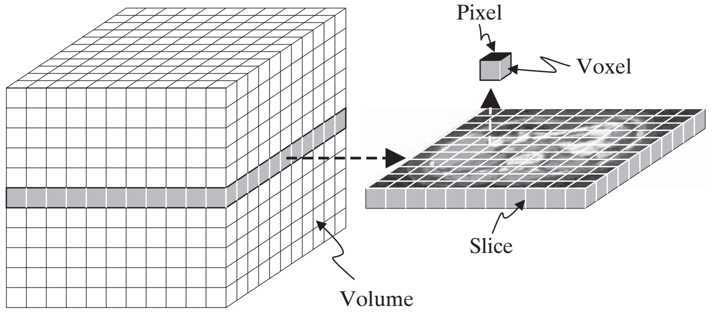


Figure 2.2
The representation of a 3-D object as a 2-D image. The 3-D object is represented by a collection of voxels; the 2-D image is made up of pixels. In this example, the image is a slice through the 3-D object.

in its radiological image, then we would say that the resolution of the system is poor. Precise definitions of resolution will be introduced in Chapter 3.

The concept of a point source in one dimension is known as the *1-D point impulse*, $\delta(x)$, which is defined by the following two properties:

$$\delta(x) = 0\,, \quad x \neq 0\,, \tag{2.2}$$

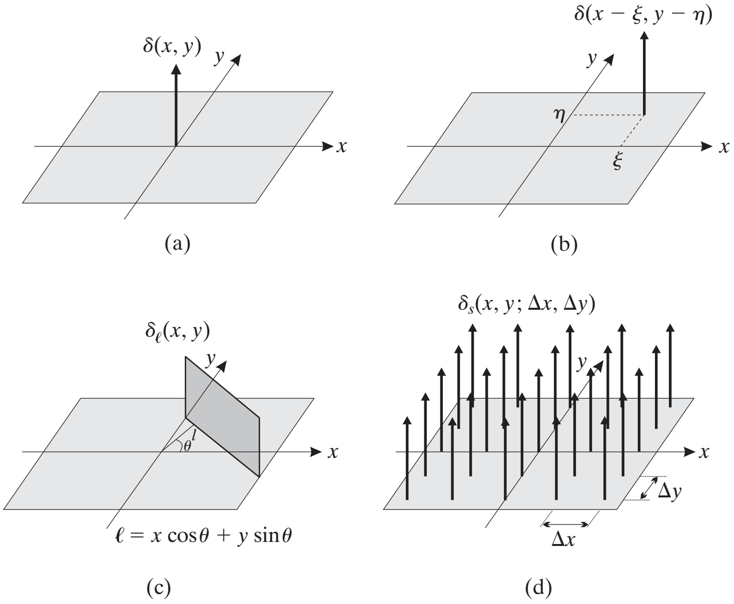
$$\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(0)\,. \tag{2.3}$$

In other references, the point impulse is also known as the *delta function*, the *Dirac function*, and the *impulse function*. The point impulse is not a function in the usual sense, but instead acts on other signals through integration, as we will see. It models the property of a point source by having infinitesimal width and unit area, which can be shown by taking $f(x) = 1$ in (2.3).

The *2-D point impulse*, $\delta(x, y)$ (also called the *2-D impulse function*, *2-D delta function*, or *2-D Dirac function*), is analogously characterized by

$$\delta(x, y) = 0\,, \quad (x, y) \neq (0, 0)\,, \tag{2.4}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(x, y)\,dx\,dy = f(0, 0)\,. \tag{2.5}$$



(a) (b)

(c) (d)

Figure 2.3
Signals derived from the point impulse: (a) point impulse $\delta(x, y)$, (b) shifted point impulse $\delta(x-\xi, y-\eta)$, (c) line impulse $\delta_\ell(x, y)$, and (d) sampling function $\delta_s(x, y; \Delta x, \Delta y)$ (note that the impulses extend to infinity in both x and y directions).

The 2-D point impulse (also not a function in the usual sense) models the property of a point source located at $(0, 0)$ by having infinitesimal width and unit volume. Figure 2.3(a) illustrates this concept.

In (2.5), the point impulse "picks off" the value of $f(x, y)$ at the location $(0, 0)$ by multiplying $f(x, y)$ with the point impulse followed by integration over all space. Suppose we shift the delta function to position (ξ, η)—that is, $\delta(x-\xi, y-\eta)$—as shown in Figure 2.3(b). Using a change of variables in (2.5), it can be shown that

$$f(\xi, \eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(x-\xi, y-\eta)\, dx\, dy\,. \tag{2.6}$$

This important property of the point impulse is known as the *sifting property*.

EXAMPLE 2.1

Consider the point impulse shifted to position (ξ, η)—that is, $\delta(x-\xi, y-\eta)$.

Question What is the nature of the function that is defined as the product of $f(x, y)$ with the shifted point impulse?

Answer Let $g(x, y) = f(x, y)\delta(x-\xi, y-\eta)$. Whenever $(x, y) \neq (\xi, \eta)$, the function $\delta(x-\xi, y-\eta)$ is zero, which means that

$$g(x, y) = 0\,, \qquad (x, y) \neq (\xi, \eta)\,.$$

When $(x, y) = (\xi, \eta)$, the shifted delta function is undefined (because its value is infinity), so $g(\xi, \eta)$ is also undefined. However, we know from (2.6) that the integral of $g(x, y)$—that is, its volume—is $f(\xi, \eta)$. Therefore, we can conclude that $g(x, y)$ is a point impulse function located at (ξ, η) having volume $f(\xi, \eta)$, or

$$g(x, y) = f(\xi, \eta)\delta(x-\xi, y-\eta)\,. \tag{2.7}$$

This result tells us that we can interpret the product of a function with a point impulse as another point impulse whose volume is equal to the value of the function at the location of the point impulse. ■

Two other properties of the point impulse will be used throughout the book. The *scaling property* of the point impulse is given by

$$\delta(ax, by) = \frac{1}{|ab|}\,\delta(x, y)\,. \tag{2.8}$$

Finally, we occasionally need to use the fact that the point impulse is an *even function*; that is,

$$\delta(-x, -y) = \delta(x, y)\,, \tag{2.9}$$

which follows from (2.8) by setting $a = b = -1$.

2.2.2 Line Impulse

When calibrating medical imaging equipment, it is sometimes easier to use a linelike rather than a pointlike object. For example, it may be easier to position a wire than a small bead to assess the resolution of a projection radiography system. For this reason, we would like a mathematical model for a line source.

Consider the set of points defined by

$$L(\ell, \theta) = \{(x, y) \,|\, x\cos\theta + y\sin\theta = \ell\}\,. \tag{2.10}$$

It can be shown that $L(\ell, \theta)$ is a line whose unit normal is oriented at an angle θ relative to the x-axis and is at distance ℓ from the origin in the direction of the unit normal. This geometry can be seen in Figure 2.3(c).

The *line impulse* $\delta_\ell(x, y)$ associated with line $L(\ell, \theta)$ is given by

$$\delta_\ell(x, y) = \delta(x\cos\theta + y\sin\theta - \ell)\,. \tag{2.11}$$

This signal, illustrated in Figure 2.3(c), is effectively a 1-D impulse function "spread out" on the line $L(\ell, \theta)$.

2.2.3 Comb and Sampling Functions

We now consider two functions that are critical to the way medical images are produced, manipulated, and stored in real-world systems. In such systems, medical images are not treated as continuous functions, but as discrete functions. We will have more to say about this in Chapter 3. Here, we introduce these two critical functions in relation to the other functions being introduced in this chapter.

We saw above that through the use of a shifted point impulse [within the double integral in (2.6)], we could "pick off" the value of a function at a single point. In nearly all medical imaging modalities, we need to pick off values not just at a single point but on a grid or matrix of points, a process that is called *sampling*. Collectively, this matrix of values represents the medical image that is observed. For example, a CT image is typically a 1,024 × 1,024 matrix of *CT numbers*, representing a physical parameter called the *linear attenuation coefficient* of a cross-section of the human body. A magnetic resonance (MR) image is typically a 256 × 256 matrix of values, representing one of several possible nuclear magnetic resonance (NMR) properties of tissues.

As a first step toward characterizing sampling mathematically, we introduce the *comb function*, given by

$$\operatorname{comb}(x) = \sum_{n=-\infty}^{\infty} \delta(x-n)\,. \tag{2.12}$$

The comb function is also known as the *shah function*. It is called the comb function because the set of shifted point impulses constituting it resembles the teeth of a comb. In 2-D, the comb function is given by

$$\operatorname{comb}(x,y) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(x-m, y-n)\,. \tag{2.13}$$

It is useful in describing signal sampling (see Section 3.6) to space the point impulses in the comb function by amounts Δx in the x direction, and Δy in the y direction. This yields the *sampling function* $\delta_s(x, y;\, \Delta x, \Delta y)$, defined by

$$\delta_s(x, y;\, \Delta x, \Delta y) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(x - m\Delta x, y - n\Delta y). \tag{2.14}$$

It can be shown using the scaling property (2.8) of the point impulse, that the sampling function is related to the comb function as follows:

$$\delta_s(x, y;\, \Delta x, \Delta y) = \frac{1}{\Delta x \Delta y} \operatorname{comb}\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right)\,. \tag{2.15}$$

The sampling function is a sequence of point impulses located at points $(m\Delta x, n\Delta y)$, $-\infty < m, n < \infty$, in the plane, as illustrated in Figure 2.3(d). The concept of the sampling function is critical for understanding *discretization*—the process of going from a continuous signal to a discrete signal.

2.2.4 Rect and Sinc Functions

Two signals that are frequently used in the study of medical imaging systems are the rect and sinc functions. The *rect function* is given by

$$\operatorname{rect}(x,y) = \begin{cases} 1, & \text{for } |x| < \dfrac{1}{2} \text{ and } |y| < \dfrac{1}{2} \\[2ex] 0, & \text{for } |x| > \dfrac{1}{2} \text{ or } |y| > \dfrac{1}{2} \end{cases}, \tag{2.16}$$

where the value at $|x| = 1/2, |y| \leq 1/2$ or $|x| \leq 1/2, |y| = 1/2$ is immaterial. (If this value is desired, we usually set it to 1/2.) The rect function is a finite energy signal, with unit total energy, that models signal concentration over a unit square centered around point $(0, 0)$. We can use the product

$$f(x, y)\,\mathrm{rect}\left(\frac{x-\xi}{X}, \frac{y-\eta}{Y}\right) \tag{2.17}$$

to select that part of signal $f(x, y)$ centered at a point (ξ, η) of the plane with width X and height Y, and set the rest to zero.

The rect function can be written as the product of two 1-D rect functions,

$$\mathrm{rect}(x, y) = \mathrm{rect}(x)\,\mathrm{rect}(y)\,, \tag{2.18}$$

where

$$\mathrm{rect}(x) = \begin{cases} 1, & \text{for } |x| < \dfrac{1}{2} \\ 0, & \text{for } |x| > \dfrac{1}{2} \end{cases}\,. \tag{2.19}$$

The 1-D rect function is plotted in Figure 2.4(a).

The *sinc function* is given by

$$\mathrm{sinc}(x, y) = \begin{cases} 1, & \text{for } x = y = 0 \\ \dfrac{\sin(\pi x)\sin(\pi y)}{\pi^2 xy}, & \text{otherwise.} \end{cases} \tag{2.20}$$

The sinc function is a finite energy signal with unit total energy. Its maximum value equals 1, and is attained at point $(0, 0)$. The sinc function can be written as the product of two 1-D sinc functions,

$$\mathrm{sinc}(x, y) = \mathrm{sinc}(x)\,\mathrm{sinc}(y), \tag{2.21}$$

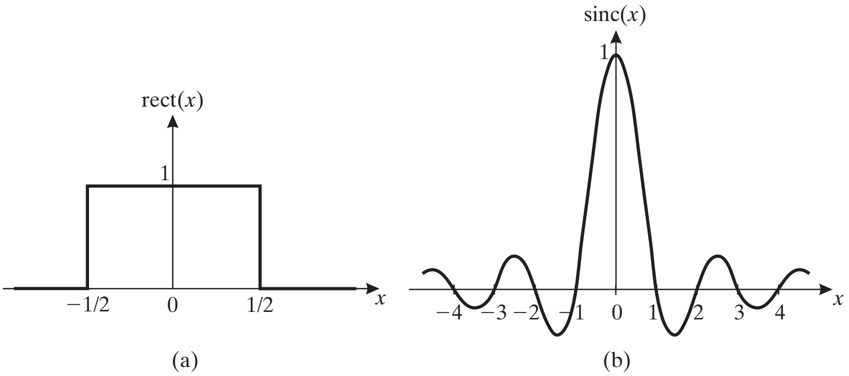


Figure 2.4
(a) The 1-D rect function rect(x), and (b) the 1-D sinc function sinc(x).

where

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x} . \tag{2.22}$$

The 1-D sinc function is plotted in Figure 2.4(b). This function consists of a main lobe and side lobes, which are smaller in height and eventually diminish to zero for large values of x. It alternates between positive and negative values by passing through zero at points $x = \pm 1, \pm 2, \ldots$.

The rect and sinc functions are useful in understanding the influence of the finite width of pixels in portraying an originally continuous signal as a discrete image.

2.2.5 Exponential and Sinusoidal Signals

Another continuous signal is the *complex exponential signal*, given by

$$e(x, y) = e^{j2\pi(u_0 x + v_0 y)} , \tag{2.23}$$

where u_0 and v_0 are two real-valued parameters and $j^2 = -1$. These parameters are usually referred to as *fundamental frequencies*, and have units that are the inverse of the units of x and y. For example, if x and y are given in mm, then u_0 and v_0 have the unit mm^{-1}.

The complex exponential signal can be decomposed into real and imaginary parts using two *sinusoidal signals*:

$$s(x, y) = \sin[2\pi(u_0 x + v_0 y)] \qquad \text{and} \qquad c(x, y) = \cos[2\pi(u_0 x + v_0 y)] . \tag{2.24}$$

Indeed,

$$\begin{aligned} e(x, y) &= e^{j2\pi(u_0 x + v_0 y)} \\ &= \cos[2\pi(u_0 x + v_0 y)] + j \sin[2\pi(u_0 x + v_0 y)] \\ &= c(x, y) + js(x, y) . \end{aligned} \tag{2.25}$$

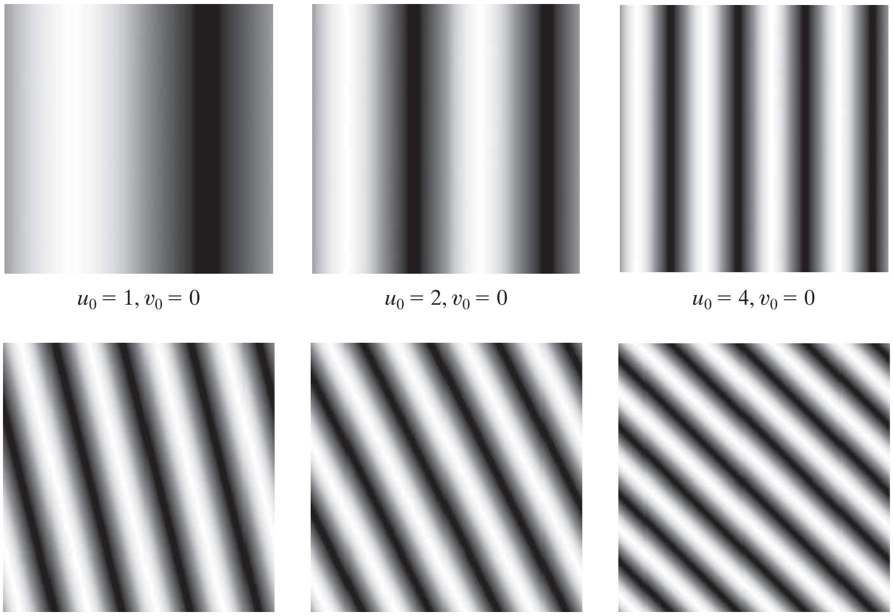
On the other hand, a sinusoidal signal can be written in terms of two complex exponential signals, since

$$s(x, y) = \sin[2\pi(u_0 x + v_0 y)] = \frac{1}{2j} e^{j2\pi(u_0 x + v_0 y)} - \frac{1}{2j} e^{-j2\pi(u_0 x + v_0 y)} , \tag{2.26}$$

$$c(x, y) = \cos[2\pi(u_0 x + v_0 y)] = \frac{1}{2} e^{j2\pi(u_0 x + v_0 y)} + \frac{1}{2} e^{-j2\pi(u_0 x + v_0 y)} . \tag{2.27}$$

The fundamental frequencies u_0 and v_0 affect the oscillating behavior of the sinusoidal signals in the x and y directions, respectively. For example, small values of u_0 result in slow oscillations in the x direction, whereas large values result in fast oscillations in the x direction. This is illustrated in Figure 2.5.

The concepts of exponential signals are particularly useful in Fourier analysis, which is used in image reconstruction and in understanding MRI.



$u_0 = 1, v_0 = 0$ $u_0 = 2, v_0 = 0$ $u_0 = 4, v_0 = 0$

$u_0 = 4, v_0 = 1$ $u_0 = 4, v_0 = 2$ $u_0 = 4, v_0 = 4$

Figure 2.5
Six instances of the sinusoidal signal $s(x, y) = \sin[2\pi(u_0x + v_0y)]$, $0 \le x, y \le 1$, for various values of the fundamental frequencies u_0, v_0. Notice that small values result in slow oscillations in the corresponding direction, whereas large values result in fast oscillations.

2.2.6 Separable Signals

Separable signals form another class of continuous signals. A signal $f(x, y)$ is a *separable signal* if there exist two 1-D signals $f_1(x)$ and $f_2(y)$ such that

$$f(x, y) = f_1(x)f_2(y) \,. \tag{2.28}$$

A 2-D separable signal that is a function of two independent variables x and y can be separated into a product of two 1-D signals, one of which is only a function of x and the other only of y. Notice that the point impulse is a separable signal, since $\delta(x, y) = \delta(x)\delta(y)$. From (2.18) and (2.21), the rect and sinc functions are separable signals as well.

Separable signals are limited in the sense that they can only model signal variations independently in the x and y directions. However, there are instances in which use of separable signals is appropriate. Of major importance is the fact that operating on separable signals is much simpler than operating on purely 2-D signals, since, for separable signals 2-D operations reduce to simpler consecutive 1-D operations.

2.2.7 Periodic Signals

A signal $f(x, y)$ is a *periodic signal* if there are two positive constants X and Y such that

$$f(x, y) = f(x + X, y) = f(x, y + Y) \,, \tag{2.29}$$

where X and Y are known as the signal *periods* in the x and y direction, respectively. From (2.29), we need to know a periodic signal only in the rectangular window $0 \leq x < X$, $0 \leq y < Y$. The sampling function $\delta_s(x, y; \Delta x, \Delta y)$ [(2.14)] is a periodic signal with periods $X = \Delta x$ and $Y = \Delta y$, whereas exponential and sinusoidal signals [(2.23) and (2.24)] are periodic with periods $X = 1/u_0$ and $Y = 1/v_0$.

2.3 Systems

A *continuous-to-continuous* (or simply *continuous*) *system* is defined as a transformation $\mathcal{S}$ of an *input* continuous signal $f(x, y)$ to an *output* continuous signal $g(x, y)$. This is illustrated in Figure 2.6, which depicts the response of a system $\mathcal{S}$ to a point impulse. In general,

$$g(x, y) = \mathcal{S}[f(x, y)], \tag{2.30}$$

which is known as the *input-output equation* of system $\mathcal{S}$. With a slight abuse of notation, we denote by $\mathcal{S}[f(x, y)]$ the value, at point (x, y), of the signal obtained by applying the transformation $\mathcal{S}$ on the *entire* signal $f(x, y)$. The importance of (2.30) is that it implies that we can predict the output of an imaging system if we know the input f and the characteristics of the imaging system $\mathcal{S}$.

Equation (2.30) is too general to be useful in practice. In a particular application, there might be many choices for $\mathcal{S}$, and the question is which one is more appropriate. For a choice of $\mathcal{S}$ to be useful, it should be mathematically tractable. It is therefore necessary to limit our choices, by considering systems obtained using transformations $\mathcal{S}$ that satisfy certain simplifying assumptions. Of course, any choice of $\mathcal{S}$ should accurately portray the actual situation.

2.3.1 Linear Systems

A simplifying assumption is that of linearity. A system $\mathcal{S}$ is a *linear system* if, when the input consists of a weighted summation of several signals, the output will also be a weighted summation of the responses of the system to each individual input signal. More precisely, for any collection $\{f_k(x, y), k = 1, 2, \ldots, K\}$ of input signals, and for any collection $\{w_k, k = 1, 2, \ldots, K\}$ of weights, we have

$$\mathcal{S}\left[\sum_{k=1}^{K} w_k f_k(x, y)\right] = \sum_{k=1}^{K} w_k \mathcal{S}\left[f_k(x, y)\right]. \tag{2.31}$$

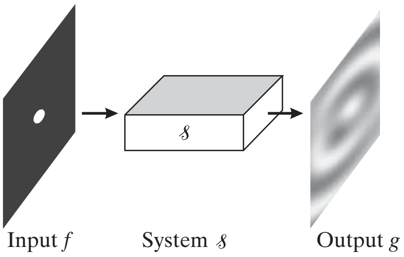


Figure 2.6
The response of a continuous system to a point impulse.

The linearity assumption leads to a class of relatively simple systems, which are tractable and enjoy a number of useful properties. Linearity is thus a critical assumption used in this book.

How reasonable is the assumption of linearity in practice? Are linear systems good mathematical approximations of the physical phenomena underlying medical imaging? It turns out that many medical imaging systems are approximately linear. In practice, this means that the image of an ensemble of signals is (nearly) identical to the sum of separate images of each signal, which is considered next.

2.3.2 Impulse Response

We now investigate how a linear system responds to an arbitrary input signal $f(x, y)$ and demonstrate the fact that linearity leads to tractable systems. However, before we pursue this investigation, it will be beneficial to first consider the output of a system $\mathcal{S}$ to a point impulse. Let us denote by $h(x, y; \xi, \eta)$ the output of system $\mathcal{S}$ to input $\delta_{\xi\eta}(x, y) = \delta(x - \xi, y - \eta)$ [that is, a point impulse located at (ξ, η)]. In this case,

$$h(x, y; \xi, \eta) = \mathcal{S}[\delta_{\xi\eta}(x, y)] \,. \tag{2.32}$$

In general, this output depends on four independent variables; it is therefore a four-dimensional signal. The response $h(x, y; \xi, \eta)$ is known as the *point spread function* (PSF) of system $\mathcal{S}$, or, equivalently, the *impulse response function*.

Let us now assume that $\mathcal{S}$ is a linear system. If $f(x, y)$ is an arbitrary input to $\mathcal{S}$, then by starting with (2.30) and applying (2.6), (2.31), and (2.32), the output $g(x, y)$ will be given by

$$g(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) h(x, y; \xi, \eta) \, d\xi \, d\eta \,. \tag{2.33}$$

The integral in (2.33) is known as the *superposition integral*. This equation shows that the PSF $h(x, y; \xi, \eta)$ uniquely characterizes a linear system; that is, we only need to know the PSF of a linear system in order to calculate its output in response to a given input.

In fact, (2.33) facilitates computation of the output $g(x, y)$ from an arbitrary but known input $f(x, y)$. If the PSF $h(x, y; \xi, \eta)$ is known for every (x, y, ξ, η), then this computation reduces to calculating the double integral in (2.33). In practice, however, knowing the PSF for every (x, y, ξ, η) is a formidable task, since this either requires that $h(x, y; \xi, \eta)$ is known analytically, which is true only in limited circumstances of usually no practical interest, or that $h(x, y; \xi, \eta)$ is experimentally measured and stored in numerical form, which is usually not practical due to the four-dimensional nature of $h(x, y; \xi, \eta)$. Therefore, additional simplification is needed at this point, which may be achieved by imposing additional conditions on system $\mathcal{S}$.

2.3.3 Shift Invariance

An additional simplifying assumption is *shift invariance*. A system $\mathcal{S}$ is shift-invariant if an arbitrary translation of the input results in an identical translation in the output. In mathematical terms, if

$$f_{x_0 y_0}(x, y) = f(x - x_0, y - y_0) \tag{2.34}$$

is a translated version of an input signal $f(x, y)$ shifted at a point (x_0, y_0), then

$$g(x - x_0, y - y_0) = \mathcal{S}[f_{x_0 y_0}(x, y)]\,, \tag{2.35}$$

where $g(x, y)$ is given by (2.30). Therefore, the response of a shift-invariant system to a translated input equals the response of the system to the actual input translated by the same amount.

Shift invariance does not require or imply linearity: a system may be shift-invariant but not linear and vice versa. However, if a linear system $\mathcal{S}$ is shift-invariant and

$$h(x, y) = \mathcal{S}[\delta(x, y)]\,, \tag{2.36}$$

then

$$\mathcal{S}[\delta_{\xi\eta}(x, y)] = h(x - \xi, y - \eta)\,. \tag{2.37}$$

In this case, the PSF is a 2-D signal. Therefore, by imposing shift invariance on a linear system, we are able to reduce the dimensionality of the PSF by a factor of two.

In practice, this means that the PSF is the same throughout the field-of-view of a shift-invariant imaging system. If we measure the PSF at one position, we can assume the same PSF at all other positions within the imaging field. Such a measurement is accomplished by using a (very) small point object (to mimic a point impulse) and (2.36), which indicates that the image of this point object *is* the PSF.

A system $\mathcal{S}$ that is both linear and shift-invariant is known as a *linear shift-invariant* (LSI) *system*. If $\mathcal{S}$ is an LSI system, then [see (2.32), (2.33), and (2.37)]

$$g(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) h(x - \xi, y - \eta)\, d\xi\, d\eta\,, \tag{2.38}$$

or, simply,

$$g(x, y) = h(x, y) * f(x, y)\,. \tag{2.39}$$

Equations (2.38) and (2.39) are known as the *convolution integral* and the *convolution equation*, respectively. Convolution plays a fundamental role in the theory of signals and systems. In this book, we almost exclusively deal with LSI systems. Treating imaging systems as LSI systems significantly simplifies our analysis of these systems, and in most cases is accurate enough for practical use.

EXAMPLE 2.2

Consider a continuous system with input-output equation

$$g(x, y) = 2f(x, y)\,. \tag{2.40}$$

Question Is this system linear and shift-invariant?

Answer If $g'(x,y)$ is the response of the system to input $\sum_{k=1}^{K} w_k f_k(x,y)$, then

$$\begin{aligned} g'(x,y) &= 2\left(\sum_{k=1}^{K} w_k f_k(x,y)\right) \\ &= \sum_{k=1}^{K} w_k 2 f_k(x,y) \\ &= \sum_{k=1}^{K} w_k g_k(x,y)\,, \end{aligned} \tag{2.41}$$

where $g_k(x,y)$ is the response of the system to input $f_k(x,y)$. Therefore, the system is linear. On the other hand, if $g'(x,y)$ is the response of the system to input $f(x-x_0, y-y_0)$, then

$$g'(x,y) = 2f(x-x_0, y-y_0) = g(x-x_0, y-y_0)\,, \tag{2.42}$$

and the system is also shift-invariant. ■

EXAMPLE 2.3

Consider a continuous system with input-output equation

$$g(x,y) = xyf(x,y)\,. \tag{2.43}$$

Question Is this system linear and shift-invariant?

Answer If $g'(x,y)$ is the response of the system to input $\sum_{k=1}^{K} w_k f_k(x,y)$, then

$$\begin{aligned} g'(x,y) &= xy\left(\sum_{k=1}^{K} w_k f_k(x,y)\right) \\ &= \sum_{k=1}^{K} w_k xy f_k(x,y) \\ &= \sum_{k=1}^{K} w_k g_k(x,y)\,, \end{aligned} \tag{2.44}$$

where $g_k(x,y)$ is the response of the system to input $f_k(x,y)$. Therefore, the system is linear. On the other hand, if $g'(x,y)$ is the response of the system to input $f(x-x_0, y-y_0)$ where $x_0 \neq 0$ and $y_0 \neq 0$, then

$$\begin{aligned} g'(x,y) &= xyf(x-x_0, y-y_0) \\ &\neq (x-x_0)(y-y_0)f(x-x_0, y-y_0)\,. \end{aligned}$$

Thus,

$$g'(x,y) \neq g(x-x_0, y-y_0)\,, \tag{2.45}$$

and the system is not shift-invariant. ■

2.3.4 Connections of LSI Systems

LSI systems can be stand-alone or connected with other LSI systems. Two types of connections are usually considered: (1) *cascade* or *serial connections* and (2) *parallel connections*.

Figure 2.7(a) illustrates two equivalent cascade connections of two LSI systems with PSFs $h_1(x,y)$ and $h_2(x,y)$, as well as the equivalent "single" LSI system. In this case, the following is true:

$$\begin{aligned} g(x,y) &= h_2(x,y) * [h_1(x,y) * f(x,y)] \\ &= h_1(x,y) * [h_2(x,y) * f(x,y)] \\ &= [h_1(x,y) * h_2(x,y)] * f(x,y)\,, \end{aligned} \tag{2.46}$$

and

$$h_1(x,y) * h_2(x,y) = h_2(x,y) * h_1(x,y)\,. \tag{2.47}$$

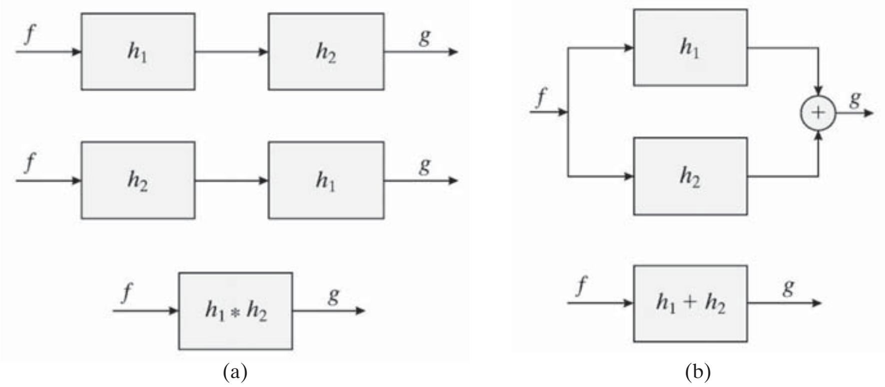
Equations (2.46) and (2.47) are two properties of convolution, known as *associativity* and *commutativity*, respectively. From the commutativity property, we see that $g(x,y) = h(x,y) * f(x,y) = f(x,y) * h(x,y)$, which leads to

$$\begin{aligned} g(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi,\eta)h(x-\xi,y-\eta)\,d\xi\,d\eta \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\xi,\eta)f(x-\xi,y-\eta)\,d\xi\,d\eta\,. \end{aligned} \tag{2.48}$$

Figure 2.7(b) illustrates a parallel connection of two LSI systems with PSFs $h_1(x,y)$ and $h_2(x,y)$, as well as the resulting equivalent single LSI system. In this case,

$$\begin{aligned} g(x,y) &= h_1(x,y) * f(x,y) + h_2(x,y) * f(x,y) \\ &= [h_1(x,y) + h_2(x,y)] * f(x,y)\,, \end{aligned} \tag{2.49}$$

which is another property of convolution, known as *distributivity*.



(a) (b)

Figure 2.7
(a) Cascade connection of two LSI systems.
(b) Parallel connection of two LSI systems.

The properties of associativity, commutativity, and distributivity associated with LSI systems significantly simplify our analysis. (We imagine you are getting the idea why we like treating medical imaging systems as LSI so much!)

EXAMPLE 2.4

Consider two LSI systems connected in cascade, with Gaussian PSFs of the form

$$h_1(x,y) = \frac{1}{2\pi\sigma_1^2} e^{-(x^2+y^2)/2\sigma_1^2} \qquad \text{and} \qquad h_2(x,y) = \frac{1}{2\pi\sigma_2^2} e^{-(x^2+y^2)/2\sigma_2^2}\,, \tag{2.50}$$

where σ_1 and σ_2 are two positive constants.

Question What is the PSF of this system?

Answer This connection is equivalent to a single LSI system with impulse response $h(x,y)$, given by

$$\begin{aligned} h(x,y) &= h_1(x,y) * h_2(x,y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\xi,\eta)h_1(x-\xi,y-\eta)\,d\xi\,d\eta \\ &= \frac{1}{4\pi^2\sigma_1^2\sigma_2^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(\xi^2+\eta^2)/2\sigma_2^2} e^{-[(x-\xi)^2+(y-\eta)^2]/2\sigma_1^2}\,d\xi\,d\eta \\ &= \frac{1}{4\pi^2\sigma_1^2\sigma_2^2}\int_{-\infty}^{\infty} e^{-\xi^2/2\sigma_2^2-(x-\xi)^2/2\sigma_1^2}\,d\xi \\ &\qquad\qquad \int_{-\infty}^{\infty} e^{-\eta^2/2\sigma_2^2-(y-\eta)^2/2\sigma_1^2}\,d\eta\,. \end{aligned} \tag{2.51}$$

However,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\xi^2/2\sigma_2^2-(x-\xi)^2/2\sigma_1^2}\,d\xi &= e^{-x^2/2(\sigma_1^2+\sigma_2^2)}\int_{-\infty}^{\infty} e^{-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left\{\xi-[\sigma_2^2/(\sigma_1^2+\sigma_2^2)]x\right\}^2}\,d\xi \\ &= e^{-x^2/2(\sigma_1^2+\sigma_2^2)}\int_{-\infty}^{\infty} e^{-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\tau^2}\,d\tau \\ &= \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\,e^{-x^2/2(\sigma_1^2+\sigma_2^2)}\,, \end{aligned} \tag{2.52}$$

in which case

$$h(x,y) = \frac{1}{2\pi(\sigma_1^2+\sigma_2^2)}\exp\left\{\frac{-(x^2+y^2)}{2(\sigma_1^2+\sigma_2^2)}\right\}\,. \tag{2.53}$$

To obtain the second equality in (2.52), we have substituted the variable $\xi - [\sigma_2^2/(\sigma_1^2+\sigma_2^2)]\,x$ with τ; whereas, to obtain the third equality, we have used the fact that

$$\int_{-\infty}^{\infty} e^{-a^2\tau^2}\,d\tau = \frac{\sqrt{\pi}}{a}\,, \qquad \text{for } a \neq 0\,. \tag{2.54}$$

The resulting PSF is Gaussian as well. ■

2.3.5 Separable Systems

Separable systems form an important class of LSI systems. As with separable signals, a 2-D LSI system with PSF $h(x, y)$ is a *separable system* if there are two 1-D systems with PSFs $h_1(x)$ and $h_2(y)$, such that

$$h(x, y) = h_1(x)h_2(y)\,. \tag{2.55}$$

A 2-D separable system consists of a cascade of two 1-D systems, one applied in the x direction and the other one in the y direction. Therefore, for a 2-D separable system, the 2-D convolution integral (2.38) can be calculated using two simpler 1-D convolution integrals, as follows:

- Compute $w(x, y) = \int_{-\infty}^{\infty} f(\xi, y)h_1(x - \xi)\, d\xi$, for every y.
- Compute $g(x, y) = \int_{-\infty}^{\infty} w(x, \eta)h_2(y - \eta)\, d\eta$, for every x.

Indeed,

$$\begin{aligned}\int_{-\infty}^{\infty} w(x, \eta)h_2(y - \eta)\, d\eta &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(\xi, \eta)h_1(x - \xi)d\xi \right] h_2(y - \eta)\, d\eta \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta)h_1(x - \xi)h_2(y - \eta)\, d\xi\, d\eta \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta)h(x - \xi, y - \eta)\, d\xi\, d\eta \quad \text{[from (2.55)]} \\ &= h(x, y) * f(x, y) = g(x, y)\,. \end{aligned} \tag{2.56}$$

The first step keeps y fixed, treats image $f(x, y)$ as a 1-D signal, and convolves it with the PSF $h_1(x)$ to obtain $w(x, y)$ for every y. The second step keeps x fixed, treats image $w(x, y)$ as a 1-D signal, and convolves it with the PSF $h_2(y)$. This is illustrated in Figure 2.8, which depicts the result of convolving an image with a

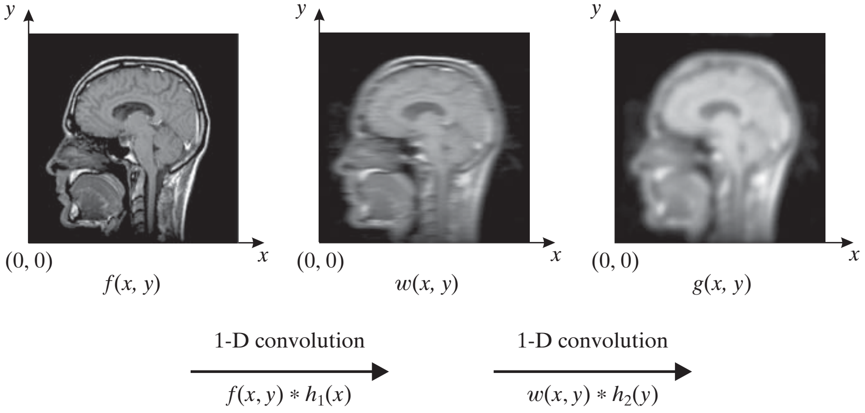


Figure 2.8
Calculation of the output of a 2-D separable system by using two 1-D steps in cascade.

2-D Gaussian PSF of the form

$$h(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}\,, \tag{2.57}$$

with $\sigma = 4$, by using two 1-D steps in cascade. This PSF is separable, with

$$h_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \qquad \text{and} \qquad h_2(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2}\,. \tag{2.58}$$

In practice, it is often faster (and easier) to execute two consecutive 1-D operations than a single 2-D operation, so the concept of separability is quite useful (when applicable).

2.3.6 Stable Systems

Stability is an important property of medical imaging systems. Informally, a medical imaging system is *stable* if small inputs lead to outputs that do not diverge. Although there are many ways to characterize system stability, we only consider bounded-input bounded-output (BIBO) stability here. A system is a *bounded-input bounded-output* (BIBO) *stable system* if, when the input is a *bounded signal*—that is, when

$$\left|f(x, y)\right| \le B < \infty\,, \qquad \text{for every } (x, y)\,, \tag{2.59}$$

for some finite B—there exists a finite B' such that

$$\left|g(x, y)\right| = \left|h(x, y) * f(x, y)\right| \le B' < \infty\,, \qquad \text{for every } (x, y)\,, \tag{2.60}$$

in which case the output will also be a bounded signal. It can be shown that an LSI system is a BIBO stable system if and only if its PSF is absolutely integrable, in which case

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|h(x, y)\right|\, dx\, dy < \infty\,. \tag{2.61}$$

In practice, it is very important to design medical imaging systems that are stable. The output of an unstable system may grow out of bound, even if the input is small. This is an undesirable problem that strongly affects the accuracy and diagnostic utility of a medical imaging system. In the physical world, most systems are stable. However, when we design a medical imaging system, we may occasionally be tempted to model it using a PSF that does not satisfy (2.61), in which case we need to find alternative stable PSFs. This will be encountered in Chapter 6, for example, where computed tomography is discussed.

2.4 The Fourier Transform

Although the convolution equation provides a means of relating the output of an LSI system to its input, there exists an alternative (and equivalent) way of viewing this relationship: the Fourier transform. The Fourier transform provides a different perspective on how signals and systems interact, and it leads to alternative tools for system analysis and implementation.

Recall that the convolution equation [(2.39)] was obtained by decomposing a signal into point impulses [see (2.6), (2.32), and (2.33)]. An alternative way to decompose a signal is in terms of complex exponential signals. It can be shown that if

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(ux+vy)}\,dx\,dy\,, \tag{2.62}$$

then

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{j2\pi(ux+vy)}\,du\,dv\,. \tag{2.63}$$

The double integral in (2.62) [as well as the function $F(u,v)$] is known as the (2-D) *Fourier transform* of $f(x,y)$, whereas the double integral in (2.63) is known as the (2-D) *inverse Fourier transform*. We use the notation

$$F(u,v) = \mathcal{F}_{\mathrm{2D}}(f)(u,v) \quad \text{and} \quad f(x,y) = \mathcal{F}_{\mathrm{2D}}^{-1}(F)(x,y) \tag{2.64}$$

to denote the Fourier transform and its inverse, respectively. The subscript "2D" is used to indicate that the Fourier transform is 2-D.

Before we can use the Fourier transform and its inverse, we need to make sure that the integrals in (2.62) and (2.63) exist. A sufficient condition for the existence of such integrals is that the signal $f(x,y)$ is continuous, or has a finite number of discontinuities, and that it is absolutely integrable. These conditions are almost always satisfied in practice.

Equation (2.63) shows that the Fourier transform produces a decomposition of a signal $f(x,y)$ into complex exponentials $e^{j2\pi(ux+vy)}$ with strength $F(u,v)$. The variables u and v are known as the x and y (spatial) *frequencies* of signal $f(x,y)$, respectively. The Fourier transform $F(u,v)$ is frequently referred to as the *spectrum* of $f(x,y)$. Since

$$e^{j2\pi(u_0x+v_0y)} = \cos[2\pi(u_0x+v_0y)] + j\sin[2\pi(u_0x+v_0y)]\,, \tag{2.65}$$

the Fourier transform provides information on the sinusoidal composition of a signal $f(x,y)$ at different frequencies. Notice the relationship between the Fourier transform and the exponential and sinusoidal signals discussed in Section 2.2.5.

In practice, the Fourier transform allows us to *separately* consider the action of an LSI system on each sinusoidal frequency. This general characterization of an LSI system then carries over to the action on arbitrary signals by considering the signals' Fourier transforms.

In general, the Fourier transform is a complex-valued signal, even if $f(x,y)$ is real-valued. It is quite common then to separately consider the Fourier transform's *magnitude* and *phase*, given by

$$|F(u,v)| = \sqrt{F_R^2(u,v) + F_I^2(u,v)}\,, \tag{2.66}$$

and

$$\angle F(u,v) = \tan^{-1}\left(\frac{F_I(u,v)}{F_R(u,v)}\right), \tag{2.67}$$

where $F_R(u,v)$ and $F_I(u,v)$ are the real and imaginary parts of $F(u,v)$, respectively [i.e., $F(u,v) = F_R(u,v) + jF_I(u,v)$], in which case

$$F(u,v) = |F(u,v)|\, e^{j\angle F(u,v)}\,. \tag{2.68}$$

The magnitude $|F(u,v)|$ and phase $\angle F(u,v)$ are usually referred to as the *magnitude spectrum* and the *phase spectrum*, respectively. Moreover, the square $|F(u,v)|^2$ of the magnitude spectrum is usually referred to as the *power spectrum*. Notice that both the magnitude and phase spectra are required to uniquely determine $f(x,y)$.

We now provide a few examples of signals and their corresponding Fourier transforms. Some basic Fourier transform pairs are summarized in Table 2.1.

EXAMPLE 2.5

Consider the point impulse $\delta(x,y)$.

Question What is its Fourier transform?

Answer From (2.62), we have

$$\begin{aligned}\mathcal{F}_{2D}(\delta)(u,v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x,y)e^{-j2\pi(ux+vy)}\,dx\,dy\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x,y)e^{-j2\pi(u0+v0)}\,dx\,dy \quad \text{[from (2.7)]}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x,y)\,dx\,dy = 1 \quad \text{[from (2.5)]}\,.\end{aligned} \tag{2.69}$$

In this case, the magnitude spectrum is 1, whereas the phase spectrum is 0. Therefore, the Fourier transform of a point impulse is constant over all frequencies, with strength equal to 1. ■

TABLE 2.1

Basic Fourier Transform Pairs

Signal	Fourier Transform
1	$\delta(u,v)$
$\delta(x,y)$	1
$\delta(x-x_0, y-y_0)$	$e^{-j2\pi(ux_0+vy_0)}$
$\delta_s(x,y;\Delta x,\Delta y)$	$\mathrm{comb}(u\Delta x, v\Delta y)$
$e^{j2\pi(u_0x+v_0y)}$	$\delta(u-u_0, v-v_0)$
$\sin[2\pi(u_0x+v_0y)]$	$\frac{1}{2j}[\delta(u-u_0,v-v_0)-\delta(u+u_0,v+v_0)]$
$\cos[2\pi(u_0x+v_0y)]$	$\frac{1}{2}[\delta(u-u_0,v-v_0)+\delta(u+u_0,v+v_0)]$
$\mathrm{rect}(x,y)$	$\mathrm{sinc}(u,v)$
$\mathrm{sinc}(x,y)$	$\mathrm{rect}(u,v)$
$\mathrm{comb}(x,y)$	$\mathrm{comb}(u,v)$
$e^{-\pi(x^2+y^2)}$	$e^{-\pi(u^2+v^2)}$

EXAMPLE 2.6

Consider the complex exponential signal

$$f(x,y) = e^{j2\pi(u_0x+v_0y)}\,. \tag{2.70}$$

Question What is its Fourier transform?

Answer From (2.62), we have

$$\begin{aligned}\mathcal{F}_{2D}(f)(u,v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(ux+vy)}\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j2\pi(u_0x+v_0y)}e^{-j2\pi(ux+vy)}\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi[(u-u_0)x+(v-v_0)y]}\,dx\,dy \\ &= \delta(u-u_0, v-v_0)\,, \end{aligned} \tag{2.71}$$

where the last step is based on the fact that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi(ux+vy)}\,dx\,dy = \delta(u,v)\,, \tag{2.72}$$

which relates the complex exponential to the point impulse. In this case, the magnitude spectrum is a point impulse located at frequency (u_0, v_0), whereas the phase spectrum is 0. When $u_0 = v_0 = 0$, the complex exponential becomes the unit-valued signal $f(x,y) = 1$, for every (x,y), in which case (2.71) implies that the Fourier transform of a unit-valued signal is a point impulse at frequency $(0,0)$. ■

A point impulse, which has an extremely concentrated profile in space, results in a uniform frequency content (i.e., a magnitude spectrum with constant value). On the other hand, a constant signal, which does not vary in space, has a spectrum that is all concentrated at frequency $(0,0)$. This extreme behavior reveals a more general property. Slow signal variation in space produces a spectral content that is primarily concentrated at low frequencies, whereas fast signal variation results in spectral content at high frequencies. This is illustrated in Figure 2.9, where three images with decreasing spatial variation (from left to right) produce spectra with decreasing high frequency content. Notice that in order to reduce the dynamic range of values in the magnitude spectrum $|F(u,v)|$ for display purposes, we calculate and display the 2-D function $\log(1 + |F(u,v)|)$ instead.

The Fourier transform of a 1-D signal and its inverse can be easily obtained from (2.62) and (2.63) by simply eliminating one of the two dimensions. In fact, if $f(x)$, $-\infty \le x \le \infty$, is a 1-D signal, then

$$F(u) = \mathcal{F}_{1\mathrm{D}}(f)(u) = \int_{-\infty}^{\infty} f(x)e^{-j2\pi ux}\,dx\,, \tag{2.73}$$

and

$$f(x) = \mathcal{F}_{1\mathrm{D}}^{-1}(F)(x) = \int_{-\infty}^{\infty} F(u)e^{j2\pi ux}\,du\,, \tag{2.74}$$

where the subscript "1D" denotes that the Fourier transform is one-dimensional.

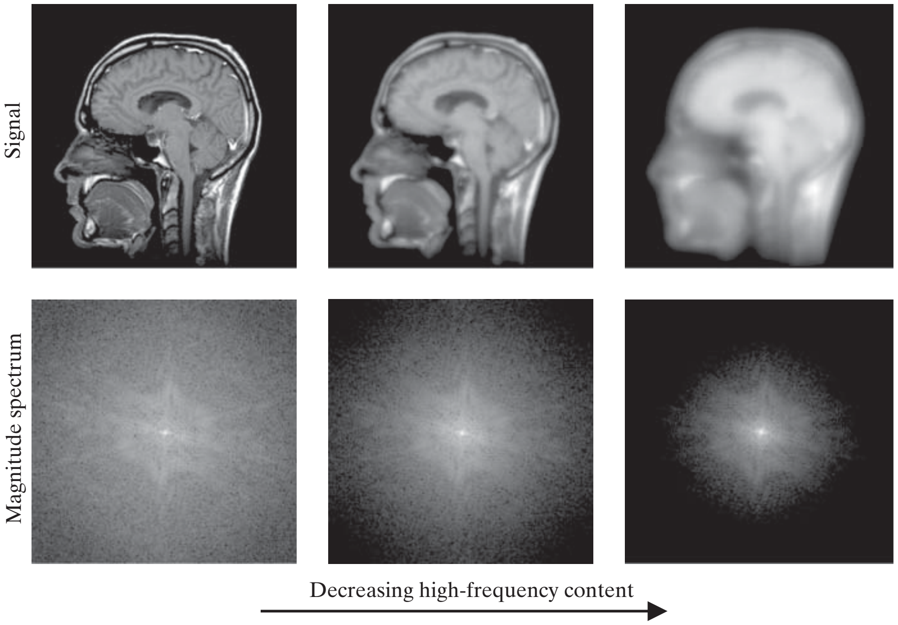


Figure 2.9
Three images of decreasing spatial variation (from left to right) and the associated magnitude spectra [depicted as $\log(1 + |F(u,v)|)$].

EXAMPLE 2.7

Consider the 1-D rect function

$$\text{rect}(x) = \begin{cases} 1, & \text{for } |x| < \frac{1}{2} \\ 0, & \text{for } |x| > \frac{1}{2} \end{cases} . \tag{2.75}$$

Question What is its Fourier transform?

Answer From (2.73), we have

$$\begin{aligned} \mathcal{F}_{1D}(\text{rect})(u) &= \int_{-\infty}^{\infty} \text{rect}(x) e^{-j2\pi ux}\, dx \\ &= \int_{-1/2}^{1/2} e^{-j2\pi ux}\, dx \\ &= \frac{1}{-j2\pi u}\, e^{-j2\pi ux}\Big|_{-1/2}^{1/2} \\ &= \frac{1}{\pi u} \frac{e^{j\pi u} - e^{-j\pi u}}{2j} \\ &= \frac{\sin(\pi u)}{\pi u} = \text{sinc}(u) \,. \end{aligned} \tag{2.76}$$

To derive (2.76), we have used the fact that if $f(x)$ is a single-valued, bounded, and integrable function on $[a, b]$ and there exists a function $F(x)$ such that $dF(x)/dx = f(x)$, for $a \le x \le b$, then

$$\int_a^x f(\xi)\, d\xi = F(\xi)\Big|_a^x = F(x) - F(a) \,, \tag{2.77}$$

for $a \le x \le b$. Therefore, the 1-D Fourier transform of the rect function is the sinc function (and vice versa). ■

2.5 Properties of the Fourier Transform

The Fourier transform satisfies a number of useful properties. Most of them are used in both theory and application to simplify calculations. The following sections discuss the most important ones. Refer to Table 2.2 for a summary.

2.5.1 Linearity

If the Fourier transforms of two signals $f(x,y)$ and $g(x,y)$ are $F(u,v)$ and $G(u,v)$, respectively, then

$$\mathcal{F}_{2D}(a_1 f + a_2 g)(u,v) = a_1 F(u,v) + a_2 G(u,v)\,, \tag{2.78}$$

where a_1 and a_2 are two constants. This property can be extended to a linear combination of an arbitrary number of signals.

TABLE 2.2

Properties of the Fourier Transform

Property	Signal	Fourier Transform
	$f(x,y)$	$F(u,v)$
	$g(x,y)$	$G(u,v)$
Linearity	$a_1 f(x,y) + a_2 g(x,y)$	$a_1 F(u,v) + a_2 G(u,v)$
Translation	$f(x - x_0, y - y_0)$	$F(u,v)e^{-j2\pi(ux_0+vy_0)}$
Conjugation	$f^*(x,y)$	$F^*(-u,-v)$
Conjugate symmetry	$f(x,y)$ is real-valued	$F(u,v) = F^*(-u,-v)$
		$F_R(u,v) = F_R(-u,-v)$
		$F_I(u,v) = -F_I(-u,-v)$
		$\lvert F(u,v)\rvert = \lvert F(-u,-v)\rvert$
		$\angle F(u,v) = -\angle F(-u,-v)$
Signal reversing	$f(-x,-y)$	$F(-u,-v)$
Scaling	$f(ax,by)$	$\frac{1}{\lvert ab\rvert}F\left(\frac{u}{a},\frac{v}{b}\right)$
Rotation	$f(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$	$F(u\cos\theta - v\sin\theta, u\sin\theta + v\cos\theta)$
Circular symmetry	$f(x,y)$ is circularly symmetric	$F(u,v)$ is circularly symmetric
		$\lvert F(u,v)\rvert = F(u,v)$
		$\angle F(u,v) = 0$
Convolution	$f(x,y) * g(x,y)$	$F(u,v)G(u,v)$
Product	$f(x,y)g(x,y)$	$F(u,v) * G(u,v)$
Separable product	$f(x)g(y)$	$F(u)G(v)$
Parseval's theorem	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lvert f(x,y)\rvert^2\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lvert F(u,v)\rvert^2\,du\,dv\,.$	

2.5.2 Translation

If $F(u, v)$ is the Fourier transform of a signal $f(x, y)$, and if

$$f_{x_0 y_0}(x, y) = f(x - x_0, y - y_0)\,, \tag{2.79}$$

then

$$\mathcal{F}_{2D}(f_{x_0 y_0})(u, v) = F(u, v) e^{-j2\pi(ux_0 + vy_0)}\,. \tag{2.80}$$

Notice that, in this case,

$$\left|\mathcal{F}_{2D}(f_{x_0 y_0})(u, v)\right| = |F(u, v)|\,, \tag{2.81}$$

and

$$\angle\mathcal{F}_{2D}(f_{x_0 y_0})(u, v) = \angle F(u, v) - 2\pi(ux_0 + vy_0)\,. \tag{2.82}$$

Therefore, translating a signal $f(x, y)$ does not affect its magnitude spectrum but subtracts a constant phase of $2\pi(ux_0 + vy_0)$ at each frequency (u, v).

2.5.3 Conjugation and Conjugate Symmetry

If $F(u, v)$ is the Fourier transform of a complex-valued signal $f(x, y)$, then

$$\mathcal{F}_{2D}(f^*)(u, v) = F^*(-u, -v)\,, \tag{2.83}$$

where $*$ denotes the complex conjugate. This is known as the *conjugation property* of the Fourier transform. When f is real-valued, its Fourier transform exhibits *conjugate symmetry*, which is defined by

$$F(u, v) = F^*(-u, -v)\,. \tag{2.84}$$

In this case, the real part $F_R(u, v)$ of $F(u, v)$ and the magnitude spectrum $|F(u, v)|$ are *symmetric functions*, whereas the imaginary part $F_I(u, v)$ and phase spectrum $\angle F(u, v)$ are *antisymmetric functions*:

$$F_R(u, v) = F_R(-u, -v) \qquad \text{and} \qquad F_I(u, v) = -F_I(-u, -v)\,, \tag{2.85}$$

$$|F(u, v)| = |F(-u, -v)| \qquad \text{and} \qquad \angle F(u, v) = -\angle F(-u, -v)\,. \tag{2.86}$$

Notice that the three magnitude spectra depicted in Figure 2.9 are symmetric around the origin.

2.5.4 Scaling

If $F(u, v)$ is the Fourier transform of a signal $f(x, y)$, and if

$$f_{ab}(x, y) = f(ax, by)\,, \tag{2.87}$$

where a and b are two nonzero constants, then

$$\mathcal{F}_{2D}(f_{ab})(u,v) = \frac{1}{|ab|} F\left(\frac{u}{a}, \frac{v}{b}\right) . \tag{2.88}$$

If we set $a = b = -1$, then the Fourier transform of signal $f(-x,-y)$ will be $F(-u,-v)$. Thus, reversing a signal in space also reverses its Fourier transform.

EXAMPLE 2.8

Detectors of many medical imaging systems can be modeled as rect functions of different sizes and locations.

Question Compute the Fourier transform of the following scaled and translated rect function:

$$f(x,y) = \text{rect}\left(\frac{x-x_0}{\Delta x}, \frac{y-y_0}{\Delta y}\right) .$$

Answer The Fourier transform of the rect function is the sinc function:

$$\mathcal{F}_{2D}(\text{rect})(u,v) = \text{sinc}(u,v) .$$

We see that $f(x,y)$ is the rect function scaled by factors Δx and Δy in each direction and then translated to (x_0, y_0). By using the scaling property of the Fourier transform, we have

$$\mathcal{F}_{2D}\left\{\text{rect}\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right)\right\} = \Delta x \Delta y \;\, \text{sinc}(\Delta x u, \Delta y v) .$$

By using the translation property, we have

$$\mathcal{F}_{2D}(f)(u,v) = \Delta x \Delta y \;\, \text{sinc}(\Delta x u, \Delta y v) e^{-j2\pi(ux_0+vy_0)} .$$

■

2.5.5 Rotation

Let us denote by $f_\theta(x,y)$ the signal

$$f_\theta(x,y) = f(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta) . \tag{2.89}$$

We note that $f_\theta(x,y)$ is a rotated version of $f(x,y)$, rotated by an angle θ around the origin $(0,0)$. If $F(u,v)$ is the Fourier transform of $f(x,y)$, then

$$\mathcal{F}_{2D}(f_\theta)(u,v) = F(u\cos\theta - v\sin\theta, u\sin\theta + v\cos\theta) . \tag{2.90}$$

If $f(x,y)$ is rotated by an angle θ, then its Fourier transform is rotated by angle θ as well.

2.5.6 Convolution

If $F(u,v)$ and $G(u,v)$ are the Fourier transforms of two signals $f(x,y)$ and $g(x,y)$, then the Fourier transform of the convolution $f(x,y) * g(x,y)$ equals the product of the individual Fourier transforms:

$$\mathcal{F}_{2D}(f * g)(u,v) = F(u,v)G(u,v)\,. \tag{2.91}$$

This property, known as the *convolution theorem*, provides a fundamental and very useful link between the space (f) and frequency (F) domains.

EXAMPLE 2.9

The convolution theorem provides a practical tool for computing the convolution of two signals, which may be difficult to conduct in the spatial domain.

Question Consider the two signals $f(x,y)$ and $g(x,y)$ given by

$$\begin{aligned} f(x,y) &= \operatorname{sinc}(Ux, Vy)\ , \\ g(x,y) &= \operatorname{sinc}(Vx, Uy)\ , \end{aligned}$$

where $0 < V \leq U$. What is the convolution $f(x,y) * g(x,y)$?

Answer From Table 2.1, we know that

$$\mathcal{F}_{2D}\{\operatorname{sinc}(x,y)\} = \operatorname{rect}(u,v)\,.$$

By using the scaling property of the Fourier transform, we have

$$F(u,v) = \mathcal{F}_{2D}(f)(u,v) = \frac{1}{UV}\operatorname{rect}\left(\frac{u}{U},\frac{v}{V}\right)$$

and

$$G(u,v) = \mathcal{F}_{2D}(g)(u,v) = \frac{1}{UV}\operatorname{rect}\left(\frac{u}{V},\frac{v}{U}\right)\,.$$

The convolution of $f(x,y)$ and $g(x,y)$ is

$$\begin{aligned} f(x,y) * g(x,y) &= \mathcal{F}_{2D}^{-1}\{F(u,v)G(u,v)\} \\ &= \mathcal{F}_{2D}^{-1}\left\{\frac{1}{(UV)^2}\operatorname{rect}\left(\frac{u}{U},\frac{v}{V}\right)\operatorname{rect}\left(\frac{u}{V},\frac{v}{U}\right)\right\} \\ &= \mathcal{F}_{2D}^{-1}\left\{\frac{1}{(UV)^2}\operatorname{rect}\left(\frac{u}{V},\frac{v}{V}\right)\right\} \\ &= \frac{1}{U^2}\mathcal{F}_{2D}^{-1}\left\{\frac{1}{V^2}\operatorname{rect}\left(\frac{u}{V},\frac{v}{V}\right)\right\} \\ &= \frac{1}{U^2}\operatorname{sinc}(Vx, Vy)\ . \end{aligned}$$

■

2.5.7 Product

If $F(u,v)$ and $G(u,v)$ are the Fourier transforms of two signals $f(x,y)$ and $g(x,y)$, the Fourier transform of the product $f(x,y)g(x,y)$ equals the convolution of the two Fourier transforms:

$$\begin{aligned} \mathcal{F}_{2D}(fg)(u,v) &= F(u,v) * G(u,v) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(\xi,\eta)F(u-\xi, v-\eta)\,d\xi\,d\eta\,. \end{aligned} \tag{2.92}$$

2.5.8 Separable Product

If f is a separable signal, such that

$$f(x, y) = f_1(x) f_2(y)\,, \tag{2.93}$$

then

$$\mathcal{F}_{2D}(f)(u, v) = F_1(u) F_2(v)\,, \tag{2.94}$$

where

$$F_1(u) = \mathcal{F}_{1D}(f_1)(u) \qquad \text{and} \qquad F_2(v) = \mathcal{F}_{1D}(f_2)(v) \tag{2.95}$$

are the 1-D Fourier transforms of $f_1(x)$ and $f_2(y)$, respectively. Therefore, the Fourier transform of a separable signal is also separable. In this case, the Fourier transform of a separable 2-D signal $f(x, y)$ can be computed by independently calculating the two 1-D Fourier transforms of $f_1(x)$ and $f_2(y)$ and then multiplying the results.

2.5.9 Parseval's Theorem

If $F(u, v)$ is the Fourier transform of a signal $f(x, y)$, then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x, y)|^2\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |F(u, v)|^2\, du\, dv\,, \tag{2.96}$$

which is known as *Parseval's theorem*. This relationship says that the total energy of a signal $f(x, y)$ in the spatial domain equals its total energy in the frequency domain. In other words, the Fourier transform (as well as its inverse) are energy-preserving ("unit gain") transformations.

2.5.10 Separability

The Fourier transform $F(u, v)$ of a 2-D signal $f(x, y)$ can be calculated using two simpler 1-D Fourier transforms, as follows:

- Compute $r(u, y) = \int_{-\infty}^{\infty} f(x, y) e^{-j2\pi ux}\, dx$, for every y.
- Compute $F(u, v) = \int_{-\infty}^{\infty} r(u, y) e^{-j2\pi vy}\, dy$, for every u.

Indeed,

$$\begin{aligned}\int_{-\infty}^{\infty} r(u, y) e^{-j2\pi vy}\, dy &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(x, y) e^{-j2\pi ux}\, dx\right] e^{-j2\pi vy}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y) e^{-j2\pi(ux+vy)}\, dx\, dy \\ &= F(u, v)\,. \end{aligned} \tag{2.97}$$

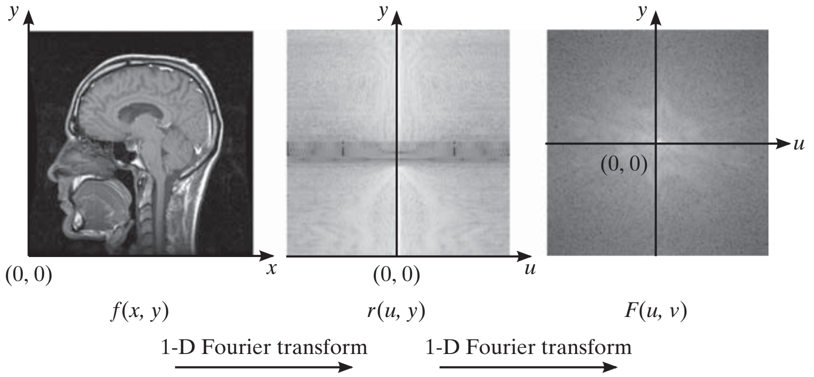


Figure 2.10 Calculation of the 2-D Fourier transform by using two 1-D Fourier transforms in cascade.

The first step fixes y, treats image $f(x, y)$ as a 1-D signal, and calculates its 1-D Fourier transform $r(u, y)$. The second step fixes u, treats $r(u, y)$ as a 1-D signal, and calculates its 1-D Fourier transform. This is illustrated in Figure 2.10. Notice that calculation of 1-D Fourier transforms is simpler than their 2-D counterparts. Therefore, this technique is strongly recommended for calculating 2-D Fourier transforms (and their inverses).

2.6 Transfer Function

In (2.38), we represented an image as the convolution of the object with the PSF. Using the Fourier transform, we can develop an equivalent representation in Fourier space. In order to do so we need to introduce the Fourier space equivalent of the PSF.

In order to determine the PSF $h(x, y)$ of an LSI system $\mathcal{S}$, we observe the output of the system to a point impulse [see (2.36)]. We now replace the point impulse with one of the complex exponential signals $e^{j2\pi(ux+vy)}$ in decomposition (2.63), for any arbitrary but fixed (u, v). As a direct consequence of the convolution integral (2.38), following a short derivation that uses (2.48), we have

$$g(x, y) = H(u, v)e^{j2\pi(ux+vy)}, \tag{2.98}$$

where

$$H(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\xi, \eta)e^{-j2\pi(u\xi+v\eta)}\, d\xi\, d\eta\,. \tag{2.99}$$

Notice that in (2.98) the output equals the input multiplied by $H(u, v)$. Recalling that (2.98) applies to any arbitrary but fixed (u, v), this means that when the input to an LSI system is a complex exponential the output will be the same complex exponential but scaled by $H(u, v)$.

From (2.62) and (2.99), it is clear that $H(u, v)$ is the Fourier transform of the PSF $h(x, y)$. $H(u, v)$ is known as the *transfer function* of the LSI system $\mathcal{S}$. $H(u, v)$ is also known as the *frequency response* or the *optical transfer function* (OTF) of system $\mathcal{S}$. The transfer function uniquely characterizes an LSI system since, given $H(u,v)$, the PSF is uniquely determined using the inverse Fourier transform:

$$h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(u,v)e^{j2\pi(ux+vy)}\,du\,dv\,. \tag{2.100}$$

The transfer function is an important tool for studying the behavior of LSI systems because of the convolution property of the Fourier transform. If an input $f(x,y)$ to an LSI system $\mathcal{S}$ with transfer function $H(u,v)$ produces an output $g(x,y)$ then, as a direct consequence of (2.39) and (2.91), we have

$$G(u,v) = H(u,v)F(u,v)\,, \tag{2.101}$$

where $F(u,v)$ and $G(u,v)$ are the Fourier transforms of $f(x,y)$ and $g(x,y)$, respectively. This equation is much simpler than the convolution equation (2.39) and allows an alternative (and simpler) method of studying the behavior of an LSI system.

EXAMPLE 2.10

Consider an idealized system whose PSF is $h(x,y) = \delta(x - x_0, y - y_0)$.

Question What is the transfer function $H(u,v)$ of the system, and what is the output $g(x,y)$ of the system to an input signal $f(x,y)$?

Answer The transfer function of the system is

$$\begin{aligned} H(u,v) &= \mathcal{F}\{h(x,y)\} \\ &= \mathcal{F}\{\delta(x - x_0, y - y_0)\} \\ &= e^{-j2\pi(ux_0+vy_0)}\,. \end{aligned}$$

For an input signal $f(x,y)$, the output is its translated version $f(x - x_0, y - y_0)$. This conclusion can be drawn from the sifting property of the δ function. Now, let us prove this using the transfer function.

The Fourier transform of the input signal $f(x,y)$ is $F(u,v) = \mathcal{F}(f)(u,v)$. The Fourier transform of the output signal $g(x,y)$ is

$$\begin{aligned} G(u,v) &= F(u,v)H(u,v) \\ &= F(u,v)e^{-j2\pi(ux_0+vy_0)}\,. \end{aligned}$$

By applying the translation property of the Fourier transform, we have $g(x,y) = \mathcal{F}^{-1}(G)(x,y)$, which is the input signal $f(x,y)$ translated to (x_0, y_0). ■

Remember that in Fourier or frequency space we are dealing with the representation of an object as sine waves of different frequencies. In medical images, these are frequencies in space—*spatial frequencies*. Low spatial frequencies represent signals that vary slowly across the image; high spatial frequencies represent signals that vary rapidly across a local region (e.g., at the edges of structures within an image).

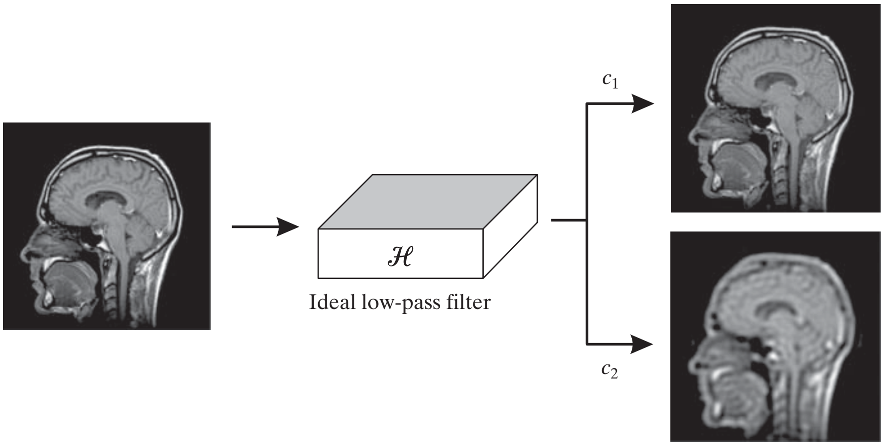


Figure 2.11
The response of an ideal low-pass filter for two values of the cutoff frequency c ($c_1 > c_2$).

As an example, consider an LSI system $\mathcal{S}$ whose transfer function $H(u, v)$ is given by

$$H(u, v) = \begin{cases} 1, & \text{for } \sqrt{u^2 + v^2} \leq c \\ 0, & \text{for } \sqrt{u^2 + v^2} > c \end{cases} . \tag{2.102}$$

This is known as an *ideal low-pass filter* with *cutoff frequency c*. From (2.101) and (2.102), we have

$$G(u, v) = \begin{cases} F(u, v), & \text{for } \sqrt{u^2 + v^2} \leq c \\ 0, & \text{for } \sqrt{u^2 + v^2} > c \end{cases} . \tag{2.103}$$

Therefore, the input spectrum will be eliminated at high spatial frequencies (u, v) given by $\sqrt{u^2 + v^2} > c$. Applying such a system to an input signal will result in *signal smoothing*, with fine signal details (details that are mainly responsible for high-frequency spectral content) being eliminated. The amount of smoothing depends directly on the value of c, with smaller values of c producing more smoothing. This is illustrated in Figure 2.11. Most imaging systems can be modeled as some sort of low-pass filter.

We should point out here that, usually, the numerical implementation of the convolution equation (2.39) is not done in the space domain [i.e., by calculating the double integral in (2.38)], but in the frequency domain, using (2.101). The main reason is the existence of an efficient algorithm (the *fast Fourier transform*, or FFT) that allows a fast computer implementation of (2.101).

2.7 Circular Symmetry and the Hankel Transform

Often, the performance of a medical imaging system does not depend on the orientation of the patient with respect to the system. This orientation independence arises from the circular symmetry of certain signals, especially the

PSF. A 2-D signal $f(x,y)$ is defined to be *circularly symmetric* if it satisfies

$$f_\theta(x,y) = f(x,y), \qquad \text{for every } \theta\,, \tag{2.104}$$

where $f_\theta(x,y)$ is a rotated version of $f(x,y)$ [see (2.89)]. It follows from (2.90) that $\mathcal{F}_{2D}(f_\theta)(u,v)$ is also circularly symmetric. If $f(x,y)$ is circularly symmetric, then it is even in both x and y, and $F(u,v)$ is even in both u and v. In addition, $F(u,v)$ is real; therefore,

$$|F(u,v)| = F(u,v) \quad \text{and} \quad \angle F(u,v) = 0\,. \tag{2.105}$$

Functions $f(x,y)$ that are circularly symmetric can be written as a function of radius only; that is,

$$f(x,y) = \mathrm{f}(r)\,, \tag{2.106}$$

where $r = \sqrt{x^2+y^2}$. Since in this case $F(u,v)$ is also circularly symmetric, $F(u,v)$ must satisfy

$$F(u,v) = \mathrm{F}(q)\,, \tag{2.107}$$

where $q = \sqrt{u^2+v^2}$. The functions $\mathrm{f}(r)$ and $\mathrm{F}(q)$ are 1-D functions representing 2-D functions that are related by the Fourier transform. It can be shown (see Problem 2.19) that $\mathrm{f}(r)$ and $\mathrm{F}(q)$ are related by

$$\mathrm{F}(q) = 2\pi \int_0^\infty \mathrm{f}(r) J_0(2\pi q r)\, r\, dr\,, \tag{2.108}$$

where $J_0(r)$ is the zero-order Bessel function of the first kind. The nth order Bessel function of the first kind is given by

$$J_n(r) = \frac{1}{\pi}\int_0^\pi \cos(n\phi - r\sin\phi)\, d\phi\,, \quad n = 0, 1, 2, \ldots\,, \tag{2.109}$$

and therefore

$$J_0(r) = \frac{1}{\pi}\int_0^\pi \cos(r\sin\phi)\, d\phi\,. \tag{2.110}$$

The relation (2.108) is called the *Hankel transform*, and we will use the notation

$$\mathrm{F}(q) = \mathcal{H}\{\mathrm{f}(r)\}\,. \tag{2.111}$$

The inverse Hankel transform is identical to the forward transform,

$$\mathrm{f}(r) = 2\pi \int_0^\infty \mathrm{F}(q) J_0(2\pi q r)\, q\, dq\,. \tag{2.112}$$

TABLE 2.3

Selected Hankel Transform Pairs

Signal	Hankel Transform
$\exp\{-\pi r^2\}$	$\exp\{-\pi q^2\}$
1	$\delta(q)/\pi q = \delta(u, v)$
$\delta(r-a)$	$2\pi a J_0(2\pi a q)$
$\mathrm{rect}(r)$	$\frac{J_1(\pi q)}{2q}$
$\mathrm{sinc}(r)$	$\frac{2\,\mathrm{rect}(q)}{\pi\sqrt{1-4q^2}}$
$\frac{1}{r}$	$\frac{1}{q}$

If a 2-D signal is circularly symmetric, the Fourier transform of the signal can be found using the Hankel transform. Some Hankel transform pairs are given in Table 2.3.

Table 2.1 presented the Fourier transform of the 2-D Gaussian function as

$$\mathcal{F}\left\{e^{-\pi(x^2+y^2)}\right\} = e^{-\pi(u^2+v^2)}, \tag{2.113}$$

which itself is a 2-D Gaussian function. These Gaussian functions are circularly symmetric, and by direct substitution of $r^2 = x^2 + y^2$ and $q^2 = u^2 + v^2$, we find the Hankel transform pair

$$\mathcal{H}\left\{e^{-\pi r^2}\right\} = e^{-\pi q^2}. \tag{2.114}$$

The Gaussian function is a good model for the blurring inherent in medical imaging systems, and we will have many occasions throughout the text to use it and its Fourier (or Hankel) transform.

The *unit disk*, defined as

$$\mathrm{f}(r) = \mathrm{rect}(r), \tag{2.115}$$

is another example of a circularly symmetric function. It can be shown (see Problem 2.20) that the Hankel transform of the unit disk is given by

$$\mathcal{H}\{\mathrm{rect}(r)\} = \frac{J_1(\pi q)}{2q}, \tag{2.116}$$

where J_1 is the first-order Bessel function of the first kind [see (2.109)]. Forming an analogy to the special relationship between the rect and sinc functions in Fourier transforms, the so-called jinc function is defined as

$$\mathrm{jinc}(q) = \frac{J_1(\pi q)}{2q}. \tag{2.117}$$

Therefore, the rect and jinc form a Hankel transform pair.

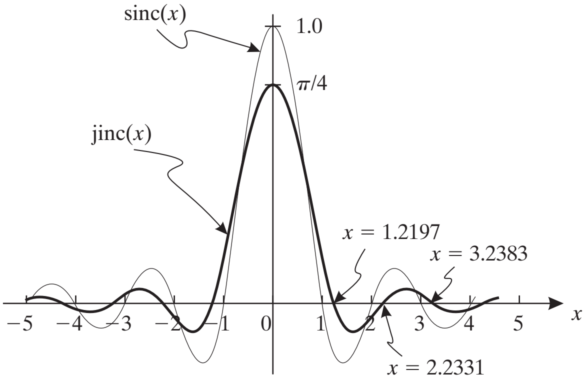


Figure 2.12
The 1-D jinc and sinc functions.

A plot of the jinc(x) function, shown with the sinc(x) function for comparison, is shown in Figure 2.12. Like the sinc, the jinc function has a central lobe with its maximum value at the origin (see Figure 2.12),

$$\text{jinc}(0) = \frac{\pi}{4}\,, \tag{2.118}$$

and has a set of zeros as $x \to \infty$. The first few zeros are at the following locations:

$$\text{jinc}(1.2197) = 0\,, \tag{2.119a}$$

$$\text{jinc}(2.2331) = 0\,, \tag{2.119b}$$

$$\text{jinc}(3.2383) = 0\,. \tag{2.119c}$$

The jinc function takes on half its peak value at $x = 0.70576$.

The scaling theorem for the Hankel transform is derived from that of the Fourier transform. First, it is recognized that the scaling parameters, a and b, must be equal, $a = b$, in order for circular symmetry to be preserved. From (2.88) it follows that

$$\mathcal{H}\left\{\text{f}(ar)\right\} = \frac{1}{a^2}\text{F}(q/a)\,. \tag{2.120}$$

EXAMPLE 2.11

In some medical imaging systems, only spatial frequencies smaller than ϱ_0 can be imaged.

Question What is the function having uniform spatial frequencies within the disk of radius ϱ_0 and what is its inverse Fourier transform?

Answer The disk of radius ϱ_0 is a circularly symmetric function, a scaled version of rect(q) given by

$$\text{F}(q) = \text{rect}\left(\frac{q}{2\varrho_0}\right).$$

Since the inverse Hankel transform is the same as the forward Hankel transform, we have

$$\text{jinc}(r) = \mathcal{H}^{-1}\left\{\text{rect}(q)\right\}\,.$$

The scaling theorem for the Hankel transform, (2.120), is then used to yield

$$\begin{aligned} f(r) &= \mathcal{H}^{-1}\left\{ \mathrm{rect}\left(\frac{q}{2\varrho_0}\right) \right\} \\ &= 4\varrho_0^2\, \mathrm{jinc}(2\varrho_0 r)\,. \end{aligned}$$ ■

2.8 Summary and Key Concepts

In this book, we take a signals and systems approach to the understanding and analysis of medical imaging systems. Signals model physical processes; systems model how medical imaging systems create new signals (images, in our case) from these original signals. In this chapter, we presented the following key concepts that you should now understand:

1. *Signals* are mathematical functions, and they can be continuous, discrete, or mixed.
2. The *point impulse* or *impulse function* is used as the input signal to characterize the response of a system, which is the impulse response.
3. *Comb* and *sampling functions*, *rect* and *sinc functions*, and *exponential* and *sinusoidal signals* are used to help describe and characterize imaging systems.
4. A system is *linear* if, when the input consists of a collection of signals, the output is a summation of the responses of the system to each individual input signal.
5. A system is *shift-invariant* if an arbitrary translation of the input signal results in an identical translation of the output.
6. A signal is *separable* if it can be represented as the product of 1-D signals.
7. The *Fourier transform* represents signals as a sum of sinusoids of different frequencies, with associated magnitude and phase.
8. The Fourier space analog of the impulse response is the *transfer function*.
9. The *output* of a linear, shift-invariant system is the convolution of the input with the impulse response; in Fourier space, it is the product of the Fourier transform of the input and the transfer function.
10. Functions that are *circularly symmetric* have Fourier transforms that are also circularly symmetric.
11. The *Hankel transform* is a one-dimensional transform that yields the Fourier transform of circularly symmetric functions.

Further Reading

Gonzalez, R.C. and Woods, R.E. *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2007.

Macovski, A. *Medical Imaging Systems*. Englewood Cliffs, NJ: Prentice Hall, 1983.

Oppenheim, A.V., Willsky, A.S., and Nawad, S.H. *Signals and Systems*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1996.

Problems

Signals and Their Properties

2.1 Determine whether the following signals are separable. Fully justify your answers.

(a) $\delta_s(x, y) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(x - m, y - n)$.

(b) $\delta_\ell(x, y) = \delta(x \cos\theta + y \sin\theta - \ell)$.

(c) $e(x, y) = \exp\{j2\pi(u_0 x + v_0 y)\}$.

(d) $s(x, y) = \sin[2\pi(u_0 x + v_0 y)]$.

2.2 Determine whether the following signals are periodic. If they are, find the smallest periods.

(a) $\delta(x, y)$.

(b) $\mathrm{comb}(x, y)$.

(c) $f(x, y) = \sin(2\pi x)\cos(4\pi y)$.

(d) $f(x, y) = \sin(2\pi(x + y))$.

(e) $f(x, y) = \sin(2\pi(x^2 + y^2))$.

(f) $f_d(m, n) = \sin\left(\frac{\pi}{5}m\right)\cos\left(\frac{\pi}{5}n\right)$.

(g) $f_d(m, n) = \sin\left(\frac{1}{5}m\right)\cos\left(\frac{1}{5}n\right)$.

2.3 We define the *energy* of a signal f in the finite rectangular window defined by $-X \le x \le X$ and $-Y \le y \le Y$, where $X, Y < \infty$, by

$$E_{XY} = \int_{-X}^{X} \int_{-Y}^{Y} |f(x, y)|^2 \, dx \, dy\,,$$

where f might be complex-valued, in which case $|\cdot|$ denotes complex magnitude (or modulus). The *total energy* is defined as

$$E_\infty = \lim_{X\to\infty} \lim_{Y\to\infty} E_{XY}\,.$$

The *power* of signal f is defined as

$$P_{XY} = \frac{1}{4XY} \int_{-X}^{X} \int_{-Y}^{Y} |f(x, y)|^2 \, dx \, dy = \frac{E_{XY}}{4XY}$$

and the *total power* as

$$P_\infty = \lim_{X\to\infty} \lim_{Y\to\infty} P_{XY}\,.$$

For each of the signals in Problem 2.1, determine the values of E_∞ and P_∞.

Systems and Their Properties

2.4 Two LSI systems are connected in cascade. Show that the overall system is an LSI system and prove equations (2.46) and (2.47).

2.5 Show that an LSI system is BIBO stable if and only if its PSF is absolutely integrable.

2.6 Determine whether the system $g(x,y) = f(x,-1) + f(0,y)$ is

(**a**) linear.

(**b**) shift-invariant.

2.7 For each system with the following input-output equation, determine whether the system is (1) linear and (2) shift-invariant.

(**a**) $g(x,y) = f(x,y)f(x - x_0, y)$.

(**b**) $g(x,y) = \int_{-\infty}^{\infty} f(x,\eta)d\eta$.

2.8 For each system with the following PSF, determine whether the system is stable.

(**a**) $h(x,y) = x^2 + y^2$.

(**b**) $h(x,y) = \exp\{-(x^2 + y^2)\}$.

(**c**) $h(x,y) = x^2 \exp\{-y^2\}$.

2.9 Consider the 1-D system whose input-output equation is given by

$$g(x) = f(x) * f(x)\,,$$

where * denotes convolution.

(**a**) Write an integral expression that gives $g(x)$ as a function of $f(x)$.

(**b**) Determine whether the system is linear.

(**c**) Determine whether the system is shift-invariant.

Convolution of Signals

2.10 Given a continuous signal $f(x,y) = x + y^2$, evaluate the following:

(**a**) $f(x,y)\delta(x-1, y-2)$.

(**b**) $f(x,y) * \delta(x-1, y-2)$.

(**c**) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x-1, y-2)f(x,3)dx\,dy$.

(**d**) $\delta(x-1, y-2) * f(x+1, y+2)$.

2.11 Consider two continuous signals $f(x,y)$ and $g(x,y)$ that are separable—that is, $f(x,y) = f_1(x)f_2(y)$ and $g(x,y) = g_1(x)g_2(y)$.

(**a**) Show that their convolution is also separable.

(**b**) Express the convolution in terms of $f_1(x)$, $f_2(y)$, $g_1(x)$, and $g_2(y)$.

2.12 By using two 1-D convolution integrals, calculate the 2-D convolution of signal $f(x,y) = x + y$ with the exponential PSF $h(x,y) = \exp\{-(x^2 + y^2)\}$.

Fourier Transforms and Their Properties

2.13 Find the Fourier transforms of the following continuous signals:

(**a**) $\delta_s(x,y) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \delta(x-m, y-n)$.

(**b**) $\delta_s(x,y;\,\Delta x, \Delta y) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \delta(x - m\Delta x, y - n\Delta y)$.

(**c**) $s(x,y) = \sin[2\pi(u_0 x + v_0 y)]$.

(d) $c(x, y) = \cos[2\pi(u_0x + v_0y)]$.

(e) $f(x, y) = \frac{1}{2\pi\sigma^2}\exp\{-(x^2 + y^2)/2\sigma^2\}$.

2.14 Suppose $F(u)$ is the Fourier transform of a 1-D real signal $f(x)$, $F(u) = \mathcal{F}[f(x)]$. Prove the following, where superscript * denotes complex conjugate:

(a) If $f(x) = f(-x)$, then $F^*(u) = F(u)$.

(b) If $f(x) = -f(-x)$, then $F^*(u) = -F(u)$.

2.15 In the previous problem, if $f(x)$ is not a real signal, can we still arrive at the same conclusions? If not, what kind of symmetric property does $F(u) = \mathcal{F}[f(x)]$ have when $f(x) = f(-x)$?

2.16 Prove the following properties of the Fourier transform:

(a) Conjugate and conjugate symmetry.

(b) Scaling.

(c) Convolution.

(d) Product.

2.17 Show that the 2-D Fourier transform of the rect function is the sinc function, whereas the 2-D Fourier transform of the sinc function is the rect function. Determine the values of E_∞ and P_∞ for the sinc function (see Problem 2.3).

2.18 Find the Fourier transform of the separable continuous signal $f(x, y) = \sin(2\pi ax)\cos(2\pi by)$.

2.19 Prove that a circularly symmetric function and its Fourier transform are related by the zero-order Bessel function of the first kind, as expressed by (2.108).

2.20 Show that the Hankel transform of a unit disk is a jinc function.

Transfer Function

2.21 A new imaging system with which you are experimenting has anisotropic properties. You measure the impulse response function as $h(x, y) = e^{-\pi(x^2+y^2/4)}$.

(a) Sketch the impulse response function.

(b) What is the transfer function?

2.22 A medical imaging system has the following line spread function, where $\alpha = 2$ radians/cm:

$$l(x) = \begin{cases} \cos(\alpha x) & |\alpha x| \le \pi/2 \\ 0 & \text{otherwise} \end{cases}.$$

(a) Suppose a bar phantom is imaged, where the bars have width w, separation w, and unity height. Assume $\pi/2\alpha \le w \le \pi/\alpha$. What are the responses of the system at the center of a bar and halfway between two bars?

(b) From the line spread function, can we tell whether the system is isotropic?

(c) Assume the system is separable with $h(x, y) = h_{1D}(x)h_{1D}(y)$. Compute its transfer function.

Applications, Extensions, and Advanced Topics

2.23 The PSF of a medical imaging system is given by

$$h(x,y) = e^{-(|x|+|y|)},$$

where x and y are in millimeters.

(**a**) Is the system separable? Explain.

(**b**) Is this system isotropic? Explain.

(**c**) What is the response of the system to the line impulse $f(x,y) = \delta(x)$?

(**d**) What is the response of the system to the line impulse $f(x,y) = \delta(x-y)$?

2.24 Consider a one-dimensional linear imaging system whose PSF is given by

$$h(x;\xi) = e^{\frac{-(x-\xi)^2}{2}},$$

which represents the response to the shifted impulse $\delta(x-\xi)$.

(**a**) Is this imaging system shift-invariant? Explain.

(**b**) Write the output image $g(x)$ of this system when the input signal is $f(x) = \delta(x+1) + \delta(x) + \delta(x-1)$?

2.25 Consider a 1-D ideal *high-pass filter* whose spectrum is shown in Figure P2.1.

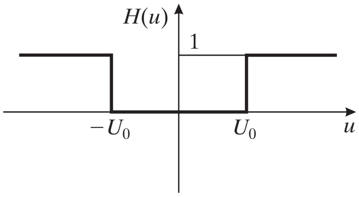


Figure P2.1
An ideal high-pass filter $H(u)$. For Problem 2.25.

(**a**) Compute the impulse response of the filter.

(**b**) Using the linearity of convolution, compute the system response to (1) a constant function $f(t) = c$ and (2) a unit step $f(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}$.

2.26 The impulse response of a 1-D LTI system is

$$h(t) = \frac{1}{T}\left[-\operatorname{rect}\left(\frac{t+0.75T}{0.5T}\right) + \operatorname{rect}\left(\frac{t}{T}\right) - \operatorname{rect}\left(\frac{t-0.75T}{0.5T}\right)\right].$$

(**a**) Plot $h(t)$ with labeled axes. Is the system stable? Is it causal?

(**b**) Find and plot the response of the system to a constant signal $f(t) = c$.

(**c**) Find and plot the response of the system to a unit step signal

$$f(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}.$$

(d) Compute the Fourier transform of $h(t)$.

(e) Plot the magnitude spectrum $|H(u)|$ for $T = 0.25$, $T = 0.1$, and $T = 0.05$.

(f) What kind of filter is this LTI system?

2.27 In CT image reconstruction, a ramp filter is used in *filtered backprojection*. The transfer function of the ramp filter is defined as

$$H(\varrho) = |\varrho| \,.$$

For practical reasons, a windowed ramp filter $\hat{H}(\varrho) = W(\varrho)|\varrho|$ is used instead of $H(\varrho)$.

(a) Assume that $W(\varrho)$ is a rectangular window defined as

$$W(\varrho) = \begin{cases} 1, & |\varrho| \leq \varrho_0 \\ 0, & \text{otherwise} \end{cases} ,$$

where ϱ_0 is the cutoff frequency. Find $\hat{h}(r) = \mathcal{F}^{-1}\{\hat{H}(\varrho)\}$.

(b) What are the responses of a ramp filter to (1) a constant function $f(r) = c$ and (2) a sinusoid function $f(r) = \sin(\omega r)$.

2.28 For the imaging system depicted in Figure P2.2, show that the output is a scaled and inverted replica of the input.

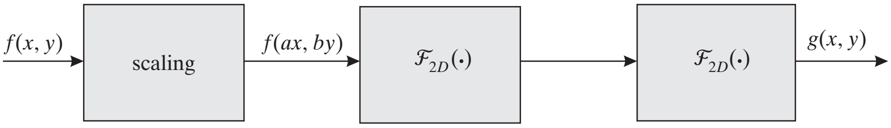


Figure P2.2
A cascade of subsystems representing the imaging system in Problem 2.28.

2.29 A continuous signal

$$f(x, y) = \begin{cases} 1, & \text{for } x = y = 0 \\ \dfrac{\sin(ax)\sin(by)}{\pi^2 xy}, & \text{otherwise} \end{cases}$$

is corrupted by an additive sinusoidal signal

$$\eta(x, y) = \cos\left[2\pi(Ax + By)\right], \quad A, B \geq 0,$$

and a set of measurements

$$g(x, y) = f(x, y) + \eta(x, y)$$

is obtained. Design an ideal low-pass filter (an LSI system whose transfer function equals 1, for $0 \leq |u| \leq U$, $0 \leq |v| \leq V$, and 0 otherwise) with impulse response $h(x, y)$, such that

$$f(x, y) = h(x, y) * g(x, y) \,.$$

Given a and b, determine all possible values of A and B for which this is possible. Explain your answers in full detail. You might use the formula

$$\int_{-\infty}^{\infty} \frac{\sin(\alpha t)}{\pi t} e^{-j\tau t} dt = \begin{cases} 1, & \text{for } |\tau| \le \alpha \\ 0, & \text{for } |\tau| > \alpha \end{cases}.$$

2.30 The *discrete-time Fourier transform* (DTFT) of a 1-D aperiodic discrete time signal with finite energy, $f(m)$, is defined as

$$F(e^{i\omega}) = \sum_{m=-\infty}^{+\infty} f(m) e^{-j\omega m}$$

and the inverse transform is

$$f(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{i\omega}) e^{j\omega m} d\omega.$$

A continuous signal $g(x) = \mathrm{rect}(x/2)$ is sampled with a given sampling period Δx_1 to give a discrete signal $g_1(m)$.

(**a**) Compute and plot the continuous Fourier transform of $g(x)$.

(**b**) Let $\Delta x_1 = 1/2$. Compute and plot $G_1(\omega)$, the DTFT of $g_1(m) = g(m\Delta x_1)$.

(**c**) Let $\Delta x_2 = 1$. Compute and plot $G_2(\omega)$, the DTFT of $g_2(m) = g(m\Delta x_2)$.

(**d**) Derive a relationship between the continuous Fourier transform of $g(x)$ and its discrete version sampled with Δx.

(**e**) Prove the convolution property of the DTFT:

$$\mathcal{F}_{\mathrm{DTFT}}\{x(m) * y(m)\} = \mathcal{F}_{\mathrm{DTFT}}\{x(m)\}\mathcal{F}_{\mathrm{DTFT}}\{y(m)\}.$$

(**f**) Verify the convolution property by computing $g_1(m) * g_2(m)$.

CHAPTER 3

Image Quality

3.1 Introduction

The primary purpose of a medical imaging system is to create images of the internal structures and functions of the human body that can be used by medical professionals to diagnose abnormal conditions, determine the underlying mechanisms that produce and control these conditions, guide therapeutic procedures, and monitor the effectiveness of treatment. The ability of medical professionals to successfully accomplish these tasks strongly depends on the *quality* of the images acquired by the medical imaging system at hand, where by "quality" we mean the degree to which an image allows medical professionals to accomplish their goals.

Image quality depends on the particular imaging modality used. With each modality, the range of image quality may be considerable, depending on the characteristics and setup of the particular medical imaging system, the skill of the operator handling the system, and several other factors, such as patient characteristics and imaging time. Studying how these factors affect image quality is an important and complicated task. This task is simplified by focusing on the following six important factors: (1) contrast, (2) resolution, (3) noise, (4) artifacts, (5) distortion, and (6) accuracy.

The ability of medical professionals to discriminate among anatomical or functional features in a given image strongly depends on *contrast*. Contrast quantifies the difference between image characteristics (e.g., intensities in shades of gray or color) of an object (or feature within an object) and surrounding objects or background. High contrast allows easier identification of individual objects in an image, whereas low contrast makes this task difficult. In the overview to Part I, we presented images from different modalities that have different contrast. For example, the brain structures in the positron emission tomography (PET) image in Figure I.4(c) are of higher contrast than the same structures in Figure I.4(a) or (b), because the actual signal coming from the patient is of intrinsically higher contrast in PET.

Sometimes, medical images are blurry and lack detail. The ability of a medical imaging system to depict details is known as *resolution*. High resolution systems create images of high diagnostic quality. Low resolution systems create images that lack fine detail. For example, while the PET image in Figure I.4(c) has higher contrast than the CT image in Figure I.4(a) or the MRI image in Figure I.4(b), it has poorer spatial resolution.

A medical image may be corrupted by random fluctuations in image intensity that do not contribute to image quality. This is known as *noise*. The source, amount, and type of noise depends on the particular imaging modality used. Object visibility is often reduced by the presence of noise, because the noise masks image features. Nuclear medicine images tend to have the highest noise, as reflected in the PET image in Figure I.4(c).

Most medical imaging systems can create image features that do not represent a valid object within, or characteristic of, the patient. These features are known as *artifacts*, and can frequently obscure important features or be falsely interpreted as abnormal findings. Medical images should not only make desired features visible but should also give an accurate impression of their shape, size, position, and other geometric characteristics. Unfortunately, for many reasons to be explained later in this book, medical imaging systems frequently introduce *distortion* of these important factors. Distortion in medical images should be corrected in order to improve the diagnostic quality of these images.

Ultimately, the quality of medical images should be judged on their utility in the context of a specific clinical application. For example, medical images that increase the chance of tumor detectability in nuclear medicine should be preferable to images with poor tumor detectability. Fundamentally, we are interested in the *accuracy* of medical images in the context of a clinical application, where "accuracy" means both conformity to truth and clinical utility.

The user of a medical imaging system is very interested in adjusting the system to produce images of the highest possible quality, while maintaining a safe environment for the patient. In order to achieve this, methods must be developed for evaluating image quality. Since image quality depends primarily on the six previously discussed factors, we need to mathematically quantify these factors and systematically study their influence on image quality. The purpose of this chapter is to provide a fundamental exposition on contrast, resolution, noise, artifacts, distortion, and accuracy. More detailed discussions on these factors, and how they affect image quality for a specific imaging modality, will be provided later.

3.2 Contrast

Contrast refers to differences between the image intensity of an object and surrounding objects or background. This difference, or *image contrast*, is itself the result of the inherent *object contrast* within the patient. In general, the goal of a medical imaging system is to accurately portray or preserve the true object contrast in the image. Particularly for detection of abnormalities, a medical imaging system that produces high contrast images is preferable to a system that produces low contrast images, since anatomical and functional features are easier to identify in high contrast images.

3.2.1 Modulation

Use of a periodic signal and its modulation is an effective way to quantify contrast. The *modulation* m_f of a periodic signal $f(x,y)$, with maximum and minimum values $f_{\max}$ and $f_{\min}$, is defined by

$$m_f = \frac{f_{\max} - f_{\min}}{f_{\max} + f_{\min}} \, . \tag{3.1}$$

Modulation quantifies the relative amount by which the amplitude (or difference) $(f_{\max} - f_{\min})/2$ of $f(x,y)$ stands out from the average value (or background) $(f_{\max} + f_{\min})/2$. In general, m_f refers to the *contrast* of the periodic signal $f(x,y)$ relative to its average value. We assume here that $f(x,y)$ has nonnegative values, in which case $0 \leq m_f \leq 1$. Of importance, $m_f = 1$ only when $f_{\min} = 0$. Thus, in practice, the usual presence of a nonzero "background" intensity in a medical image reduces image contrast. If $m_f = 0$ (in which case, $f_{\min} = f_{\max}$), we say that $f(x,y)$ has no contrast. If $f(x,y)$ and $g(x,y)$ are two periodic signals with the same average value, we say that $f(x,y)$ has more contrast than $g(x,y)$ if $m_f > m_g$.

3.2.2 Modulation Transfer Function

The way a medical imaging system affects contrast can be investigated by imaging a *sinusoidal object* $f(x,y)$ of the form

$$f(x,y) = A + B\sin(2\pi u_0 x) \, , \tag{3.2}$$

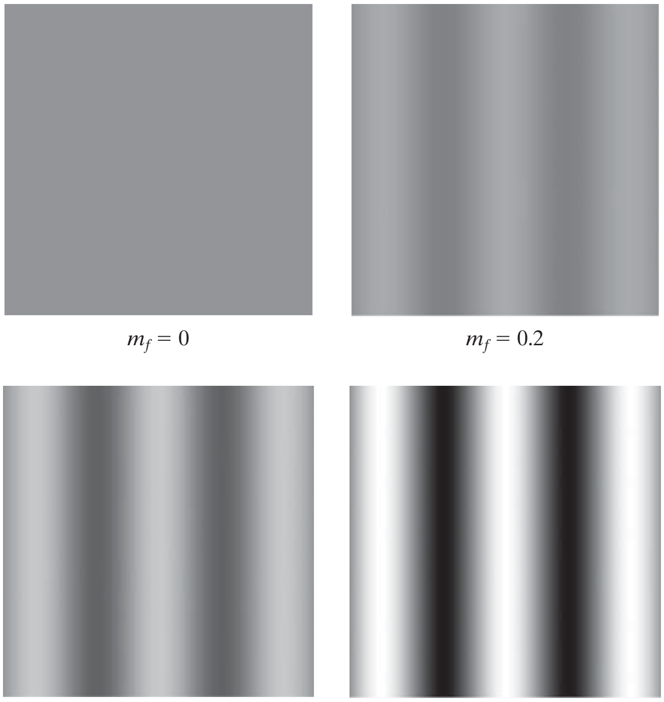
where A and B are two nonnegative constants such that $A \geq B$. This is a sinusoidal object that varies only in the x direction with spatial frequency u_0. Notice that $f_{\max} = A + B$ and $f_{\min} = A - B$, so the modulation of $f(x,y)$ is given by

$$m_f = \frac{B}{A} \, . \tag{3.3}$$

Figure 3.1 depicts four instances of $f(x,y)$ for the cases when $m_f = 0$, 0.2, 0.5, 1. Notice that as modulation increases, it becomes much easier to distinguish differences in shades of gray in the image of $f(x,y)$; in other words, contrast increases.

We are now interested in determining how an linear shift-invariant (LSI) imaging system with point spread function (PSF) $h(x,y)$ affects the modulation of $f(x,y)$; that is, we are interested in mathematically relating the modulation m_g of the output $g(x,y)$ to the modulation m_f of the input $f(x,y)$. In order to simplify our discussion, we assume that $h(x,y)$ is circularly symmetric [see (2.104)]. Since

$$f(x,y) = A + B\sin(2\pi u_0 x) = A + \frac{B}{2j}\left[e^{j2\pi u_0 x} - e^{-j2\pi u_0 x}\right] , \tag{3.4}$$



$m_f = 0$ $m_f = 0.2$

$m_f = 0.5$ $m_f = 1$

Figure 3.1
Four instances of a sinusoidal signal $f(x, y) = A + B\sin(2\pi u_0 x)$, with $m_f = B/A = 0, 0.2, 0.5, 1$.

the output $g(x, y)$ of the system is given by

$$g(x, y) = AH(0, 0) + B\,|H(u_0, 0)|\sin(2\pi u_0 x)\,. \tag{3.5}$$

Notice that the recorded image $g(x, y)$ of the sinusoidal object $f(x, y)$ is also sinusoidal with (the same) frequency u_0. From (3.5), $g_{\max} = AH(0, 0) + B|H(u_0, 0)|$ and $g_{\min} = AH(0, 0) - B|H(u_0, 0)|$, in which case the modulation of $g(x, y)$ is given by

$$m_g = \frac{B|H(u_0, 0)|}{AH(0, 0)} = m_f \frac{|H(u_0, 0)|}{H(0, 0)}\,. \tag{3.6}$$

The modulation m_g of $g(x, y)$ depends on the spatial frequency u_0.

The way an LSI medical imaging system affects modulation, and therefore contrast, is illustrated in Figure 3.2. The output modulation m_g is a scaled version of the input modulation m_f, the scaling factor being the magnitude spectrum $|H(u_0, 0)|$ of the medical imaging system under consideration. If $H(0, 0) = 1$ and $|H(u_0, 0)| < 1$, then $m_g < m_f$, and since both signals $f(x, y)$ and $g(x, y)$ have the same average value, the output $g(x, y)$ will have less contrast than the input $f(x, y)$.

The ratio of the output modulation to the input modulation as a function of spatial frequency is called the *modulation transfer function* (MTF), and [by rearranging (3.6)] is given by

$$\mathrm{MTF}(u) = \frac{m_g}{m_f} = \frac{|H(u, 0)|}{H(0, 0)}\,. \tag{3.7}$$

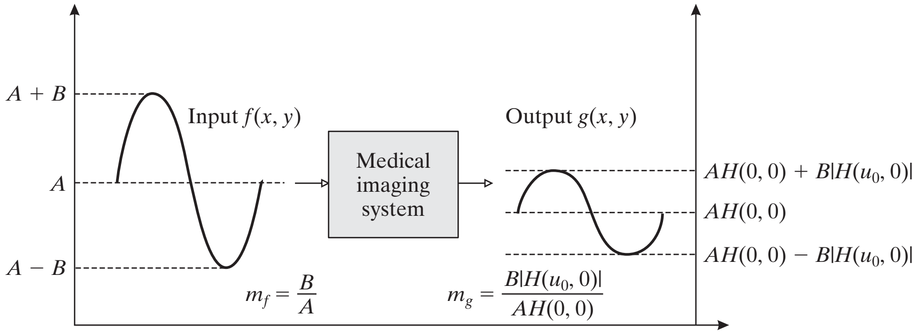


Figure 3.2
Basic principles for determining the modulation of the output of an LSI medical imaging system from the modulation of the input, when the input object is sinusoidal.

This shows that the MTF of a medical imaging system is, in essence, the "frequency response" of the system, and it can be directly obtained from the Fourier transform of the PSF of the system (remember that $H(u, v) = \mathcal{F}_{2D}\{h(x, y)\}$). Since $|H(u, 0)| = |H(-u, 0)|$, the MTF is usually considered only at nonnegative frequencies. Since the MTF, which characterizes contrast, can be mathematically related to the PSF, which characterizes blurring (or resolution), we can assume that blurring reduces contrast. We will have more to say about this in Section 3.3.3.

The MTF quantifies degradation of contrast as a function of spatial frequency. For most medical imaging systems,

$$0 \le \mathrm{MTF}(u) \le \mathrm{MTF}(0) = 1, \qquad \text{for every } u\,, \tag{3.8}$$

with the MTF becoming significantly less than unity (or even zero) at high spatial frequencies.

EXAMPLE 3.1

A typical MTF is depicted in Figure 3.3. Note that $0 \le \mathrm{MTF}(u) \le 1$, with the maximum value attained at $u = 0$. Note also that the MTF monotonically decreases to zero with increasing frequency, becoming zero at spatial frequencies larger than 0.8 mm^{-1}.

Question What can we learn about the contrast behavior of an imaging system with this MTF?

Answer At spatial frequency 0.6 mm^{-1}, the MTF takes a value of 0.5. This means that the contrast of a sinusoidal object at spatial frequency 0.6 mm^{-1} is reduced by half when imaged through this system. Moreover, since the MTF is zero at all spatial frequencies larger than 0.8 mm^{-1}, any sinusoidal input with frequency larger than 0.8 mm^{-1} will be imaged as a constant output of zero contrast. ■

As indicated above, we can think of loss of contrast as the result of the blurring action of a medical imaging system. This is illustrated in Figure 3.4, which depicts the outputs of three radiographic imaging systems with increasingly poorer MTFs. Here, a poorer MTF is one that drops to zero at lower spatial frequencies. A poorer MTF results in less contrast.

It should be noted here that the PSF of a medical imaging system need not be *isotropic*—that is, equivalent in all (2-D or 3-D) directions. In a nonisotropic

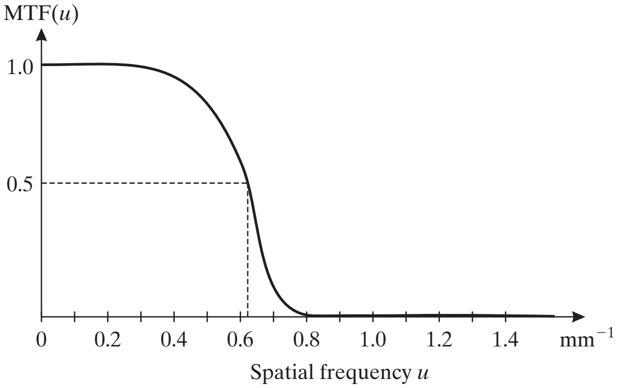


Figure 3.3
A typical MTF of a medical imaging system.

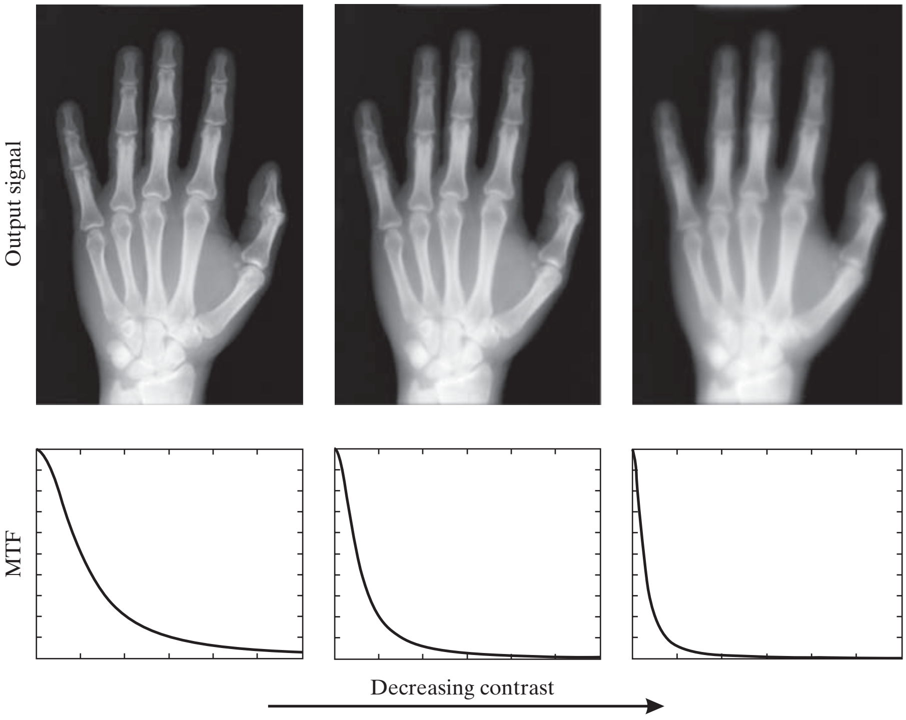


Figure 3.4
Impact of the MTF of a medical imaging system on contrast. Clearly, a poorer MTF results in lower contrast. Original image courtesy of GE Healthcare.

system, the profile through the PSF changes with orientation; thus, the system has an orientation-dependent response. In the nonisotropic (2-D) case, the MTF is rotationally dependent, and (3.7) can be generalized to

$$\mathrm{MTF}(u,v) = \frac{m_g}{m_f} = \frac{|H(u,v)|}{H(0,0)}\,, \tag{3.9}$$

in which case

$$m_g = m_f \frac{|H(u,v)|}{H(0,0)}\,. \tag{3.10}$$

For a typical nonisotropic medical imaging system,

$$0 \leq \mathrm{MTF}(u,v) = \frac{|H(u,v)|}{H(0,0)} \leq \mathrm{MTF}(0,0) = 1, \qquad \text{for every } u, v. \tag{3.11}$$

3.2.3 Local Contrast

The identification of some specific object or feature within an image is only possible if its value differs from that of surrounding areas. The definition of modulation or contrast for sinusoidal signals can be adapted for use in this situation as well. It is common in many imaging modalities (e.g., nuclear medicine) to refer to an object of interest (e.g., a tumor in the liver) as the *target*, as illustrated in Figure 3.5. Suppose that the target has a nominal image intensity of f_t. The target (i.e., the tumor) is surrounded by other tissues (i.e., the liver tissue), called the *background*, which may obscure our ability to see or detect the target. Suppose that the background has a nominal image intensity of f_b. The difference between the target and its background is captured by the *local contrast*, defined as

$$C = \frac{f_t - f_b}{f_b}. \tag{3.12}$$

The definition of local contrast in (3.12) differs from the definition of modulation in (3.1) for sinusoidal signals in that the intensities f_t and f_b may be selected locally—for example, within the liver—and they need not be the maximum and minimum intensities within the image as a whole. For example, f_t could be taken to be the average image intensity within the tumor, whereas f_b could be taken as the average image intensity within the liver. The tumor intensity could be larger than f_b or less than f_b. If f_t is less than f_b, then C is negative and, in certain circumstances, it may be convenient to report its absolute value.

EXAMPLE 3.2

Consider an image showing an organ with intensity I_o and a tumor with intensity $I_t > I_o$.

Question What is the local contrast of the tumor? If we add a constant intensity $I_c > 0$ to the image, what is the local contrast? Is the local contrast improved?

Answer By definition, the local contrast of the tumor is

$$C = \frac{I_t - I_o}{I_o}.$$

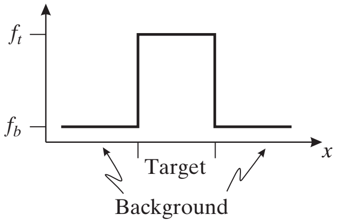


Figure 3.5
Local contrast scenario.

If we add a constant intensity I_c to the image, the intensities of the background and the target become $f_b = I_o + I_c$, and $f_t = I_t + I_c$. The local contrast of the processed image is

$$C' = \frac{(I_t + I_c) - (I_o + I_c)}{I_o + I_c} = \frac{I_t - I_o}{I_o + I_c} = C\frac{I_o}{I_o + I_c} < C.$$

So, the local contrast is worse if we add a constant intensity I_c to the image. ■

3.3 Resolution

Resolution is another basic measure of image quality. For our purposes, resolution can be thought of as the ability of a medical imaging system to accurately depict two distinct events in space, time, or frequency as separate. In this case, we talk about *spatial*, *temporal*, or *spectral resolution*, respectively. Resolution can also be thought of as the degree of smearing, or blurring, a medical imaging system introduces to a single event in space, time, or frequency. These two ways of looking at resolution are related, because the less smearing a system introduces, the closer in space, time, or frequency two events can be and still be distinguished as separate. Therefore, a high resolution medical imaging system is characterized by low smearing, whereas a low resolution system is characterized by high smearing. In this section, we mostly focus on spatial resolution, although a brief discussion of temporal and spectral resolution is also provided. Therefore, when we talk about “resolution,” we mean spatial resolution, unless otherwise specified.

3.3.1 Line Spread Function

As suggested above, we can think of resolution as the degree of smearing, or blurring, a medical imaging system introduces to a single event (e.g., a point) in space. This is the traditional PSF described in Chapter 2, and the response of an imaging system to a point source (i.e., a point impulse) is often used to characterize resolution. As an alternative, we can consider the response of a medical imaging system to a line impulse (see Section 2.2.2).

Consider an LSI medical imaging system with isotropic PSF $h(x, y)$ that is normalized to 1. Suppose that a line source, passing through the origin of the spatial domain, is imaged through the system. We will mathematically represent this line source by the *line impulse* $f(x, y) = \delta_\ell(x, y)$ [see (2.11)]. Since the system is isotropic, it is sufficient to consider the response to a vertical line through the origin; from (2.11), we see that for this case, $f(x, y) = \delta(x)$. Then, the output $g(x, y)$ of the system will be given by

$$\begin{aligned} g(x, y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\xi, \eta) f(x - \xi, y - \eta)\, d\xi\, d\eta \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} h(\xi, \eta)\delta(x - \xi)\, d\xi\right] d\eta \\ &= \int_{-\infty}^{\infty} h(x, \eta)\, d\eta\,, \end{aligned} \tag{3.13}$$

where, in the third equality, we have used the 1-D analog of (2.6).

The resulting image $g(x, y)$ is only a function of x, say $l(x)$. This is known as the *line spread function* (LSF) of the system under consideration, and it can be used to quantify resolution. This function is directly related to the PSF $h(x, y)$ since, from (3.13), we have

$$l(x) = \int_{-\infty}^{\infty} h(x, \eta)\, d\eta. \tag{3.14}$$

Notice that, since the PSF $h(x, y)$ is assumed to be isotropic, $l(x)$ is symmetric [i.e., $l(x) = l(-x)$], and since the PSF is normalized to 1,

$$\int_{-\infty}^{\infty} l(x)\, dx = 1. \tag{3.15}$$

Moreover, the 1-D Fourier transform $L(u)$ of the LSF $l(x)$ is related to the transfer function $H(u, v)$ of the system, since

$$\begin{aligned} L(u) &= \mathcal{F}_{1D}[l](u) \\ &= \int_{-\infty}^{\infty} l(x) e^{-j2\pi ux}\, dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, \eta) e^{-j2\pi ux}\, dx\, d\eta \\ &= H(u, 0). \end{aligned} \tag{3.16}$$

Therefore, the values of the transfer function along the horizontal line that passes through the origin of the frequency domain are adequate for determining the 1-D Fourier transform of the LSF, and hence the LSF itself. Since the PSF $h(x, y)$ is assumed to be isotropic, the transfer function is isotropic as well. Thus, the LSF is adequate for determining the PSF of the system. Indeed, from the LSF $l(x)$, we can calculate the 1-D Fourier transform $L(u)$, and $H(u, 0) = L(u)$. However, since the transfer function $H(u, v)$ is isotropic, the values of $H(u, v)$ along any line that passes through the origin of the frequency domain will be the same as the values of $H(u, 0)$.

3.3.2 Full Width at Half Maximum

Given the LSF (or the PSF) of a medical imaging system, its resolution can be quantified using a measure called the *full width at half maximum* (FWHM). This is the (full) width of the LSF (or the PSF) at one-half its maximum value. The FWHM is usually expressed in millimeters. Provided there is no geometric scaling, the FWHM equals the minimum distance that two lines (or points) must be separated in space in order to appear as separate in the recorded image. This is depicted in Figure 3.6, where we see the profiles of two points through a 1-D imaging system with PSF $h(x)$. The points move closer to each other as we go from (a) to (d). In a linear system, the observed profile is the sum of the individual profiles from the two points. In practice, a residual "dip" between the points must occur in the (summed) profile for the two points to be visualized or resolved as separate. The profile depicted in Figure 3.6(c) shows the separation

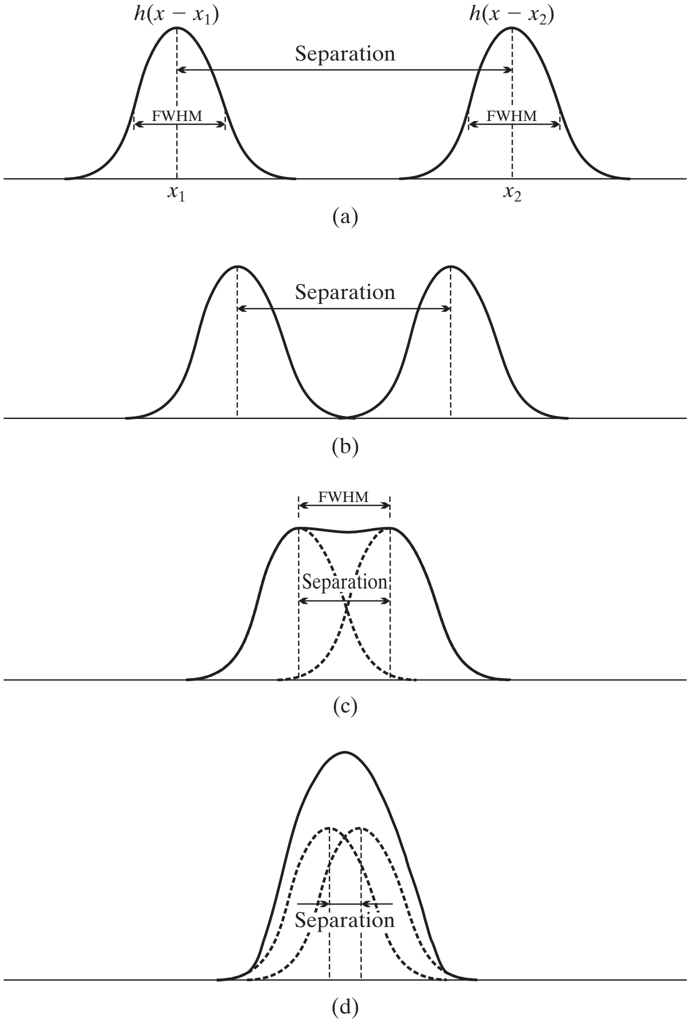


Figure 3.6
An example of the effect of system resolution on the ability to differentiate two points. The FWHM equals the minimum distance that the two points must be separated in order to be distinguishable.

distance at which the two points are just distinguishable. Notice that, in this case, the two points are separated by the FWHM. Therefore, a decrease in the FWHM indicates an improvement in resolution.

3.3.3 Resolution and Modulation Transfer Function

Another way to quantify the resolution of a medical imaging system is as the smallest separation (in mm) between two adjacent maxima (or minima) in a sinusoidal input that can be resolved in the image. Consider the case when the input of a medical imaging system is sinusoidal of the form $f(x, y) = B\sin(2\pi ux)$, with amplitude B and frequency u. From (3.5) and (3.7), the output of the system is given by

$$g(x, y) = \mathrm{MTF}(u)H(0, 0)B\sin(2\pi ux)\,. \tag{3.17}$$

Notice that the separation between two adjacent maxima (or minima) of the sinusoidal input $f(x, y)$ is $1/u$. The recorded image $g(x, y)$ is also sinusoidal, with $1/u$ being the separation between two adjacent maxima (or minima) as well. However, the amplitude of the output image equals the amplitude of the input multiplied by the MTF at spatial frequency u. In practice, $\text{MTF}(u) \neq 0$, for every $u \leq u_c$, and $\text{MTF}(u) = 0$, for every $u > u_c$, for some spatial cutoff frequency u_c, in which case, $g(x, y) = 0$, for every $u > u_c$. In this case, the resolution of the system will be $1/u_c$.

EXAMPLE 3.3

The MTF depicted in Figure 3.3 becomes zero at spatial frequencies larger than $0.8\ \text{mm}^{-1}$.

Question What is the resolution of this system?

Answer The resolution of a system with such an MTF is $1/(0.8\ \text{mm}^{-1}) = 1.25$ mm. Fine structures of an object to be imaged by such a system with spatial frequencies larger than $0.8\ \text{mm}^{-1}$ cannot be seen at the output of the system. ■

From our discussion, we find that the MTF can be effectively used to compare two competing medical imaging systems in terms of their contrast and resolution. If the MTFs of the two systems under consideration are of a similar shape but have a different cutoff frequency u_c, we can conclude that the system with higher MTF values will be better in terms of contrast and resolution. For example, the radiographic imaging system depicted in the first panel in Figure 3.4 is better in terms of contrast and resolution than the system depicted in the third panel.

If the MTF curves are of different shapes, the situation is more complicated. Figure 3.7 depicts two MTF curves that correspond to two competing medical imaging systems: SYSTEM 1 and SYSTEM 2. SYSTEM 1 has better low frequency contrast, and is thus better for imaging coarse details, while SYSTEM 2 has a better high frequency contrast, and is thus better for imaging fine details. Since contrast, as quantified by the MTF, is a function of spatial frequency, we can make this sort of frequency-by-frequency comparison. Spatial resolution, as described here, is not frequency-dependent, so it becomes harder to directly compare MTFs in the context of "better resolution." As described above, the FWHM of the PSF or LSF is the most direct metric of resolution. Further

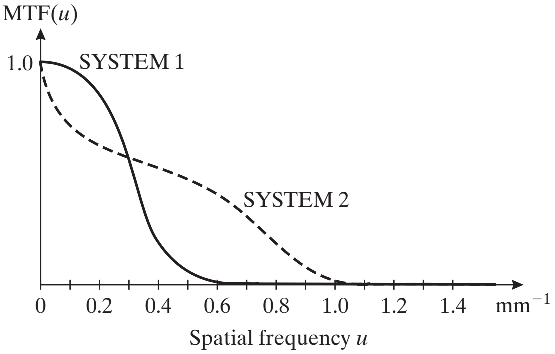


Figure 3.7
MTF curves of two competing medical imaging systems.

understanding of a system's resolution from its MTF usually comes from MTF values at higher spatial frequencies and the cutoff frequency u_c.

The MTF can be directly obtained from the LSF. Indeed, from (3.7) and (3.16),

$$\mathrm{MTF}(u) = \frac{|L(u)|}{L(0)}, \qquad \text{for every } u. \tag{3.18}$$

Therefore, the MTF equals the (normalized) magnitude of the 1-D Fourier transform of the LSF. The next example shows that this relationship can be used to determine the FWHM of a medical imaging system directly from its MTF.

EXAMPLE 3.4

Sometimes, the PSF, LSF, or MTF can be described by a mathematical function by either fitting observed data or by making simplifying assumptions about its shape. Assume that the MTF of a medical imaging system is given by

$$\mathrm{MTF}(u) = e^{-\pi u^2}\,. \tag{3.19}$$

Question What is the FWHM of this system?

Answer By using the 1-D inverse Fourier transform and the fact that $|L(u)| = \mathrm{MTF}(u)$, we have $l(x) = e^{-\pi x^2}$. The FWHM will then be given by FWHM $= 2x_0$, where x_0 is such that

$$e^{-\pi x_0^2} = \frac{1}{2}\,. \tag{3.20}$$

This results in FWHM $= 2\sqrt{\ln 2/\pi}$. ■

3.3.4 Subsystem Cascade

Medical imaging systems are often modeled as a cascade of LSI subsystems, as introduced in Section 2.3.4. Accordingly, the recorded image $g(x, y)$ can be modeled as the convolution of the input object $f(x, y)$ with the PSF of the first subsystem, followed by the convolution with the PSF of the second subsystem, etc. (see Section 2.3.4). For example, in the case of K subsystems with PSFs $h_1(x, y), h_2(x, y), \ldots, h_K(x, y)$,

$$g(x, y) = h_K(x, y) * \cdots * (h_2(x, y) * (h_1(x, y) * f(x, y)))\,. \tag{3.21}$$

The quality of the overall system, in terms of contrast and resolution, can be predicted by considering the quality of each subsystem.

If resolution is quantified using the FWHM, then the FWHM of the overall system can be determined approximately from the FWHMs $R_1, R_2, \ldots, R_K$ of the individual subsystems, by

$$R = \sqrt{R_1^2 + R_2^2 + \cdots + R_K^2}\,. \tag{3.22}$$

Notice that the overall FWHM R is dominated by the largest (i.e., the poorest resolution) term. Thus, small improvements in any given subsystem's resolution do not often yield noticeable improvements in overall system resolution.

The following example shows that, when a medical imaging system is composed of subsystems with Gaussian PSFs, (3.22) gives the exact value of the FWHM of the overall system.

EXAMPLE 3.5

Consider a 1-D medical imaging system with PSF $h(x)$ composed of two subsystems with Gaussian PSFs of the form

$$h_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{-x^2}{2\sigma_1^2}\right\} \quad \text{and} \quad h_2(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{\frac{-x^2}{2\sigma_2^2}\right\}. \tag{3.23}$$

Question What is the FWHM of this system?

Answer The FWHMs R_1 and R_2 associated with the two subsystems h_1 and h_2, respectively, are given by $R_1 = 2x_1$ and $R_2 = 2x_2$, where x_1 and x_2 are such that

$$h_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{-x_1^2}{2\sigma_1^2}\right\} = \frac{1}{2\sqrt{2\pi}\sigma_1}, \tag{3.24}$$

and

$$h_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{\frac{-x_2^2}{2\sigma_2^2}\right\} = \frac{1}{2\sqrt{2\pi}\sigma_2}. \tag{3.25}$$

After algebraic manipulation of (3.24) and (3.25),

$$R_1 = 2\sigma_1\sqrt{2\ln 2} \quad \text{and} \quad R_2 = 2\sigma_2\sqrt{2\ln 2}. \tag{3.26}$$

Following Example 2.4, it can be shown that the PSF $h(x)$ of the overall system is given by

$$h(x) = h_1(x) * h_2(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left\{\frac{-x^2}{2(\sigma_1^2+\sigma_2^2)}\right\}. \tag{3.27}$$

The FWHM R of this system is given by $R = 2x_0$, where x_0 is such that

$$h(x_0) = \exp\left\{\frac{-x_0^2}{2(\sigma_1^2+\sigma_2^2)}\right\} = 0.5, \tag{3.28}$$

from which we obtain

$$R = 2\sqrt{\sigma_1^2+\sigma_2^2}\,\sqrt{2\ln 2}. \tag{3.29}$$

■

If contrast and resolution are quantified using the MTF, then the MTF of the overall system will be given by

$$\text{MTF}(u,v) = \text{MTF}_1(u,v)\text{MTF}_2(u,v)\cdots\text{MTF}_K(u,v), \tag{3.30}$$

in terms of the MTFs $\mathrm{MTF}_k(u,v)$, $k = 1, 2, \ldots, K$, of the individual subsystems. This is a direct consequence of (3.9) and the fact that the frequency response $H(u,v)$ of the overall system is given by

$$H(u,v) = H_1(u,v)H_2(u,v)\cdots H_K(u,v), \tag{3.31}$$

which is the product of the frequency responses $H_k(u,v)$, $k = 1, 2, \ldots, K$, of the individual subsystems.

Figure 3.8 depicts the MTF curves of three subsystems of a medical imaging system together with the MTF curve of the overall system. If one subsystem has a small value of the MTF at some spatial frequency, then the MTF of the overall system will be small at that frequency as well. In other words, the MTF of the overall system will always be less than the MTF of each subsystem. This follows from (3.30), which suggests that, for every $k = 1, 2, \ldots, K$,

$$\mathrm{MTF}(u,v) \leq \mathrm{MTF}_k(u,v), \qquad \text{for every } u, v, \tag{3.32}$$

provided that $\mathrm{MTF}_k(u,v) \leq 1$, for every $k = 1, 2, \ldots, K$. Therefore, the overall quality of a medical imaging system, in terms of contrast and resolution, will be inferior to the quality of each subsystem.

From the previous discussion, the resolution of a medical imaging system can be specified by either its PSF or its LSF. In some medical imaging systems, we are able to create an object consisting of a very small "point" or "line" of some material. For example, we may create a very small and highly radioactive point or line source for measuring the resolution of a nuclear medicine camera. When a medical imaging system is mathematically modeled using an LSI system, its output to these objects will be the PSF and the LSF, respectively. This provides a practical way for calculating these two important response functions.

From our discussion, we can see that spatial resolution and image contrast are tightly linked, since the Fourier transforms of the PSF and the LSF, which are measures of resolution, yield the MTF, which is a measure of contrast. Indeed, spatial resolution can be thought of as the ability of an imaging system to preserve object contrast in the image, since blurring, due to poor resolution, is what actually reduces contrast.

In a nonisotropic system, the profile through the PSF changes with orientation; thus, the system has an orientation-dependent resolution. A good example

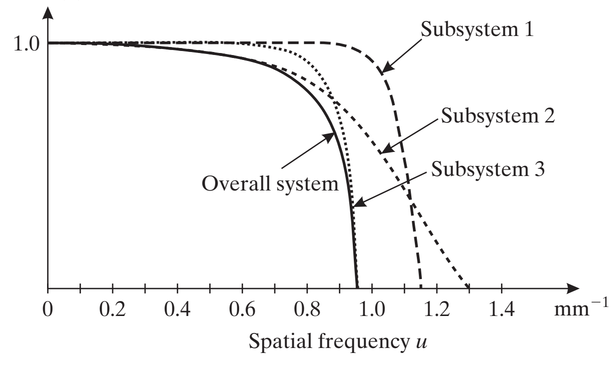


Figure 3.8
MTF curves of three subsystems of a medical imaging system and the MTF curve of the overall system.

of this situation is in ultrasound imaging systems where the *range resolution* (along the transducer axis) is usually substantially better than the *lateral resolution* (orthogonal to the transducer axis). (See Chapters 10 and 11 for more details.) In the nonisotropic case, the MTF is given by (3.9).

It is also possible for a medical imaging system to be linear but not shift-invariant. Then, the resolution is spatially dependent. For example, this situation exists in ultrasound systems because the acoustic energy spreads out with increasing distance from the transducer. This makes the FWHM increase with depth, which corresponds to a degradation of lateral resolution. Magnification in pinhole collimators, or focused collimators in nuclear medicine, also leads to spatially dependent resolution, in which case the image plane resolution is related, in a precise geometric way, to actual resolution within the object.

3.3.5 Resolution Tool

Images with better resolution are preferable to images with poorer resolution, since they contain more image details. Resolution can be quantified in terms of the ability of a system to image details of a given test pattern. For example, one common way to measure resolution for a particular system is to image the so-called *resolution tool* or *bar phantom*, such as the one depicted in Figure 3.9. This tool is composed of groups of parallel lines of a certain width, separated by gaps having the same width as the line width (yielding an overall duty cycle of 50 percent). Each group is characterized by the density of such lines, measured in *line pairs per millimeter* (lp/mm). The tool is imaged through the system under consideration, and system resolution is reported as the frequency (in lp/mm) of the finest line group that can be resolved at the output. For example, the resolution might be 6–8 lp/mm for a projection radiography system and 2 lp/mm for a CT scanner.

3.3.6 Temporal and Spectral Resolution

The preceding discussion on spatial resolution and the concepts embodied in the PSF, LSF, and FWHM apply equally well to temporal and spectral resolution.

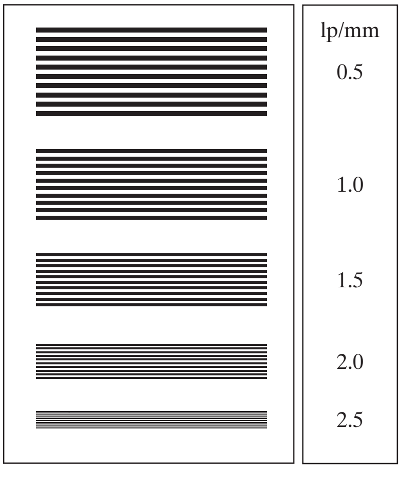


Figure 3.9
A resolution tool or bar phantom.

Temporal resolution is the ability to distinguish two events in time as being separate. *Spectral resolution* is the ability to distinguish two different frequencies (or, equivalently, energies). Conceptually, we could create a frequency distribution histogram of the number of observed events as a function of time or energy from a single-time or single-energy process (i.e., a point impulse in time or energy) to yield the equivalent of a PSF. As with the PSF, the ideal system response would be a delta function and the actual FWHM (in time or energy) would quantify the resolution. In other words, the concept of PSF applies equally well to events in time or frequency as it does to events in space.

3.4 Noise

An unwanted characteristic of medical imaging systems is *noise*. Noise is a generic term that refers to any type of random fluctuation in an image, and it can have a dramatic impact on image quality; image quality decreases as noise increases. The source and amount of noise depend on the imaging method used and the particular medical imaging system at hand. For example, in projection radiography, x-rays arrive at the detector in discrete packets of energy, called *quanta* or *photons*. The discrete nature of their arrival leads to random fluctuations, called *quantum mottle*, which give an x-ray image a textured or grainy appearance. On the other hand, in MRI, radio frequency pulses generated by nuclear spin systems are sensed by antennas connected to amplifiers. Since these signals have very low power, they compete with signals being generated in the antenna from natural thermal vibrations. Thermal vibrations are unpredictable—that is, random—and therefore comprise one source of noise in magnetic resonance images. The effect of increasing noise is shown in Figure 3.10.

The source of noise in a medical imaging system depends on the physics and instrumentation of the particular modality, subjects that will be developed in depth later in this book. The main objective of this section is to provide an introductory exposition of the tools used to mathematically characterize noise. A general way to characterize noise is to consider it as the numerical outcome of a random event or experiment. In nuclear medicine, for example, certain radioactive sources emit gamma ray photons that are received and recorded by a detector. Although governed by fixed physical properties (such as photon energy and decay rate), the specific nature of radioactive decay is random: photons are emitted at random times in random directions. We think of the noise as the deviation from a nominal value that would be predicted from

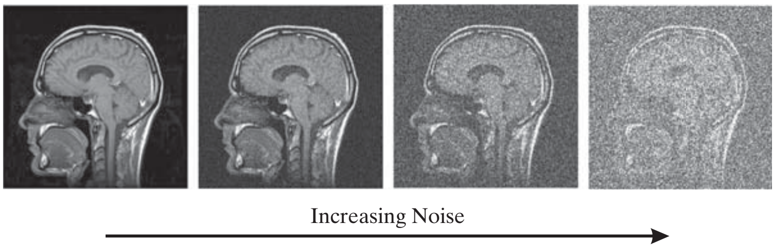


Figure 3.10
The effect of noise on image quality: image quality decreases rapidly with increasing noise contamination.

purely deterministic arguments. This deviation, arising from the random nature of radioactive emissions, accounts for the noise that is present in all nuclear medicine images.

3.4.1 Random Variables

The numerical quantity associated with a random event or *experiment* is called a *random variable*. Different repetitions of the experiment may produce different observed values—that is, the experiment has a random outcome. A random variable is mathematically described by $P_N(\eta)$, its *probability distribution function* (PDF), given by

$$P_N(\eta) = \Pr[N \leq \eta]\,, \tag{3.33}$$

where $\Pr[\cdot]$ denotes *probability*. The PDF gives the probability that random variable N will take on a value less than or equal to η. Notice that $0 \leq P_N(\eta) \leq 1$, $P_N(-\infty) = 0$, $P_N(\infty) = 1$, and $P_N(\eta_1) \leq P_N(\eta_2)$, for $\eta_1 \leq \eta_2$.

3.4.2 Continuous Random Variables

If $P_N(\eta)$ is a continuous function of η, then N is a *continuous* random variable. This random variable is uniquely specified by its *probability density function* (pdf),[1]

$$p_N(\eta) = \frac{dP_N(\eta)}{d\eta}\,. \tag{3.34}$$

Any pdf satisfies the following three properties:

$$p_N(\eta) \geq 0\,, \tag{3.35}$$

$$\int_{-\infty}^{\infty} p_N(\eta)\,d\eta = 1\,, \tag{3.36}$$

$$P_N(\eta) = \int_{-\infty}^{\eta} p_N(u)\,du\,. \tag{3.37}$$

In practice, the pdf of a random variable may not be known. Instead, a random variable is often characterized by its *expected value*

$$\mu_N = \mathrm{E}[N] = \int_{-\infty}^{\infty} \eta p_N(\eta)\,d\eta\,, \tag{3.38}$$

also called its *mean*, and its *variance*

$$\sigma_N^2 = \mathrm{Var}[N] = \mathrm{E}[(N-\mu_N)^2] = \int_{-\infty}^{\infty} (\eta-\mu_N)^2 p_N(\eta)\,d\eta\,, \tag{3.39}$$

---

[1] By convention, the abbreviation for probability density function is (lower case) pdf, and the abbreviation for probability distribution function is (upper case) PDF.

where E[·] and Var[·] denote expectation and variance, respectively. The square root σ_N of the variance is called the *standard deviation* of N.

The mean can be thought of as the average value of the random variable, whereas the standard deviation can be thought of as the "average" variation of the values of the random variable about its mean. The larger the standard deviation, the "more random" the random variable. As the standard deviation approaches zero, the observed values of the random variable more tightly cluster around the mean and, in the limit, the random variable becomes a constant that equals μ_N.

Uniform Random Variable A random variable N is said to be *uniform* over the interval $[a, b]$ if its pdf is of the form

$$p_N(\eta) = \begin{cases} \dfrac{1}{(b-a)}, & \text{for } a \leq \eta < b \\ 0, & \text{otherwise} \end{cases} . \tag{3.40}$$

In this case, the distribution function is given by

$$P_N(\eta) = \begin{cases} 0, & \text{for } \eta < a \\ \dfrac{\eta - a}{b - a}, & \text{for } a \leq \eta \leq b\,, \\ 1, & \text{for } \eta > b \end{cases} \tag{3.41}$$

and the expected value and variance are given by

$$\mu_N = \frac{a+b}{2} \qquad \text{and} \qquad \sigma_N^2 = \frac{(b-a)^2}{12}\,, \tag{3.42}$$

respectively.

Gaussian Random Variable If the pdf of a random variable N is given by

$$p_N(\eta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\eta-\mu)^2/2\sigma^2}\,, \tag{3.43}$$

then N is a *Gaussian* random variable. In this case, the distribution function is given by

$$P_N(\eta) = \frac{1}{2} + \operatorname{erf}\left(\frac{\eta - \mu}{\sigma}\right)\,, \tag{3.44}$$

where erf(x) denotes the *error function* given by the integral

$$\operatorname{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-u^2/2}\, du\,. \tag{3.45}$$

The expected value and variance are given by

$$\mu_N = \mu \qquad \text{and} \qquad \sigma_N^2 = \sigma^2\,, \tag{3.46}$$

respectively. The integral in (3.45) cannot be evaluated in closed form, but it is usually tabulated in mathematics handbooks and can be numerically evaluated using most mathematics, statistics, and engineering software packages.

EXAMPLE 3.6

Distribution functions can be found by integrating their corresponding density function using (3.37).

Question Can (3.44) be proved by direct integration?

Answer The distribution function of a Gaussian random variable with mean μ and variance σ^2 is

$$\begin{aligned}
P_N(\eta) &= \int_{-\infty}^{\eta} p_N(\tau)\, d\tau \\
&= \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\tau-\mu)^2/2\sigma^2}\, d\tau \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\frac{\eta-\mu}{\sigma}} e^{-t^2/2}\sigma\, dt, \quad \text{let } t = \tfrac{\tau-\mu}{\sigma} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-t^2/2}\, dt + \operatorname{erf}\left(\frac{\eta-\mu}{\sigma}\right) \\
&= \frac{1}{2} + \operatorname{erf}\left(\frac{\eta-\mu}{\sigma}\right).
\end{aligned}$$

To get the last equality, we use the fact that $\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ is the pdf of the standard Gaussian random variable and is symmetric around $t = 0$. ■

In general, the mean and variance do not uniquely specify a random variable. This means that, given μ_N and σ_N^2, there might be more than one pdf that produces the same mean and variance. However, in the case of a Gaussian random variable, the pdf is uniquely specified by its mean and variance.

Usually, noise in medical imaging systems is the result of a summation of a large number of independent noise sources. According to the *central limit theorem* of probability, a random variable that is the sum of a large number of independent causes tends to be Gaussian. Therefore, it is often natural to model noise in medical imaging system by means of a Gaussian random variable.

3.4.3 Discrete Random Variables

When the random variable N takes only values $\eta_1, \eta_2, \ldots, \eta_k$, it is said to be a *discrete random variable*. This random variable is uniquely specified by the *probability mass function* (PMF) $\Pr[N = \eta_i]$, for $i = 1, 2, \ldots, k$, where $\Pr[N = \eta_i]$ is the probability that random variable N will take on the particular value η_i. The PMF satisfies the following three properties:

$$0 \leq \Pr[N = \eta_i] \leq 1, \quad \text{for } i = 1, 2, \ldots, k\,, \tag{3.47}$$

$$\sum_{i=1}^{k} \Pr[N = \eta_i] = 1\,, \tag{3.48}$$

$$P_N(\eta) = \Pr[N \leq \eta] = \sum_{\text{all } \eta_i \leq \eta} \Pr[N = \eta_i]\,. \tag{3.49}$$

It is permissible for $k \to \infty$, meaning that there will be an infinite (but still countable) number of possible outcomes.

In the case of a discrete random variable, the mean value and variance are given by

$$\mu_N = \mathrm{E}[N] = \sum_{i=1}^{k} \eta_i \Pr[N = \eta_i]\,, \tag{3.50}$$

and

$$\sigma_N^2 = \mathrm{Var}[N] = \mathrm{E}[(N - \mu_N)^2] = \sum_{i=1}^{k} (\eta_i - \mu_N)^2 \Pr[N = \eta_i]\,, \tag{3.51}$$

respectively. Notice that the integrals in (3.38) and (3.39) have been replaced by sums in (3.50) and (3.51).

Poisson Random Variable Let N be a discrete random variable that takes values $0, 1, \ldots,$ and has the PMF

$$\Pr[N = k] = \frac{a^k}{k!} e^{-a}\,, \qquad \text{for } k = 0, 1, \ldots, \tag{3.52}$$

where $a > 0$ is a real-valued parameter. N is said to be a *Poisson random variable*, and it turns out that its mean equals its variance and

$$\mu_N = a\,, \tag{3.53}$$

$$\sigma_N^2 = a\,. \tag{3.54}$$

Poisson random variables play an important role in medical imaging systems, and most particularly in radiographic and nuclear medicine imaging. For example, they are used to statistically characterize the distribution of photons counted per unit area by an x-ray image intensifier, or to characterize the photon counts produced by a radiotracer in nuclear medicine.

EXAMPLE 3.7

In x-ray imaging, the Poisson random variable is used to model the number of photons that arrive at a detector in time t, which is a random variable referred to as a *Poisson process* and given the notation $N(t)$. The PMF of $N(t)$ is given by

$$\Pr[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

where λ is called the *average arrival rate* of the x-ray photons.

Question What is the probability that there is no photon detected in time t?

Answer The probability that there is no photon detected in time t is

$$\Pr[N(t) = 0] = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.$$

■

EXAMPLE 3.8

For the Poisson process of Example 3.7, the time that the first photon arrives is a random variable, say T.

Question What is the pdf $p_T(\tau)$ of random variable T?

Answer Assume that the first photon arrives in the interval $t < T < t + \Delta t$. For very small Δt, we have

$$\text{Prob}[t < T < t + \Delta t] \approx p_T(t)\Delta t.$$

For this to happen, it must be true that no photons arrived in the interval $[0, t]$ and exactly one photon arrived in the interval $[t, t + \Delta t]$. We learned from Example 3.7 that

$$\text{Prob}[\text{no photon detected in time } t] = e^{-\lambda t}.$$

The probability that exactly one photon will be detected in the interval $[t, t + \Delta t]$ is

$$\text{Prob}[\text{One photon detected in interval } [t, t + \Delta t]] = \frac{(\lambda \Delta t)^1}{1!} e^{-\lambda \Delta t} = \lambda \Delta t \, e^{-\lambda \Delta t}.$$

Using the Taylor series expansion for the exponential function

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

and the fact that Δt is very small (and will go to zero in the limit) permits the approximation

$$\text{Prob}[\text{One photon detected in interval } [t, t + \Delta t]] \approx \lambda \Delta t$$

where second-order terms in Δt have been dropped. Putting all this together yields

$$\begin{aligned} p_T(t)\Delta t &\approx \text{Prob}[\text{no photon detected in time } t] \\ &\quad \cdot \text{Prob}[\text{One photon detected in interval } [t, t + \Delta t]] \\ &\approx e^{-\lambda t} \cdot \lambda \Delta t \, . \end{aligned}$$

As $\Delta t \to 0$, all approximations get tighter, and so we recognize by dividing both sides of the above equation by Δt that

$$p_T(t) = \lambda e^{-\lambda t}, \qquad t \in [0, \infty).$$

The random variable T corresponding to this pdf is called the *exponential random variable.* ■

3.4.4 Independent Random Variables

It is usual in imaging experiments to consider more than one random variable at a time. The theory that is required in order to characterize a collection of random variables follows from the theory of a single random variable, with parallel definitions of distribution, density, and mass functions. In this book, however, we require only a simplified discussion related to *sums of independent random variables*. Loosely speaking, a collection of random variables are *independent* if knowledge of some of the random variables (i.e., making a partial observation) tells you nothing, statistically speaking, about the remaining random variables.

Consider the collection of random variables N_1, $N_2, \ldots N_m$, having the pdf's $p_1(\eta)$, $p_2(\eta), \ldots, p_m(\eta)$, respectively. The sum of these random variables S is another random variable having another pdf, $p_S(\eta)$. It is always the case that the mean of S is precisely the sum of the means of N_1, $N_2, \ldots, N_m$. That is,

$$\mu_S = \mu_1 + \mu_2 + \cdots + \mu_m \,, \tag{3.55}$$

where $\mu_1, \ldots, \mu_m$ are the means of the pdf's given above. Independence is not required for this to be true.

When the random variables are independent, we can go a step further. In this case, the variance of S is the sum of the individual variances,

$$\sigma_S^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_m^2 \,. \tag{3.56}$$

We emphasize that the variances, not the standard deviations, are added. Also, in the case of independence, it is possible to determine the pdf of S by

$$p_S(\eta) = p_1(\eta) * p_2(\eta) * \cdots * p_m(\eta) \,, \tag{3.57}$$

where $*$ is convolution. The facts that the variances add and that the pdf of the sum can be determined easily are powerful results following from the independence of random variables. We will use these facts to derive approximate noise and signal-to-noise ratio (SNR) expressions for CT imaging in Chapter 6.

EXAMPLE 3.9

Consider the sum S of two independent Gaussian random variables N_1 and N_2, each having a mean of zero and variance of σ^2.

Question What are the mean, variance, and pdf of the resulting random variable?

Answer The mean of S is zero since

$$\mu_S = \mu_1 + \mu_2 = 0 + 0 = 0 \,,$$

and the variance of S is $2\sigma^2$ since

$$\sigma_S^2 = \sigma_1^2 + \sigma_2^2 = \sigma^2 + \sigma^2 = 2\sigma^2 \,.$$

Using the fact that the convolution of two Gaussian waveforms yields a Gaussian waveform, we conclude that the pdf of S has a Gaussian form. Furthermore, since the Gaussian pdf is characterized completely by its mean and variance, we conclude that

$$p_S(\eta) = \frac{1}{\sqrt{2\pi\sigma_S^2}} \exp\left\{\frac{-\eta^2}{2\sigma_S^2}\right\} = \frac{1}{\sqrt{4\pi\sigma^2}} \exp\left\{\frac{-\eta^2}{4\sigma^2}\right\} .$$

■

3.5 Signal-to-Noise Ratio

In this book, we often assume that the output of a medical imaging system is a random variable G (or a collection of random variables), composed of two components, f and N. Component f, which is usually referred to as *signal*, is the (deterministic or nonrandom) "true" value of G, whereas N is a random fluctuation or error component due to noise. The identification of an abnormal condition within the human body most often depends on how "close" an observed value g of G, characteristic to that condition, is to its true value f.

A useful way to quantify this is by means of the *signal-to-noise ratio* (SNR). The SNR describes the relative "strength" of signal f with respect to that of noise N. Higher SNR values indicate that g is a more accurate representation of f, whereas lower SNR values indicate that g is less accurate. Therefore, higher image quality requires that the output of a medical imaging system be characterized by high SNR.

One way of thinking about the "signal" is that it is the modulation or contrast in the image, as discussed in Section 3.2, whereas "noise" is the unwanted, random fluctuations discussed in Section 3.4. As discussed above, blurring reduces contrast and thus SNR; noise also reduces SNR. Figure 3.11 depicts the effects of both blur and noise on SNR; as one moves from the upper left to the lower right, SNR decreases.

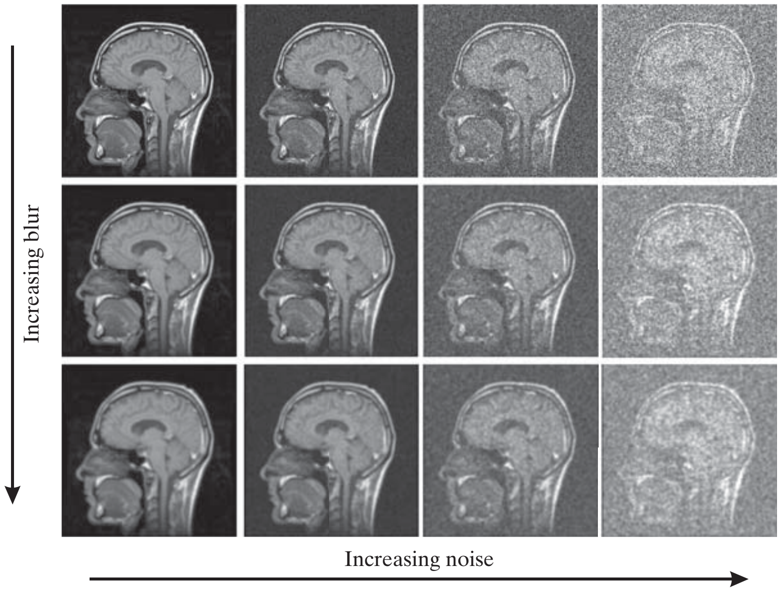


Figure 3.11
The effect of increasing blur and noise on SNR.

3.5.1 Amplitude SNR

Most frequently, the SNR is expressed as the ratio of signal amplitude to noise amplitude:

$$\mathrm{SNR_a} = \frac{\mathrm{Amplitude}(f)}{\mathrm{Amplitude}(N)}\,. \tag{3.58}$$

We refer to $\mathrm{SNR_a}$ as the *amplitude* SNR. Notice that there are many ways to specify what we mean by "signal" and "noise." Moreover, exact definition of the amplitude SNR depends on specifying what we mean by "signal amplitude" and "noise amplitude." Thus, the amplitude SNR is case-dependent, and its definition must be specifically adapted to the particular situation at hand.

EXAMPLE 3.10

In projection radiography, the number of photons G counted per unit area by an x-ray image intensifier follows a Poisson distribution as in (3.52). In this case, we may consider signal f to be the average photon count per unit area (i.e., the mean of G) and noise N to be the random variation of this count around the mean, whose amplitude is quantified by the standard deviation of G.

Question What is the amplitude SNR of such a system?

Answer From (3.52) to (3.54), it follows that the amplitude SNR is given by

$$\mathrm{SNR_a} = \frac{\mu_G}{\sigma_G} = \frac{\mu}{\sqrt{\mu}} = \sqrt{\mu}\,. \tag{3.59}$$

This quantity is known as the *intrinsic SNR* of x-rays, and we will have much more to say about this in Part II of the book. This tells us something very important and practical: The greater the average number μ of photons, the larger the amplitude SNR, and the smaller the *relative* amplitude of random fluctuations in G. Therefore, a higher x-ray exposure generally improves the quality of radiographic images. Keep in mind, however, that greater exposure to ionizing radiation means more risk of radiogenic cancer. ■

3.5.2 Power SNR

Another way to express the SNR is as the ratio of signal power to noise power:

$$\mathrm{SNR_p} = \frac{\mathrm{power}(f)}{\mathrm{power}(N)}\,. \tag{3.60}$$

We refer to $\mathrm{SNR_p}$ as the *power SNR*. Notice that exact definition of the power SNR depends on specifying what we mean by "signal power" and "noise power." Thus, as with the amplitude SNR, the power SNR is case-dependent, and its definition must be specifically adapted to the particular situation at hand.

EXAMPLE 3.11

If $f(x,y)$ is the input to a noisy medical imaging system with PSF $h(x,y)$, then the output at (x,y) may be thought of as a random variable $G(x,y)$ composed of signal $h(x,y)*f(x,y)$ and noise $N(x,y)$, with mean $\mu_N(x,y)$ and variance $\sigma_N^2(x,y)$.

Question What is the power SNR of such a system?

Answer In this framework, the power SNR at the output of this system is given by

$$\mathrm{SNR_p} = \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y) * f(x,y)|^2 \, dx \, dy}{\sigma_N^2}, \tag{3.61}$$

where the noise power is specified by the variance σ_N^2. Usually, it is assumed that there is no correlation between noise values and that for every (x, y)

$$\mu_N(x,y) = 0 \qquad \text{and} \qquad \sigma_N(x,y) = \sigma_N. \tag{3.62}$$

This type of noise is known as *white noise*. ■

White noise is a crude approximation of reality, but it is convenient to use since it leads to mathematically simple models. Keep in mind that, in most cases, it is more accurate to consider correlated noise and assume some mathematical expression for such correlation. In this case, if we assume that the noise mean and variance do not depend on (x, y) (a common assumption that characterizes so-called *wide-sense stationary noise*), then it can be shown that

$$\mathrm{SNR_p} = \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y) * f(x,y)|^2 \, dx \, dy}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{NPS}(u,v) \, du \, dv}, \tag{3.63}$$

where

$$\begin{aligned}\mathrm{NPS}(u,v) = \lim_{x_0,y_0\to\infty} \frac{1}{4x_0y_0} \mathrm{E}\Bigg[\Bigg|\int_{-x_0}^{x_0}\int_{-y_0}^{y_0} &[N(x,y) - \mu_N] \\ &\exp\left(-j2\pi(ux+vy)\right) \, dx \, dy\Bigg|^2\Bigg],\end{aligned} \tag{3.64}$$

is known as the *noise power spectrum* (NPS). From (3.63) and Parseval's theorem, given by (2.96), we have

$$\mathrm{SNR_p} = \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |H(u,v)|^2 \; |F(u,v)|^2 \, du \, dv}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{NPS}(u,v) \, du \, dv} \tag{3.65}$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{SNR_p}(u,v)\mathrm{NPS}(u,v) \, du \, dv}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{NPS}(u,v) \, du \, dv}, \tag{3.66}$$

where [see also (3.9)]

$$\mathrm{SNR_p}(u,v) = \frac{|H(u,v)|^2|F(u,v)|^2}{\mathrm{NPS}(u,v)} = \frac{\mathrm{MTF}^2(u,v)}{\mathrm{NPS}(u,v)}\,|F(u,v)|^2H^2(0,0) \tag{3.67}$$

is called the *frequency-dependent power SNR*. The frequency-dependent power SNR quantifies, at a given frequency, the relative "strength" of signal to that of noise at the output of the LSI system under consideration. From (3.66) and (3.67), $\mathrm{SNR_p}(u,v)$ provides a relationship between contrast, resolution, noise, and image quality. For a given output noise level (i.e., a fixed NPS) and a given input $f(x,y)$, better contrast and resolution properties (i.e., a larger MTF) result in better image quality (i.e., a higher output power SNR).

3.5.3 Differential SNR

Consider an object (or target) of interest placed on a background. Let f_t and f_b be the average image intensities within the target and background, respectively. A useful choice for SNR is obtained by taking the "signal" to be the difference in average image intensity values between the target and the background integrated over the area A of the target, and by taking the "noise" to be the random fluctuation of image intensity from its mean over an area A of the background. This leads to the *differential signal-to-noise ratio* ($\mathrm{SNR_{diff}}$), given by

$$\mathrm{SNR_{diff}} = \frac{A(f_t - f_b)}{\sigma_b(A)}\,, \tag{3.68}$$

where $\sigma_b(A)$ is the standard deviation of image intensity values from their mean over an area A of the background. From (3.12), we have

$$\mathrm{SNR_{diff}} = \frac{CAf_b}{\sigma_b(A)}\,, \tag{3.69}$$

which relates the differential SNR to contrast.

EXAMPLE 3.12

Consider the case of projection radiography. We may take f_b to be the average photon count per unit area in the background region around a target, in which case, $f_b = \lambda_b$, where λ_b is the mean of the underlying Poisson distribution governing the number of background photons counted per unit area. Notice that, in this case, $\sigma_b(A) = \sqrt{\lambda_b A}$.

Question What is the average number of background photons counted per unit area, if we want to achieve a desirable differential SNR?

Answer From (3.69),

$$\mathrm{SNR_{diff}} = \frac{CA\lambda_b}{\sqrt{A\lambda_b}} = C\sqrt{A\lambda_b}\,. \tag{3.70}$$

From (3.70), the differential SNR is approximately proportional to contrast as well as to the square root of the object area multiplied by radiation exposure (characterized by the average photon count per unit area λ_b). To achieve a desirable differential SNR, it is required that

$$\lambda_b = \frac{\mathrm{SNR}_{\mathrm{diff}}^2}{C^2 A}\,. \tag{3.71}$$

This relationship was first suggested by Albert Rose and is known as the *Rose model*. Clearly, to maintain good image quality (i.e., to obtain images with high SNR), high radiation dose is required when viewing small, low-contrast objects. ■

3.5.4 Decibels

The SNR is sometimes given in *decibels* (dB). When the SNR is the ratio of amplitudes, such as with the amplitude SNR or the differential SNR, then

$$\text{SNR (in dB)} = 20 \times \log_{10} \text{SNR (ratio of amplitudes)}\,. \tag{3.72}$$

When the SNR is the ratio of powers, such as with the power SNR, then

$$\text{SNR (in dB)} = 10 \times \log_{10} \text{SNR (ratio of powers)}\,. \tag{3.73}$$

3.6 Sampling

To electronically sense, store, and process continuous signals using computers, we must transform them into collections of numbers. This transformation, called *discretization* or *sampling*, means that we only retain representative signal values and discard the rest. There are many ways to do this. In this book, we focus our attention on the so-called *rectangular sampling* scheme. According to this scheme, a 2-D continuous signal is replaced by a discrete signal whose values are the values of the continuous signal at the vertices of a 2-D rectangular grid. More precisely, given a 2-D continuous signal $f(x, y)$, rectangular sampling generates a 2-D discrete signal $f_d(m, n)$, such that

$$f_d(m, n) = f(m\Delta x, n\Delta y), \qquad \text{for } \ m, n = 0, 1, \ldots. \tag{3.74}$$

In (3.74), Δx and Δy are the *sampling periods* in the x and y directions, respectively. This is illustrated in Figure 3.12. Notice that $f_d(m, n)$ forms an array of numbers that contains the values of the 2-D continuous signal $f(x, y)$ at the discrete points $(m\Delta x, n\Delta y)$. The inverses $1/\Delta x$ and $1/\Delta y$ of Δx and Δy are referred to as the *sampling frequencies* in the x and y direction, respectively. As shown below, these sampling frequencies are related to the comb and sampling functions presented in Chapter 2.

Ideally, we would like to keep as few samples as possible. This strategy minimizes the number of detectors (or the scanning time needed by a single detector) and reduces signal storage and subsequent processing requirements. So we ask: Given a 2-D continuous signal $f(x, y)$, what are the maximum possible values for Δx and Δy such that $f(x, y)$ can be reconstructed from the 2-D discrete signal $f_d(m, n)$, obtained by (3.74)? As illustrated in Figure 3.12, we may be tempted to perform *coarse sampling* as opposed to *fine sampling*. However, if

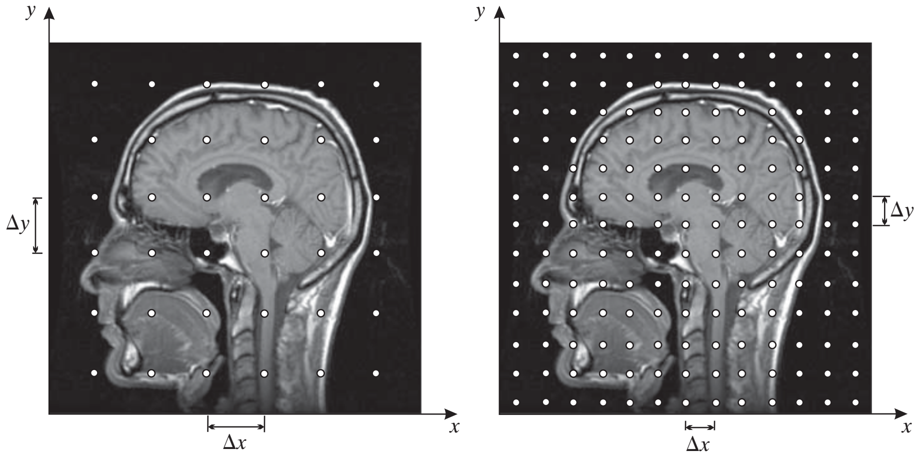


Figure 3.12
Illustration of coarse (left) and fine (right) rectangular sampling schemes. Although coarse sampling results in fewer samples, it may not allow reconstruction of the original continuous signal from these samples.

we give in to this temptation, we will end up (in the extreme case) with only one sample, which is clearly not enough to represent a continuous signal in general (unless the signal is constant).

Sampling a continuous signal with too few samples results in a type of signal corruption called *aliasing*, in which higher frequencies "take the alias of" lower frequencies. A signal sampled with too few samples cannot be reconstructed from its discrete representation; the best possible continuous reconstruction from these samples will be corrupted by aliasing. The visual appearance of this artifact depends on the signal's spectrum, but generally it appears as new high-frequency patterns where none should exist. The spatial frequency of these patterns is always lower than it should be but may be relatively high compared with other spatial information in the image. An example of this phenomenon is depicted in Figure 3.13. The improperly sampled image, as shown in Figure 3.13(b), is corrupted by a high-frequency texture that is not present in the original image. When proper anti-aliasing is used, as described in Section 3.6.3, sampling will produce an image with reduced resolution but no aliasing artifacts, as shown in Figure 3.13(c). Obviously, it is vital that medical imaging systems sample without aliasing.

3.6.1 Signal Model for Sampling

It is intuitive to believe that a slowly varying signal could be reconstructed from fewer samples than a rapidly varying signal, which may require finer sampling at regions of rapid signal variation. Since signal variation in space is directly associated with frequency content, we suspect that there must be a direct relationship between the appropriate values of sampling periods Δx and Δy and the frequency content of the signal under consideration. Here, we prove this intuition mathematically, and derive a recipe for choosing appropriate values for Δx and Δy. We do this by using the comb function $\text{comb}(x, y)$, given by (2.13), and its close relative, the sampling function $\delta_s(x, y; \Delta x, \Delta y)$, given by (2.14).

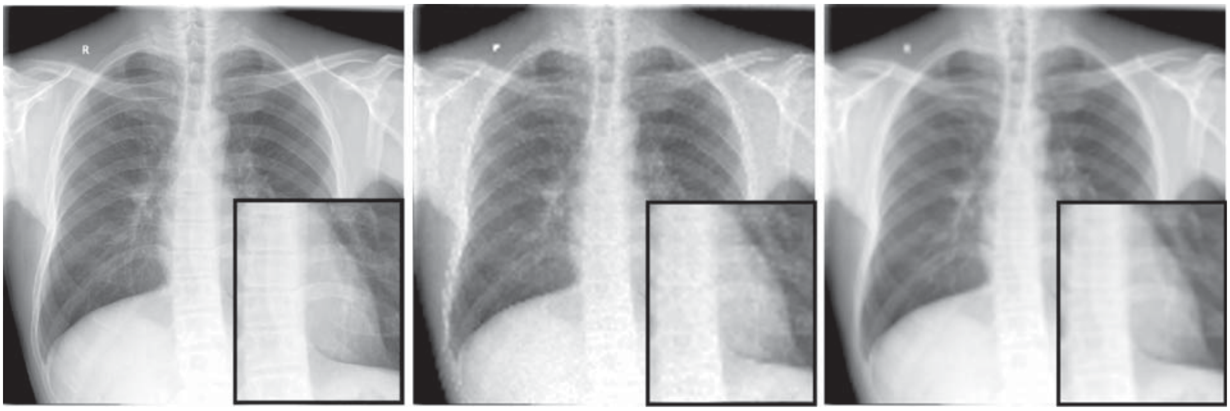

(a) (b) (c)

Figure 3.13
(a) Original chest x-ray image and sampled images, (b) without, and (c) with anti-aliasing. Inserts show a zoomed region.

Consider multiplying a continuous signal $f(x, y)$ by the sampling function. We have

$$\begin{aligned} f_s(x, y) &= f(x, y)\delta_s(x, y;\, \Delta x, \Delta y) \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(x, y)\delta(x - m\Delta x, y - n\Delta y) \quad \text{[from (2.14)]} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m\Delta x, n\Delta y)\delta(x - m\Delta x, y - n\Delta y) \quad \text{[from (2.7)]} \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f_d(m, n)\delta(x - m\Delta x, y - n\Delta y)\,, \end{aligned} \tag{3.75}$$

where we have used (3.74) for the final step. It follows that, given the discrete signal $f_d(m, n) = f(m\Delta x, n\Delta y)$, we can calculate the continuous signal $f_s(x, y)$ regardless of the sampling periods Δx and Δy. Thus, if we can reconstruct $f(x, y)$ from $f_s(x, y)$, then $f(x, y)$ can also be reconstructed from $f_d(m, n)$. The most important point here is that, in order to understand the effects of sampling, we need only look at the *continuous* signal $f_s(x, y)$ and its relationship to $f(x, y)$.

Since $f_s(x, y)$ is the product of two functions, its Fourier transform is the convolution of the Fourier transforms of the two functions. This is a consequence of the product property (2.92) of the Fourier transform. Therefore,

$$\begin{aligned} F_s(u, v) &= F(u, v) * \operatorname{comb}(u\Delta x, v\Delta y) \\ &= F(u, v) * \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(u\Delta x - m, v\Delta y - n) \quad \text{[from (2.13)]} \end{aligned}$$

$$= \frac{1}{\Delta x \Delta y} F(u, v) * \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(u - m/\Delta x, v - n/\Delta y) \quad \text{[from (2.8)]}$$

$$= \frac{1}{\Delta x \Delta y} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} F(u, v) * \delta(u - m/\Delta x, v - n/\Delta y)$$

$$= \frac{1}{\Delta x \Delta y} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\xi, \eta) \delta(u - m/\Delta x - \xi, v - n/\Delta y - \eta) \, d\xi \, d\eta \right]$$

$$= \frac{1}{\Delta x \Delta y} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} F(u - m/\Delta x, v - n/\Delta y) \quad \text{[from (2.6)]}, \qquad (3.76)$$

where we have used the fact that the Fourier transform of the sampling function is given by

$$\mathcal{F}_{2D}(\delta_s(x, y; \Delta x, \Delta y)) = \text{comb}(u\Delta x, v\Delta y), \qquad (3.77)$$

as indicated in Table 2.1.

From (3.76), it is clear that the spectrum $F_s(u, v)$ of $f_s(x, y)$ is calculated by shifting the spectrum $F(u, v)$ of $f(x, y)$ to locations $(m/\Delta x, n/\Delta y)$, for all m and n, adding all shifted spectra and dividing the result by $\Delta x \Delta y$. This is illustrated in Figure 3.14. If the shifted spectra in $F_s(u, v)$ do not overlap, then the original spectrum $F(u, v)$ of $f(x, y)$, and thus $f(x, y)$ itself, can be recovered by filtering $f_s(x, y)$ to "pick" one of the (equivalent) spectra. It is customary to use a low-pass filter [see (2.102)] to capture the spectrum centered at the origin, like the one outlined with a gray box in Figure 3.14(b).

3.6.2 Nyquist Sampling Theorem

In order that the spectra in $F_s(u, v)$ do not overlap, it is first necessary that the spectrum of $f(x, y)$ be zero outside a rectangle in frequency space. Such signals are called *band-limited*. If the highest frequencies present in $f(x, y)$ in the x and y directions are U and V, respectively, then if

$$\Delta x \leq \frac{1}{2U} \qquad \text{and} \qquad \Delta y \leq \frac{1}{2V}, \qquad (3.78)$$

the spectrum $F(u, v)$ can be reconstructed from the spectrum $F_s(u, v)$, in which case $f(x, y)$ can be reconstructed from $f_s(x, y)$ and thus from the samples $f_d(m, n)$.

If $\Delta x > 1/2U$ or $\Delta y > 1/2V$, there is overlap of "high" frequencies of $F(u, v)$ in $F_s(u, v)$, producing aliasing. Aliasing is illustrated in Figure 3.14(c). In this case, the spectrum $F(u, v)$ cannot be recovered from the spectrum $F_s(u, v)$ and, therefore, $f(x, y)$ cannot be reconstructed from its samples $f_d(m, n)$.

In summary, we have the following important result: A 2-D continuous band-limited signal $f(x, y)$, with cutoff frequencies U and V, can be uniquely

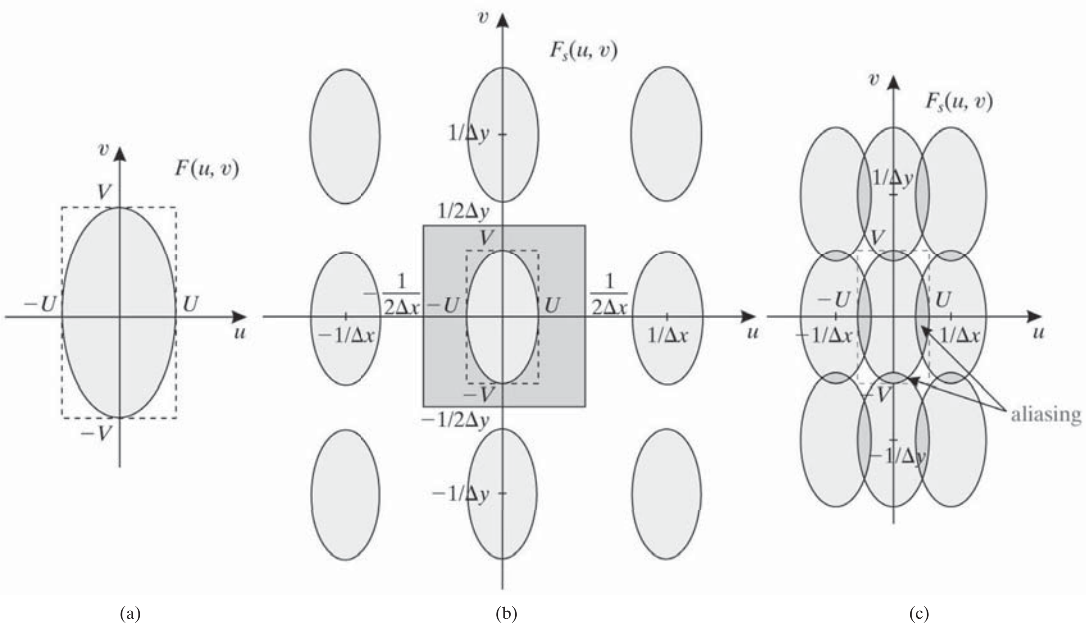


Figure 3.14
(a) The spectrum $F(u, v)$ of a band-limited continuous signal $f(x, y)$ with cutoff frequencies U and V. (b) The spectrum $F_s(u, v)$ of signal $f_s(x, y)$ obtained by sampling $f(x, y)$ with sampling periods $\Delta x < 1/2U$ and $\Delta y < 1/2V$. In this case, $F(u, v)$ can be perfectly reconstructed from $F_s(u, v)$. The spectrum $F(u, v)$ takes value one within the light gray area and zero outside. (c) The spectrum $F_s(u, v)$ of $f_s(x, y)$ obtained by sampling $f(x, y)$ with sampling periods $\Delta x > 1/2U$ and $\Delta y > 1/2V$. In this case, $f_s(x, y)$ experiences aliasing and $F(u, v)$ cannot be reconstructed perfectly in general.

determined from its samples $f_d(m, n) = f(m\Delta x, n\Delta y)$, if and only if the sampling periods Δx and Δy satisfy

$$\Delta x \leq \frac{1}{2U} \qquad \text{and} \qquad \Delta y \leq \frac{1}{2V}\,.$$

This is known as the *sampling theorem* (or the *Nyquist sampling theorem*, after its discoverer). To avoid aliasing, the maximum allowed values for Δx and Δy are given by

$$(\Delta x)_{\max} = \frac{1}{2U} \qquad \text{and} \qquad (\Delta y)_{\max} = \frac{1}{2V}\,, \tag{3.79}$$

and are known as the *Nyquist sampling periods*.

From our previous discussion, aliasing-free sampling requires band-limited continuous signals. From (3.79), the minimum number of samples required for aliasing-free sampling is directly proportional to the cutoff frequencies U and V. This verifies the fact that slowly varying signals, which are characterized by small values of U and V, require fewer samples than rapidly varying signals, which are characterized by large values of U and V. From the preceding discussion, we can see why undersampled signals tend to show high-frequency artifacts: The overlap of high-frequency spectra in the aliased Fourier transform artificially boosts high-frequency content.

3.6.3 Anti-Aliasing Filters

In medical imaging systems, there is an inherent trade-off between the number of samples acquired by a detector and image quality. Acquiring a large number of samples typically produces the highest image resolution but is almost always either expensive or time-consuming. Reducing the number of samples may lead to aliasing, which will introduce unwanted artifacts into the image. An alternative is to first filter the continuous signal using a low-pass filter, and then sample with fewer samples. In this case, the image is degraded by blurring rather than aliasing artifacts, which is usually preferable. Such a low-pass filter is called an *anti-aliasing filter*, and must be applied before sampling.

Anti-aliasing filtering is often an inherent part of a medical imaging instrument because the true continuous signal is blurred by the impulse response of the system. It is useful to consider two sources of blurring, that caused by the inherent physics and geometry of the system and that caused by the detectors themselves. Ideally, the physics and geometry cause only a small amount of blurring and therefore a high frequency signal is presented to the detectors. Although this situation is highly dependent on the modality, it would be fairly accurate to say that very little anti-aliasing is accomplished by the physics and geometry of the system.

Anti-aliasing is more often accomplished by the detectors, which (in most cases) integrate the incident signal over the area of each detector. This additional integration, which is still part of the overall PSF of the system, provides additional low-pass filtering prior to sampling. Since the detectors are not point samplers, it is incorrect to model the discretization of a medical imaging system as the multiplication of the incident signal $f(x,y)$ with the sampling function $\delta_s(x,y;\,\Delta x,\Delta y)$. Instead, the effect of detector integration can be modeled by first convolving $f(x,y)$ with the PSF $h(x,y)$ of the detector. The overall detector system PSF is the result of the cascade of PSFs that characterize the individual components of resolution, including the detector's geometric resolution. Here we focus on a specific component of resolution: any sampling process that is not point sampling. In many cases, the sampling is *area sampling* rather than point sampling. This occurs when the detector is either intrinsically digital (i.e., it is a pixelated detector) or its output is represented by an array of pixels. In either case, the PSF of the area sampling component can be represented by a rect function [(2.16)], and the sampling process can be mathematically modeled by

$$f_d(m,n) = f_s(m\Delta x, n\Delta y)\,, \tag{3.80}$$

$$f_s(x,y) = [h(x,y) * f(x,y)]\,\delta_s(x,y;\,\Delta x,\Delta y)\,. \tag{3.81}$$

Since integration is a low-pass process, this resembles the anti-aliasing filtering approach described above: The continuous signal $f(x,y)$ is first filtered by a low-pass filter with PSF $h(x,y)$, from which $f_d(m,n)$ is obtained by sampling.

Note that the part of the overall PSF due to the other individual components of resolution can be included in this analysis simply by considering $h(x,y)$ to be the overall PSF of the system rather than just the PSF of the area sampling component.

EXAMPLE 3.13

Consider a medical imaging system with sampling period Δ in both the x and y directions.

Question What is the highest frequency allowed in the images so that the sampling is free of aliasing? If an anti-aliasing filter, whose PSF is modeled as a rect function, is used and we ignore all the side lobes of its transfer function, what are the widths of the rect function?

Answer From the Nyquist sampling theorem, we know that if the image contains frequency higher than $1/2\Delta$, then there will be aliasing. So the highest allowed frequency is $1/2\Delta$.

If an anti-aliasing filter is used, the highest frequency of the image being sampled is the cutoff frequency of the filter (here, we assume the transfer function of the filter has value 0 outside the cutoff frequency). In this example, the anti-aliasing filter is modeled as a rect function. Its transfer function is given by the sinc function (we ignore the magnitude of the rect function, since it does not change the cutoff frequency),

$$H(u,v) = \mathrm{sinc}(\Delta_x u, \Delta_y v)\,,$$

where Δ_x and Δ_y are the widths of the filter in the x and y directions, respectively.

If we ignore the side lobes of $H(u,v)$, the cutoff frequencies of the filter are the first zeros of the sinc function, which are $1/\Delta_x$ and $1/\Delta_y$. For the given sampling periods Δ, we must have

$$\frac{1}{\Delta_x} \le \frac{1}{2\Delta} \qquad \text{and} \qquad \frac{1}{\Delta_y} \le \frac{1}{2\Delta},$$

which is equivalent to

$$\Delta_x \ge 2\Delta \text{ and } \Delta_y \ge 2\Delta.$$

■

Figure 3.13 shows the impact of sampling with and without anti-aliasing. When the sampling period is too small, there is spectral overlap and the resulting image has artifacts that are most typically seen as high-frequency patterns, as are visible in Figure 3.13(b). Anti-aliasing low-pass filters the image before sampling, making the image blurrier but devoid of artifacts, as is apparent in Figure 3.13(c).

3.7 Other Effects

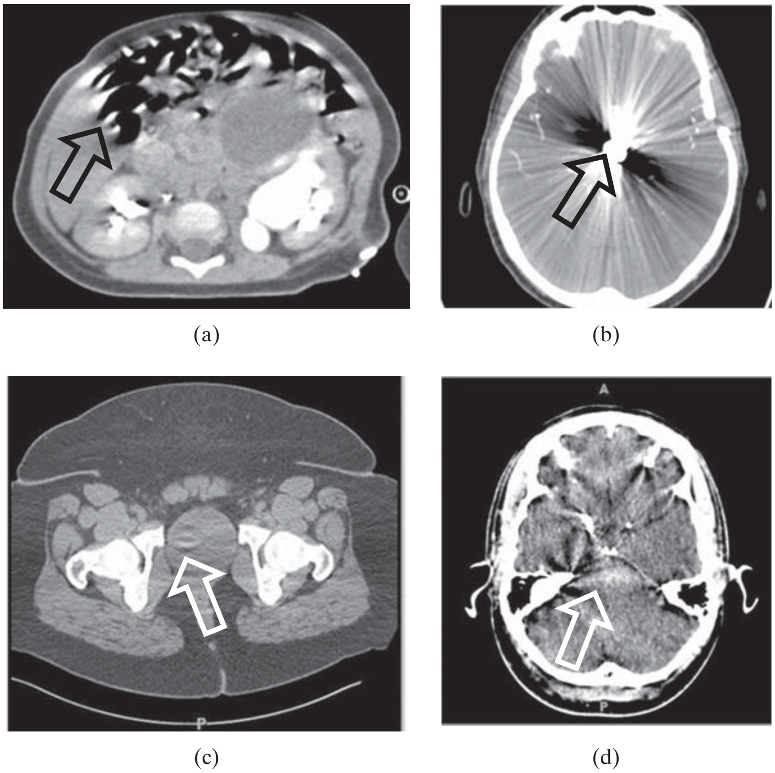
3.7.1 Artifacts

A problem that frequently affects image quality is the creation of image features known as *artifacts* that do not represent valid anatomical or functional objects. Artifacts can obscure important targets, and they can be falsely interpreted as valid image features. Moreover, they can impair correct detection and characterization of features of interest by adding "clutter" to images.

Artifacts are caused by a variety of reasons and can appear at any step of the imaging process. For example, in projection radiography, artifacts can be generated by the x-ray source, by restricting the x-ray beam in order to avoid exposing parts of a patient that need not be imaged, and by nonuniformities in the x-ray image intensifier over the imaging area.

In CT, artifacts may arise from patient motion, which produces streak artifacts throughout the image, known as *motion artifacts*; Figure 3.15(a) shows an example. Another typical artifact in CT is known as a *star artifact*. This artifact is generated by the presence of metallic materials in the patient, which results in incomplete projections. An example of this artifact is depicted in Figure 3.15(b). Another artifact is the so-called *beam hardening artifact*. This artifact shows up as broad dark bands or streaks in the image, and it is due to significant beam attenuation caused by certain materials. An example of this artifact is depicted in Figure 3.15(d). Finally, a common artifact in CT is the so-called *ring artifact*, illustrated in Figure 3.15(c). This artifact is caused by detectors that go out of calibration and do not properly record incoming data.

There are many reasons for medical images to be corrupted by artifacts. Evaluation and possibly removal of artifacts should be part of any high quality medical imaging system. Good design, proper calibration, and maintenance of medical imaging systems may control and even eliminate artifacts. The specific



(a) (b)

(c) (d)

Figure 3.15
Examples of artifacts in CT: (a) motion artifact, (b) star artifact, (c) ring artifact, and (d) beam hardening and partial voluming artifact. Courtesy of Vince Blasko and Beatrice Mudge, Department of Radiology, Johns Hopkins Hospital.

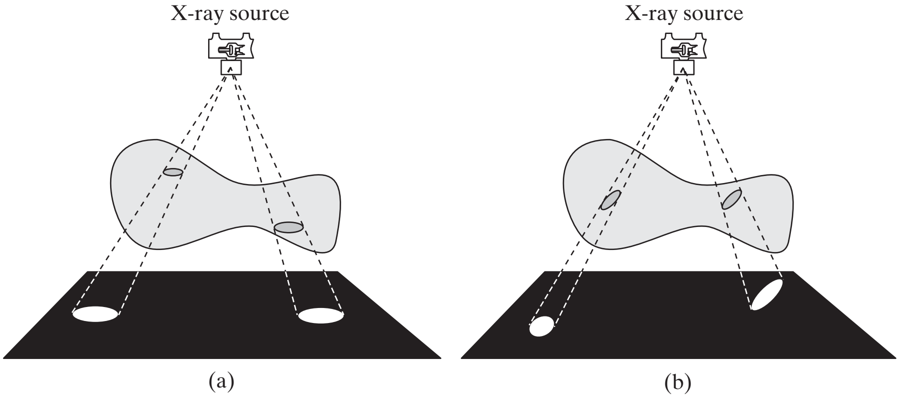


Figure 3.16
(a) Size distortion in a radiographic imaging system due to magnification: although the sizes of the two dark objects are different, their projections are the same. (b) Shape distortion in a radiographic imaging system due to x-ray beam divergence: although the two dark objects are the same, the shapes of their projections are different.

artifacts that appear in each particular imaging modality will be discussed in subsequent chapters.

3.7.2 Distortion

Medical imaging systems often introduce *distortion*, another factor affecting image quality. Distortion is geometrical in nature and refers to the inability of a medical imaging system to give an accurate impression of the shape, size, and/or position of objects of interest.

In projection radiography, for example, *size distortion* can be present due to *magnification* caused by the distance of the x-ray source from the object being imaged. This is illustrated in Figure 3.16(a). Notice that although the sizes of the two dark objects are different their projections are the same.

On the other hand, *shape distortion* can be generated as a result of unequal magnification of the object being imaged. One common cause of shape distortion is the fact that anatomical structures lie at different levels within the human body. In projection radiography, shape distortion is also caused by the divergence of the x-ray beam. This is illustrated in Figure 3.16(b). Notice that although the two dark objects are the same the shapes of their projections are different.

Unfortunately, distortion can be very difficult to determine and correct. In order to evaluate distortion, knowledge of the actual shape and size of the object being imaged is required. Moreover, a good understanding of the imaging geometry is necessary. Developing methods for correcting distortion is very important for increasing image quality and improving diagnosis. Distortions introduced by each particular imaging modality will be discussed in later chapters.

3.8 Accuracy

The preceding discussions in this chapter have focused on quantitative metrics of image quality. Even so, we said at the outset that image quality ultimately must be judged in the context of a specific clinical application. Medical images are used for diagnosis ("Is the disease present?"), prognosis ("How will the disease progress, and what is the expected outcome?"), treatment planning ("Which treatment will work best?"), and treatment monitoring ("Is the treatment reversing the disease, and to what extent?"). Fundamentally, we are interested in the *accuracy* of medical images in the context of these clinical applications. Here, "accuracy" means both conformity to truth (i.e., freedom from error) and clinical utility. In practice, we are usually interested in *quantitative accuracy* and *diagnostic accuracy*.

3.8.1 Quantitative Accuracy

Sometimes, we are interested in the numerical value of a given anatomic or functional feature within an image. For example, we may wish to measure tumor dimensions from a radiograph, or estimate glucose metabolic rate from a nuclear medicine image. In such situations, we need to know the error in our measurement. This error, or difference from the true value, arises from two sources: *bias*, which represents a systematic, reproducible difference from the truth, and *imprecision*, which represents a random, measurement-to-measurement variation.

It is helpful to separate the component of error due to bias from that due to imprecision. If we can establish that the measurement is precise (i.e., reproducible), then we can correct for systematic errors through the use of a calibration standard that converts the measured value to the true value. In practice, error usually arises from both components, and our measurements are never error-free.

3.8.2 Diagnostic Accuracy

We can conceptualize a medical image as representing some parameter (or set of parameters) of interest in a patient. For the sake of simplicity, we assume that the diagnostic process involves a test that extracts this parameter (or set of parameters), which is then used to classify a given patient as either normal or diseased. We further conceptualize two Gaussian distributions of the parameter (or set of parameters), one for normal patients and another for diseased patients, with (usually) some degree of overlap. This implies that some normal patients may be classified as diseased and vice versa.

In a clinical setting, we are interested in two parameters:

- *Sensitivity*, also known as the *true-positive fraction*; this is the fraction of patients with disease who the test (e.g., the medical image) calls abnormal.
- *Specificity*, also known as the *true-negative fraction*; this is the fraction of patients without disease who the test (e.g., the medical image) calls normal.

In practice, sensitivity and specificity are established in a group of patients through the use of a 2 × 2 contingency table, as shown in Figure 3.17. Here, *a* and *b* are, respectively, the number of diseased and normal patients who the

		Disease	
		+	−
Test	+	a	b
	−	c	d

Figure 3.17
A contingency table.

test calls abnormal, whereas c and d are, respectively, the number of diseased and normal patients who the test calls normal. In this case, the sensitivity and specificity are given by

$$\text{sensitivity} = \frac{a}{a+c} \qquad \text{and} \qquad \text{specificity} = \frac{d}{b+d}\,. \tag{3.82}$$

Typically, the ultimate diagnosis that confirms the presence or absence of disease is through a more-invasive test than medical imaging, a “gold standard” that is used to verify the accuracy of (noninvasive) medical imaging procedures. The *diagnostic accuracy* (DA) is the fraction of patients that are diagnosed correctly, and is given by

$$\text{DA} = \frac{a+d}{a+b+c+d}\,. \tag{3.83}$$

In order to maximize diagnostic accuracy, both sensitivity and specificity must be maximized. In practice, because of overlap in the distribution of parameter values between normal and diseased patients, a threshold must be established to call a study “abnormal,” as shown in Figure 3.18. A lower threshold implies that more studies will be called abnormal, thus increasing sensitivity but decreasing specificity. A higher threshold implies that fewer studies will be called abnormal, thus increasing specificity but decreasing sensitivity. One way of graphically depicting this relation is via a *receiver operating characteristic* (ROC) curve, which plots sensitivity as a function of 1 − specificity (see Problem 3.29).

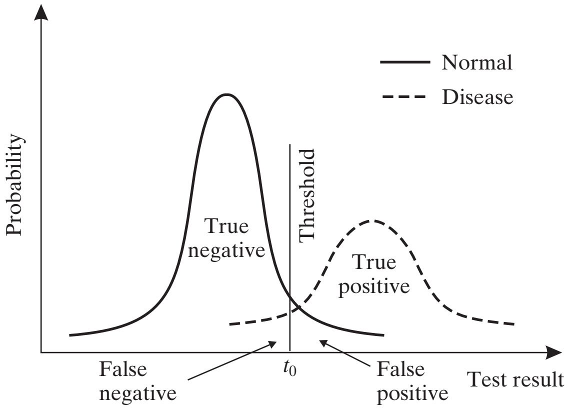


Figure 3.18
Probability distributions of the test results for normal subjects and patients with certain disease.

In practice, the threshold must be chosen as a balance between sensitivity and specificity. The choice of threshold for a specific test depends on the relative cost-of-errors in calling a normal patient abnormal or a diseased patient normal. The threshold also depends on the prevalence or proportion of all patients who have the disease, because in practice we are interested in two other parameters:

- *Positive predictive value* (PPV), which is the fraction of patients called abnormal who actually have the disease
- *Negative predictive value* (NPV), which is the fraction of persons called normal who do not have the disease

Notice that

$$\mathrm{PPV} = \frac{a}{a+b} \qquad \text{and} \qquad \mathrm{NPV} = \frac{d}{c+d}, \tag{3.84}$$

and that both depend on *prevalence* (PR), which is given by

$$\mathrm{PR} = \frac{a+c}{a+b+c+d}. \tag{3.85}$$

EXAMPLE 3.14

Diagnostic accuracy alone cannot tell how good a diagnostic method is. This example shows that a bad diagnostic method can still achieve a high DA.

Question Consider a group of 100 patients, among which 10 are diseased and 90 are normal. We simply label all patients as normal. Construct the contingency table for this test and determine the sensitivity, specificity, and diagnostic accuracy of the test.

Answer The contingency table for the test is shown below.

		Disease	
		+	−
Test	+	0	0
	−	10	90

From the table, we have $a = b = 0$, $c = 10$, and $d = 90$. So the sensitivity, specificity, and diagnostic accuracy of the test are

$$\text{Sensitivity} = \frac{a}{a+c} = 0,$$

$$\text{Specificity} = \frac{d}{b+d} = 1.0,$$

$$\mathrm{DA} = \frac{a+d}{a+b+c+d} = 0.9.$$

It can be seen that although all patients with the disease are diagnosed wrong, we still have a relatively high diagnostic accuracy of 0.9. This is simply because the patients without the disease comprise the majority of patients in the group being studied. This would be considered a flawed experimental design. ■

3.9 Summary and Key Concepts

Image quality characterizes the performance of a medical imaging system and directly affects clinical utility. It is assessed through a combination of specific performance parameters as measured in images. In this chapter, we presented the following key concepts that you should now understand:

1. *Image quality* refers to the degree to which an image allows a radiologist to accomplish the clinical goals of the imaging study.
2. The six most important factors influencing image quality are *contrast*, *resolution*, *noise*, *artifacts*, *distortion*, and *accuracy*.
3. *Contrast* refers to the difference in image intensity of an object or target and surrounding objects or background.
4. *Resolution* is the ability of an imaging system to distinguish and depict two signals that differ in space, time, or energy as distinct.
5. *Noise* is any random fluctuation in an image; noise generally interferes with the ability to detect a signal in an image.
6. Continuous signals are transformed by *sampling* or *discretization* into discrete signals in order to be digitally represented.
7. *Aliasing* is caused by improper sampling of a continuous signal, yielding artifacts in the resulting digital image.
8. *Artifacts* are false signals in an image that do not represent any valid structural or functional signal in the patient.
9. *Distortion* is any geometric inaccuracy in size or shape.
10. *Quantitative accuracy* refers to the accuracy, compared with the truth, of numerical values obtained from an image; *diagnostic accuracy* refers to the accuracy of interpretations and conclusions about the presence or absence of disease drawn from image patterns.

Further Reading

Barrett, H.H. and Swindell, W. *Radiological Imaging: The Theory of Image Formation, Detection, and Processing*. New York, NY: Academic Press, 1981.

Carlton, R.R. and Adler, A.M. *Principles of Radiographic Imaging: An Art and a Science*, 5th ed. Clifton Park, NY: Delmar Cengage Learning, 2012.

Cherry, S.R., Sorenson, J.A., and Phelps, M.E. *Physics in Nuclear Medicine*, 4th ed. Philadelphia, PA: W. B. Saunders, 2012.

Cunningham, I.A. "Applied Linear Systems Theory." pp. 79–159 in *Handbook of Medical Imaging. Vol. 1, Physics and Psychophysics*. Bellingham, WA: SPIE Press, 2000.

Dobbins, J.T. "Image Quality Metrics for Digital Systems." pp. 161–222 in *Handbook of Medical Imaging. Vol. 1, Physics and Psychophysics*. Bellingham, WA: SPIE Press, 2000.

Evans, A.L. *The Evaluation of Medical Images*. Bristol, England: Adam Hilger, 1981.

Krestel, E. *Imaging Systems for Medical Diagnostics*. Berlin, Germany: Siemens Aktiengesellschaft, 1990.

Papoulis, A. and Pillai, S.U. *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, NY: McGraw-Hill, 2002.

Smith, N.B. and Webb, A. *Introduction to Medical Imaging: Physics, Engineering, and Clinical Applications*. Cambridge, UK: Cambridge University Press, 2011.

Wolbarst, A.B. *Physics of Radiology*, 2nd ed. Norwalk, CT: Appleton and Lange, 2005.

Problems

Contrast

3.1 Prove Equation (3.5).

3.2 Consider an LSI medical imaging system with PSF given by

$$h(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

(**a**) Calculate the MTF associated with this system.

(**b**) Plot the MTF as a function of frequency.

(**c**) If a sinusoidal object $f(x,y) = 2 + \sin(\pi x)$ is imaged through the system, what is the percentage change in modulation caused by this system?

3.3 Let H_1 and H_2 be two 1-D LTI systems whose PSFs are $h_1(x)$ and $h_2(x)$:

$$h_1(x) = e^{-x^2/5}, \quad h_1(x) = e^{-x^2/10}.$$

(**a**) Find the MTF of system H_1.

(**b**) Find the MTF of the cascade system of H_1 and H_2.

3.4 Show that the MTF of a nonisotropic medical imaging system is given by Equation (3.9).

3.5 Consider an organ with a tumor is imaged. In the resulting image, the organ has intensity I_o and the tumor has intensity $I_t > I_o$. Which of the following methods can improve the local contrast if the organ is treated as background:

(**a**) Multiplying the image by a constant α.

(**b**) Subtracting a constant $0 < I_s < I_o$ from the image.

Resolution

3.6 A new imaging system with which you are experimenting has anisotropic properties. You measure the impulse response function as $h(x,y) = e^{-\pi(x^2+y^2/4)}$. What is the FWHM of the system as a function of polar angle θ?

3.7 A medical imaging system has the following line spread function, where $\alpha = 2$ radians/cm:

$$l(x) = \begin{cases} \cos(\alpha x) & |\alpha x| \le \pi/2 \\ 0 & \text{otherwise} \end{cases}.$$

(a) Find the FWHM.

(b) Determine the resolution of this imaging system in lines per cm.

3.8 Consider a one-dimensional linear imaging system whose PSF is given by

$$h(x;\xi) = e^{\frac{-(x-\xi)^2}{2}},$$

which represents the response to the shifted impulse $\delta(x-\xi)$.

(a) What is the PSF of a system having the same basic form of PSF as that above but whose FWHM is one half as large?

(b) Does the resolution of the system improve when you make the above change in FWHM? Why?

(c) What characteristics should be expected from the MTF in an imaging system that has high contrast (low blurring)?

3.9 Consider again the LSI medical imaging system of Problem 3.2.

(a) Calculate its *line spread function* (LSF).

(b) What is the FWHM associated with this system?

3.10 Consider a one-dimensional medical imaging system which is composed of two subsystems with PSFs given by

$$h_1(x) = e^{-x^2/2} \quad \text{and} \quad h_2(x) = e^{-x^2/200}.$$

(a) What is the FWHM associated with each subsystem?

(b) What is the FWHM associated with the overall system?

(c) Which subsystem most greatly affects the FWHM of the overall system?

3.11 A bar phantom is imaged by an LSI system, which is modeled as an ideal moving average system. The PSF of the system is

$$h(x,y) = \mathrm{rect}\left(\frac{x}{\Delta}, \frac{y}{\Delta}\right).$$

(a) If the bar separation of the bar phantom is Δ, what is the output of the image system?

(b) If the bar separation is 0.5Δ, what is the output of the imaging system?

(c) Derive the relation of the contrast of the output image with the bar separation and draw conclusion on the resolution of the imaging system.

Random Variables and Noise

3.12 Show that if N is a random variable with mean μ_N and standard deviation σ_N then

$$M = \frac{N - \mu_N}{\sigma_N}$$

is a random variable with mean $\mu_M = 0$ and standard deviation $\sigma_M = 1$.

3.13 Suppose N random variables X_i, $i = 1, \cdots, N$, are independent with mean μ_i and variance $\sigma_i^2, i = 1, \cdots, N$. Show that the mean and the variance of the random variable $X = \sum_{i=1}^{N} X_i$ are given by

$$\mu = \sum_{i=1}^{N} \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{N} \sigma_i^2 .$$

3.14 In the above problem, if X_i, $i = 1, \cdots, N$, are not independent, will the equalities for the mean and variance of the sum of the random variables still hold? Explain.

3.15 Show that the expected value and variance of a uniform random variable X over the interval (a, b) are given by

$$\mu_X = \frac{a+b}{2} \quad \text{and} \quad \sigma_X^2 = \frac{(b-a)^2}{12} .$$

3.16 Consider two medical imaging systems with PSFs $h_1(x, y)$ and $h_2(x, y)$ and MTFs $\mathrm{MTF}_1(u, v)$ and $\mathrm{MTF}_2(u, v)$, respectively, such that

$$\mathrm{MTF}_1(u, v) \leq \mathrm{MTF}_2(u, v) .$$

Show that the system with the larger MTF, and thus with better contrast and resolution properties, is characterized by a larger output power SNR, given by (3.63), and is thus better in terms of image quality.

3.17 Consider the system shown in Figure P3.1, in which an image $f(x, y)$ is corrupted with zero mean white noise $n(x, y)$ with variance σ_n^2. The corrupted image is input to a system with PSF $h(x, y)$ to get an output image $g(x, y)$.

(a) What are the mean and variance of the noise in the output $g(x, y)$?

(b) What are the power SNR for the input and output images of the system?

(c) If the system does not change $f(x, y)$ in any way, under what condition(s) will the system improve the SNR?

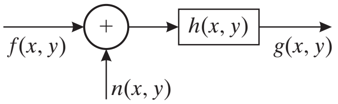


Figure P3.1
A system with additive noise.

Sampling Theory

3.18 A signal $f(t)$ is defined as

$$f(t) = \begin{cases} \sin(2\pi t/T), & 0 \leq t \leq T \\ 0, & \text{otherwise} \end{cases} .$$

We sample the signal with a sampling period of $\Delta T = 0.25T$.

(a) What is $f_s(t)$, and what is $f_d(m)$?

(b) Define a new signal $f_b(t)$ as

$$f_b(t) = f_d(k), \quad \text{for } k\Delta T \le t < (k+1)\Delta T.$$

Sketch $f_b(t)$ and find its Fourier transform.

(c) Repeat the above using a sampling period of $\Delta T = 0.5T$.

3.19 The Nyquist sampling periods for 1-D band-limited signals $f(x)$ and $g(x)$ are Δ_f and Δ_g, respectively. Find the Nyquist sampling periods for the following signals:

(a) $f(x - x_0)$, where x_0 is a given constant.

(b) $f(x) + g(x)$.

(c) $f(x) * f(x)$.

(d) $f(x)g(x)$.

(e) $|f(x)|$.

3.20 We want to sample the 2-D continuous signal $f(x, y) = \exp\{-\pi(x^2 + y^2)\}$ by means of a rectangular sampling scheme to obtain 1.5 samples per millimeter. Determine the PSF $h(x, y)$ of an ideal low-pass anti-aliasing filter with the maximum possible frequency content. What percentage of the spectrum energy (see Problem 2.3) of $f(x, y)$ is preserved by this filter? In practice, can we sample $f(x, y)$ alias-free without using an anti-aliasing filter?

3.21 In a medical imaging system, small detectors having width w can be used individually or be grouped together into larger (2x2) detectors. Two example groupings are shown in Figure P3.2. Assume that the small detectors have uniform response over their spatial extent and that the detector array is stationary.

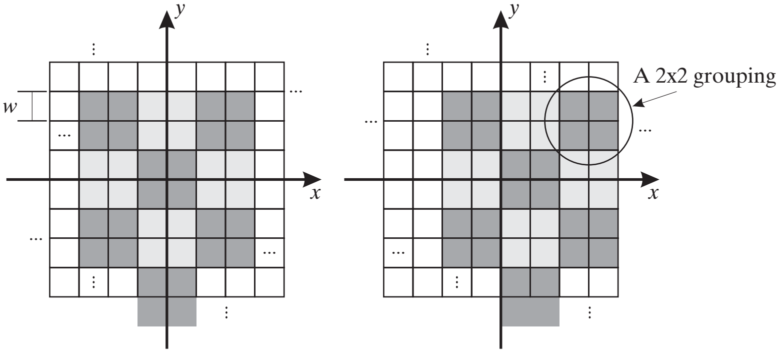


Figure P3.2
Grouping detectors.

(a) Find the impulse response, the MTF, and the (horizontal) FWHM of the system when the small detectors are used.

(b) Consider the small detectors, and assume that a sinc function is zero except for its main lobe. What property of the imaged object guarantees that no aliasing will occur?

(c) Describe a sequential imaging process, involving groupings like those shown in the picture, which would guarantee that no aliasing occurs for any object.

(d) What is the resolution (FWHM) of the imaging system in part (c)?

(e) Can the sequential grouping you described in part (c) be done after an image is acquired using the small detectors? Explain.

3.22 Consider the following 1-D imaging systems, system 1 and system 2, with PSFs given by

$$h_1(x) = \text{rect}(2x),$$
$$h_2(x) = e^{-\pi x^2}.$$

(a) Calculate the full width half maximum of each system. Which has better spatial resolution?

(b) Calculate the MTF of each system.

You wish to image the following signal

$$f(x) = \cos(4\pi x)$$

with each system. Let $g_1(x)$ be the output of system 1 when $f(x)$ is the input, and let $g_2(x)$ be the output of system 2 when $f(x)$ is the input.

(c) Calculate $g_1(x)$ and $g_2(x)$. Which system should you use to image this signal? Hint: you should not have to calculate any convolution integrals.

(d) What is the largest sampling interval you can use to sample $g_1(x)$ and $g_2(x)$ such that you do not introduce distortion through aliasing?

Artifacts, Distortion, and Accuracy

3.23 Compare artifacts with noise, and state the common properties and differences between them.

3.24 In projection radiography, size distortion is an artifact that cannot be ignored. Suppose the x-ray source is a perfect point source located at origin and the detector plane is the $x = d$ plane. A ball centered on the x-axis between the source and the detector plane is imaged by the system. Assuming that the radius of the ball is $r < d/2$, derive the relation between the radius R of the image of the ball on the detector plane and the location of the ball. If the source-detector distance is fixed, what measure(s) can be taken to reduce the size distortion? What is the smallest ratio R/r that can be achieved?

3.25 A medical imaging system has a geometric distortion, which is well modeled as

$$x = \xi + \frac{1}{50}\xi\eta^2,$$
$$y = \eta,$$

where x and y are the coordinates in the image plane and ξ and η are coordinates in the physical domain.

(a) If we take measurements on rectangular grids in the image plane, find a way to correct the geometric distortion.

(b) If we want to take measurements on rectangular grids in the physical domain, on what points should we take measurements in the image domain?

3.26 Suppose the pdfs of the test result for patients with and without a disease are modeled as Gaussian with different mean values and variances, as shown in Figure 3.18. We design a diagnostic test by selecting a threshold t_0. For a patient whose test value is below t_0, we call it normal. If the test value is above t_0, we call is diseased. By selecting different thresholds, we obtain different *diagnostic tests*. Denote the means and variances of the test result of the normal and diseased subjects as μ_0, σ_0^2, and μ_1, σ_1^2. And assume $\mu_0 < \mu_1$.

(a) Find expressions for the pdfs of the test value for normal and diseased subjects.

(b) If we choose $t_0 = (\mu_0 + \mu_1)/2$, compute the sensitivity and the specificity.

(c) Derive an expression for the sensitivity as a function of the threshold value.

(d) Derive an expression for the diagnostic accuracy as a function of the threshold value.

Applications, Extensions and Advanced Topics

3.27 Show that a medical imaging system composed of several subsystems will be inferior to each individual subsystem in terms of contrast and resolution.

3.28 In most modern medical imaging systems, images are measured at discrete grid points. Interpolation is needed when we want to treat the images as functions defined on a continuous domain. For 1-D signals, linear interpolation is one of the simplest. It can easily be extended to bilinear interpolation in 2-D and trilinear interpolation in 3-D. Linear interpolation is defined as

$$f(x) = \frac{x_1 - x}{x_1 - x_0} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1), \quad x_0 \le x \le x_1 .$$

In 2-D, bilinear interpolation for $f(p)$ from $f(A)$, $f(B)$, $f(C)$, and $f(D)$ is done in the following steps (see Figure P3.3):

- Using linear interpolation to get $f(E)$ from $f(A)$ and $f(B)$.
- Using linear interpolation to get $f(F)$ from $f(C)$ and $f(D)$.
- Using linear interpolation to get $f(P)$ from $f(E)$ and $f(F)$.

(a) Derive an explicit expression for $f(P)$ using $f(A)$, $f(B)$, $f(C)$, and $f(D)$.

(b) Prove that the results are the same whether we get $f(P)$ from $f(E)$ and $f(F)$, or from $f(G)$ and $f(H)$.

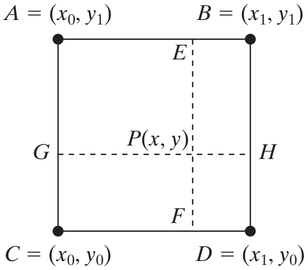


Figure P3.3
Bilinear interpolation.

(c) In Problem 3.25, if we take measurements on rectangular grids in the image plane at $(m\Delta x, n\Delta y)$, with $\Delta x = \Delta y = 1$ and m, n are integers, how should we interpolate the value for $\xi = 3$ and $\eta = 3.5$ in the physical domain?

3.29 As we mentioned in Example 3.14, diagnostic accuracy alone is not enough to determine whether a test is good. In this problem, we will introduce the concept of *receiver operating characteristic* (ROC) curves with a simple example. Consider the situation described in Problem 3.26; suppose the mean and variance of the test value for the normal subjects are $\mu_0 = 2$ and $\sigma_0^2 = 1$, with arbitrary units. The mean and variance of the test value for the diseased subjects are $\mu_1 = 8$ and $\sigma_1^2 = 4$, with arbitrary units.

(a) For threshold value t_0 in the interval $\mu_0 \leq t_0 \leq \mu_1$, plot the curve of sensitivity (vertical axis) against 1 − specificity (horizontal axis). This is the ROC curve.

(b) What is the ROC curve for a perfect diagnostic test?

(c) For a given ROC curve, if we define the optimal test to be the one that is closest to the perfect test, determine the optimal test (threshold) for the above case.

This page intentionally left blank

PART II

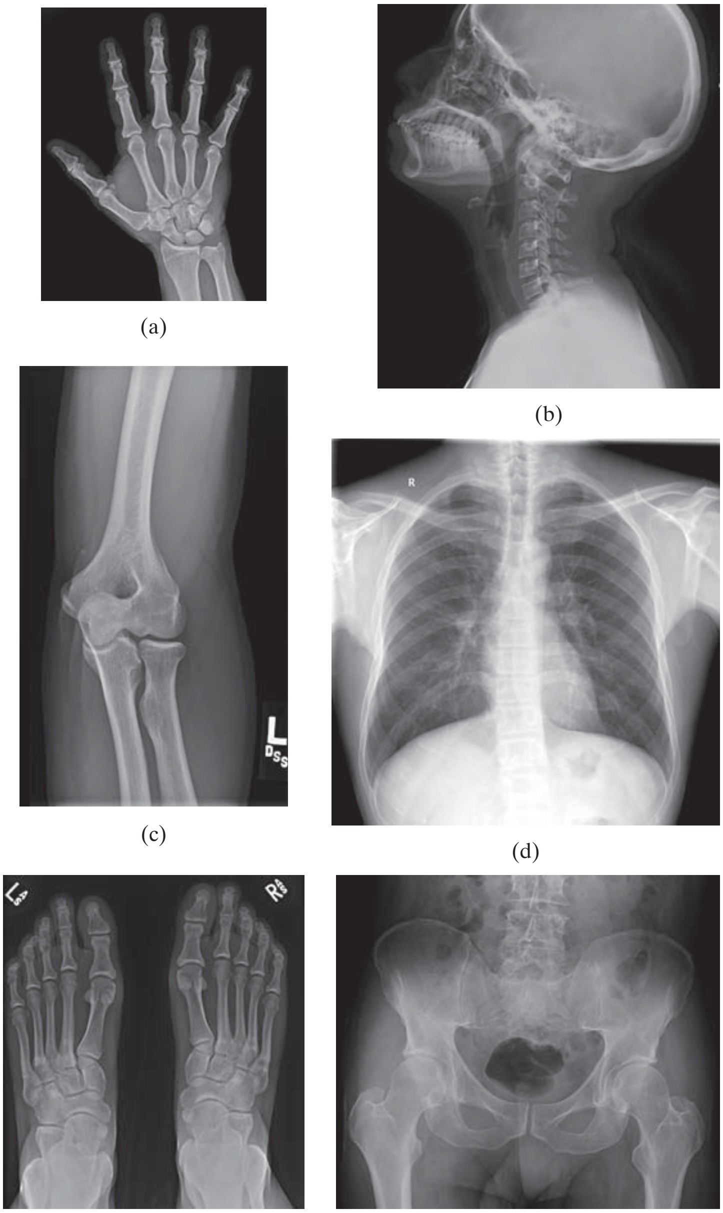
Radiographic Imaging

Overview

Ionizing radiation—radiation capable of ejecting electrons from atoms—forms the basis of a number of important imaging modalities. In some cases, we are interested in the transmission of ionizing radiation through the body. In other cases, we are interested in the emission of ionizing radiation from the body. In this part of the book, we consider the two main imaging modalities that make use of the transmission of ionizing radiation through the body: projection radiography and computed tomography.

Projection radiography and computed tomography rely on the *transmission* of ionizing radiation through the body. Various tissues and organs within the body *attenuate* or decrease the intensity of the beam of ionizing radiation as it passes through the body. Thus, even if the beam entering the body is of uniform intensity, the beam exiting the body contains "shadows" of tissues and organs as a "latent image" of varying beam intensity. Since it is the physical characteristics of the tissues or organs (e.g., effective atomic number and density) that determine that tissue's attenuation abilities, the resulting images depict *structures* within the body, and projection radiography and computed tomography are considered *anatomical* imaging modalities, since they portray anatomy.

In Figure 1.1(b), we showed the most classic projection radiograph—a "chest x-ray." In Figure II.1, we show several other transmission images. Figure II.1(a) is also a projection radiograph—this time, of a hand. Notice the dominance of bone in this image and the homogeneity of soft tissues. This is because soft tissues do not attenuate x-rays as well as bone and the difference in attenuation between different types of soft tissues is very small. Figure II.1(b) shows a radiograph of the head and neck. Because the shoulders are wide in this lateral (left-to-right) view and they contain bone, the attenuation there is quite large, and the image is washed out (white) there. The brain, being all soft tissue, is virtually devoid of features in this image. This lack of detail in the brain was of great frustration to radiologists until computed tomography (CT) and magnetic resonance imaging (MRI) burst onto the scene in the 1970s and 1980s.



(a) (b) (c) (d) (e) (f)

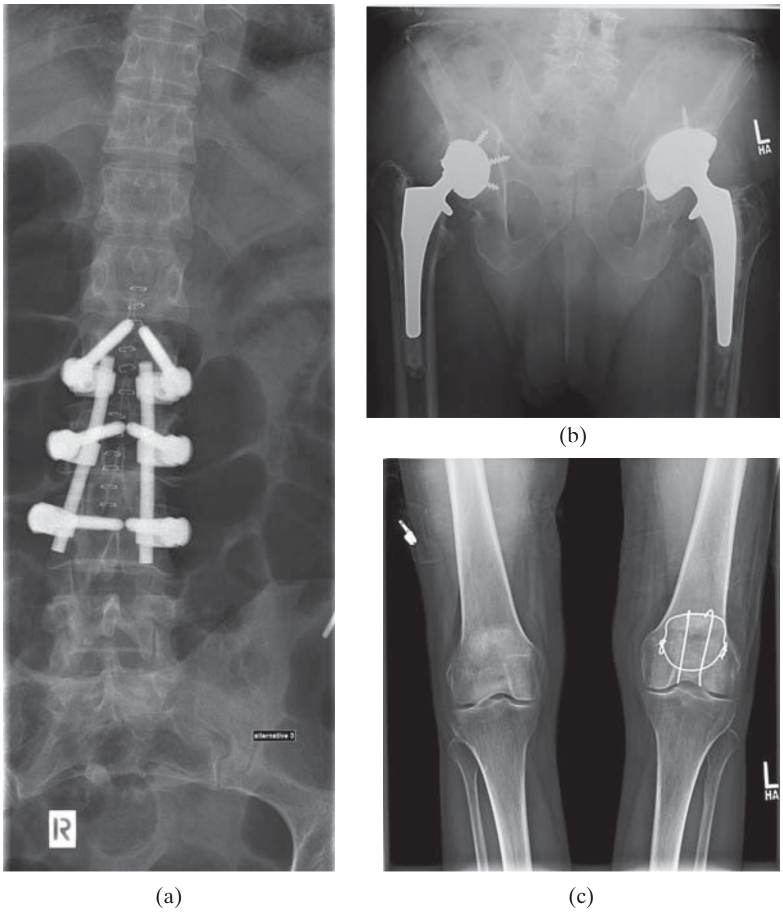
Figure II.1
Representative x-ray transmission images of various parts of the body: (a) hand, (b) head and neck, (c) knee, (d) chest, (e) feet, and (f) pelvis. All images courtesy of GE Healthcare.

It is instructive to look at the bones around the knee in Figure II.1(c). Why do they appear to have bright outlines? We know that bones are filled with marrow, which is a soft tissue. Therefore, around the edges, the x-rays have to travel farther in the bone itself and therefore they experience more attenuation, yielding a brighter image. If the bones were solid, then they would be brighter in the middle on a radiograph.

Figure II.1(d) shows a chest x-ray, a staple of modern radiography. In one single, quick acquisition the physician can examine the lungs, the heart, and the diaphragm as well as all the bones in the torso including the spine. Notice that the entire rib cage appears in this image—displaying the ribs on both the front and back of the body—revealing very evidently the overlaying structure that is characteristic of x-ray radiographs. (The posterior ribs are the ones that are nearly horizontal in the image.)

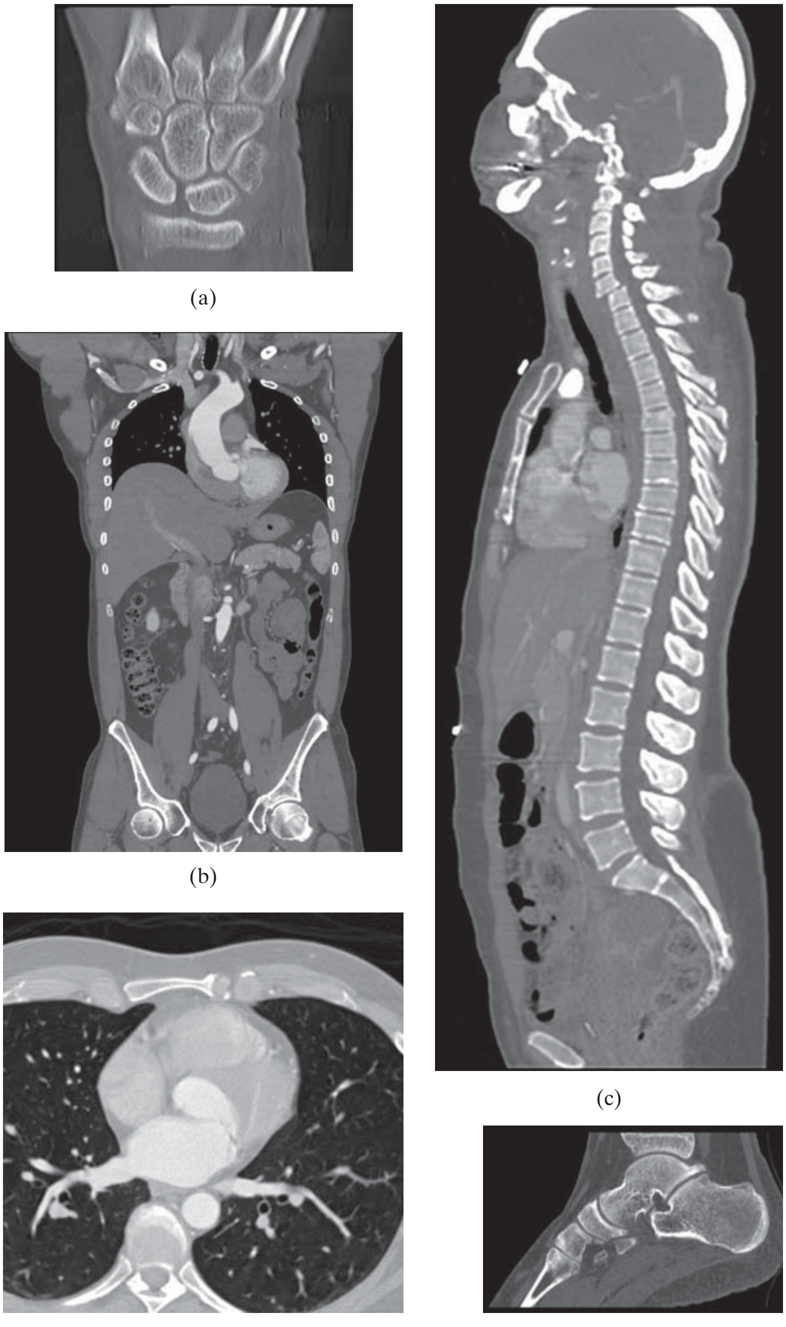
Figure II.1(e) shows a radiograph of human feet. In this image, we can recognize a faint outline of the soft tissue—looking like a conventional footprint. We see the white outlines of bone and also the washed out appearance of many overlaying bones in the ankle region including the fibula and tibia. This image also shows the symbols “L” and “R” indicating which foot is the left and which is the right, respectively. In this case, left is on our left and right is on our right, as if we were looking down at our own feet. The markers are made from lead and are placed by the radiographer near the appropriate anatomy. Without these markers, it is sometimes difficult or impossible to know left from right—and this can lead to medical errors.

Figure II.1(f) shows a radiograph of the human pelvis. This image shows bony structures as usual and faint evidence of soft tissues. Bowel gas is fairly



(a) (b) (c)

Figure II.2
X-ray images of (a) the spine shown a surgical fixation device, (b) the pelvis showing two artificial hip joints, and (c) the knee showing bone fixation wires. Image (a) courtesy of Philips Healthcare. Images (b) and (c) courtesy of GE Healthcare.



(a)

(b)

(c)

(d)

(e)

Figure II.3
Various CT images of the human body: (a) wrist, (b) torso and abdomen, (c) complete spine, (d) chest, and (e) ankle. Image (a) courtesy of Dr. Avneesh Chhabra, Department of Radiology, Johns Hopkins Hospital. Images (b) and (e) courtesy of GE Healthcare. Image (c) courtesy of Philips Healthcare. Image (d) courtesy of Osirix.

prominent in this image as well as darker regions that are caused by the negligible x-ray attenuation of air.

Projection radiography and computed tomography both rely on an *x-ray tube* to produce the beam of ionizing radiation, and radiation detectors on the other side of the patient to detect the radiation that is passed through without being fully absorbed. In the case of projection radiography, the detector is sometimes a piece of film, which then also serves as the display device. Increasingly often, the detector is a digital device that is capable of storing a latent image of the x-rays that hit the detector and either instantly or after a delay transmitting the latent image to a computer. In the case of computed tomography, an x-ray tube with a collection of detectors on the opposite side of the patient rotate together around the patient, producing x-ray images at a number of angles. These images, stored in a computer, form the *projection data* used to *reconstruct* cross-sectional or transaxial images through the patient.

Projection radiography is perhaps the most commonly performed medical imaging approach. It is used any time a physician needs an image of structures within the body, especially to obtain an overall picture of the torso and to examine bones. It is also commonly used during and after orthopedic surgery to verify the placement of artificial joints and various fixation devices, as shown in Figure II.2.

CT is distinguished from projection radiography in that it removes overlaying structures by reconstructing cross sections of the body rather than imaging projections. CT also produces images with lower resolution than conventional radiographs, and typically requires higher x-ray dose to the patient. Figure II.3 shows various CT images of the human body.

Modern CT scanners invariably acquire stacks of two-dimensional cross sections which can be handled by a computer as a three-dimensional (3-D) dataset. Because of this capability, CT images can be *reformatted* by the computer into arbitrary cross sections that are different than those that are natively acquired by the scanner. All of the images in Figure II.3 are reformatted with the exception of II.3(d). For example, in the large sagittal image shown in Figure II.3(c) the data were acquired axially while sliding the patient on a table through the axially-oriented CT scanner.

Figure II.3(d) shows a cross section of the chest (including the heart in which all four chambers can be identified). The fact that the blood appears bright—as bright as bone—is because a contrast agent has been introduced. Also, it might be noticed that many of the bones appearing in these images have a bright ring around them, which appears somewhat similar to that of radiographs. Because the structures are not superposed in CT, this ring appearance actually reflects differences in the make-up of the tissues. In this case, cortical bone has a higher attenuation than cancellous bone and bone marrow, and is therefore brighter on these images.

CHAPTER

4

Physics of Radiography

4.1 Introduction

This chapter provides an introduction to the physics behind x-ray imaging modalities. Broadly speaking, these modalities can be divided into two types, projection radiography and computed tomography, which we cover in the next two chapters. These modalities do not involve radioactivity (a common misconception), so we delay the introduction of radioactivity until Chapter 7, where it is needed for the presentation of the nuclear medicine modalities in Chapters 8 and 9.

As we pointed out in Chapter 1, x-rays were discovered in 1895 by Roentgen while working with a Crooke's tube (a precursor to the modern x-ray tube). He noticed that "rays of mysterious origin," which he named *x-rays*, caused a fluorescent plate to glow. Upon investigation, he determined that these rays were in fact produced by the tube. The first radiograph of a human being—the hand of Roentgen's wife—was made by Roentgen within a month of the discovery. Roentgen's discovery of x-rays and their immediate application to imaging the human body truly mark the birth of medical imaging.

Today, we know that x-rays are electromagnetic (EM) waves whose frequencies are much higher than those of visible light. They are only one form of *ionizing radiation*—that is, radiation capable of ejecting electrons from atoms—used in medical imaging, however. Other forms of ionizing radiation used in medical imaging include *particulate radiation* and *gamma rays*. Both particulate radiation and gamma rays can be products of radioactive decay, and we will have more to say about this in Chapter 7. It also turns out that an *electron beam*, which is a form of particulate radiation, is used in the production of x-rays. So, we need to study certain aspects of particulate radiation in this chapter as well. Furthermore, gamma rays, like x-rays, are high-frequency EM waves. So, in studying the propagation properties of x-rays, we are also implicitly studying the same for gamma rays.

This chapter provides a necessary background for all imaging modalities that use ionizing radiation, including projection radiography (Chapter 5), computed tomography (Chapter 6), planar scintigraphy (Chapter 8), and emission computed tomography (Chapter 9). We start by describing the basic physical properties of the atom and the concepts of excitation and ionization. We then describe ionizing radiation in a more general context, concluding with its measurement and interaction with biological tissues.

4.2 Ionization

Ionization is the ejection of an electron from an atom, creating a *free electron* and an *ion*. Radiation that carries enough energy to ionize an atom is called *ionizing radiation*. Here, the term *radiation* covers a broad range of physical phenomena. For example, visible light, x-rays, gamma rays, and electron beams are all examples of radiation. High-energy EM waves such as x-rays and gamma rays are ionizing radiations; light is not. Particulate radiations such as an electron beam may be ionizing depending on the energy each particle possesses. We now discuss the structure of atoms and the process of ionization in more detail.

4.2.1 Atomic Structure

Today's concept of atomic structure is based on a quantum mechanical picture of the atom (from the 1920s), which in turn arose from the Bohr model of the hydrogen atom (in 1913). An atom consists of a nucleus of protons and neutrons, which together are called *nucleons*, surrounded by orbiting electrons. A common visualization of atomic structure, sometimes called the *planetary atom*, is shown in Figure 4.1. The *atomic number* Z is equal to the number of protons in the nucleus and defines the element. Each proton has a positive charge equal in magnitude and opposite in charge to that of each electron. Since an atom as a whole is electrically neutral, there are an equal number of electrons and protons in an atom; hence, Z also represents the number of electrons in the atom. A short table of the elements is given in Table 4.1.

The *mass number* A of an atom is equal to the number of nucleons in the nucleus. The term *nuclide* refers to any unique combination of protons and neutrons forming a nucleus. Nuclides are typically denoted by either ${}^{A}_{Z}X$ or X-A, where X is the element symbol. For example, ${}^{12}_{6}\mathrm{C}$ and C-12 are both symbols for the most abundant carbon atom. Notice that the element symbol and atomic number are redundant.

As indicated in Table 4.1, the number of neutrons in a given nucleus is approximately equal to the number of protons; and certain combinations are stable, while others are unstable. Unstable nuclides are called *radionuclides*,

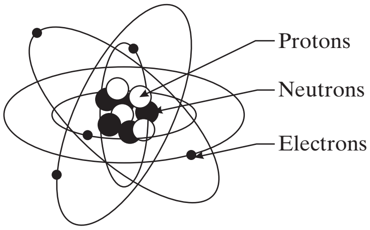


Figure 4.1
The planetary atom: a common (and useful) visualization of atomic structure.

TABLE 4.1

Abbreviated Table of the Elements

Element	Symbol	Atomic Number (Z)	Mass Numbers of Stable Isotopes (A)	Mass Numbers of Unstable Isotopes (A)
Hydrogen	H	1	1, 2	3
Helium	He	2	3, 4	5, 6, 8
Lithium	Li	3	6, 7	5, 8, 9, 11
Beryllium	Be	4	9	6, 7, 8, 10, 11, 12
Boron	B	5	10, 11	8, 9, 12, 13
Carbon	C	6	12, 13	9, 10, 11, 14, 15, 16
Nitrogen	N	7	14, 15	12, 13, 16, 17, 18
Oxygen	O	8	16, 17, 18	13, 14, 15, 19, 20

From Johns and Cunningham, 1983.

and their atoms are *radioactive*. Radionuclides are statistically likely to undergo radioactive decay, which causes a rearrangement of the nucleus, which in turn gives off energy and results in a more stable nucleus. For example, ${}^{14}_{6}\mathrm{C}$ or C-14 denotes a radioactive carbon atom; it is statistically likely to decay into ${}^{14}_{7}\mathrm{N}$, a stable nitrogen atom, during which a beta particle will be emitted from the nucleus. We will have a lot more to say about radionuclides and radioactivity in Chapters 7, 8, and 9.

The electrons orbiting the nucleus are organized into so-called orbits or *shells*. The K shell is closest to the nucleus, the L shell is next, then the M shell, and so on. Electrons are restricted to specific quantum states within each shell. Only one electron is permitted to be in each state, and this leads to a maximum number of electrons per shell, given by $2n^2$, where n is the shell number ($K = 1$, $L = 2$, etc.). For example, only two electrons are permitted in the K shell, 8 in the L shell, 18 in the M shell, and 32 in the N shell. Each atom has a so-called *ground state* configuration for its electrons, which corresponds to the lowest energy configuration of the atom—nature's preferred arrangement of electrons in a given atom. Generally speaking, in the ground state, the electrons will be in the lowest orbital shells and within the lowest energy quantum states within each shell. A diagram of the arrangement of electrons in shells for the carbon atom is shown in Figure 4.2.

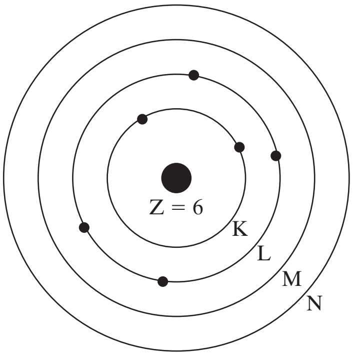


Figure 4.2
The arrangement of electrons into shells for the carbon atom.

4.2.2 Electron Binding Energy

It is energetically more favorable for an electron to be *bound* in an atom rather than to be *free*. In other words, the total energy of the atom is less than the total energy of the atom (minus the electron) and the (free) electron. The difference between these two energies is called the *electron binding energy*. Binding energy is usually specified in units of *electron volts* (eV). Recall that one electron volt is equal to the kinetic energy gained by an electron when accelerated across one volt potential. For comparison with more conventional units of energy, $1\ \text{eV} = 1.6 \times 10^{-12}$ ergs $= 1.6 \times 10^{-19}$ J.

A given electron's binding energy depends on the element to which the electron is bound and the shell within which the given electron resides. For any given element, electron binding energy decreases with increasing shell number. The binding energy of the sole electron in a hydrogen atom is 13.6 eV, the smallest binding energy among all "lighter" atoms (i.e., those having smaller atomic numbers). There are even smaller electron binding energies in the outer orbits of the "heavier elements." For example, some electrons in the O-shell of mercury atoms have binding energies equal to 7.8 eV. For our purposes, however, it is sufficient to consider an "average" binding energy of the electrons in a given atom or even in a given molecule. For example, the average binding energy for air is about 34 eV. Metals have significantly larger average electron binding energies. For example, the average binding energy of lead is about 1 keV and for tungsten it is about 4 keV.

4.2.3 Ionization and Excitation

If radiation (particulate or EM) transfers energy to an orbiting electron (in an atom of the material the radiation is passing through) which is equal to or greater than that electron's binding energy, then the electron is ejected from the atom. This process, illustrated in Figure 4.3, is called *ionization*. It yields an *ion* (an atom with a +1 charge in this case) and the electron that came from it, which together are called an *ion pair*. By convention, radiation with energy greater than or equal to 13.6 eV is considered *ionizing*; all other radiations are considered *nonionizing*.

It is possible for a single ionizing radiation emission to ionize many atoms along its path before its energy is exhausted. The ionizing radiation we use in

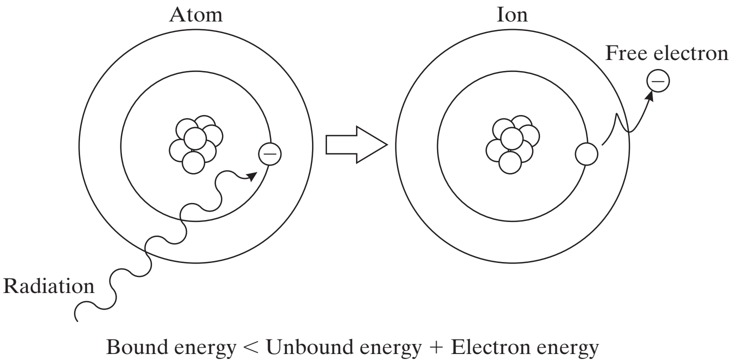


Figure 4.3
After an ionization event, the ejected electron plus the ion have more energy than the original atom.

medical imaging has energies ranging from about 25 keV to 500 keV. Broadly speaking, the "easiest" atom to ionize is a hydrogen atom, since it has the smallest average electron binding energy. Tungsten may be about the "most difficult" to ionize in the substances normally encountered in medical imaging. Using the range of ionizing radiations energies given above and the binding energies for these two elements, we can conclude that a single ionizing particle or ray used in medical imaging is capable of ionizing between 10 and 40,000 atoms before its energy is exhausted.

If an ionizing particle or ray transfers some energy to a bound electron but less than the electron's binding energy, then the electron is raised to a higher energy state—to a more outer orbit—but is not ejected. This process is called *excitation*. In both ionization and excitation, an electron shell is left with a "hole" that must be filled in order to return the atom to a lower energy state. The filling of these open holes comprises an important source of secondary radiation called *characteristic radiation*, which we will explore below.

EXAMPLE 4.1

Suppose an electron is accelerated within a vacuum from a heated cathode held at ground potential to an anode held at 120 kV (DC).

Question If the anode is made of tungsten, what is the maximum number of tungsten atoms that can be ionized on average?

Answer The electron will have kinetic energy equal to 120 keV upon reaching the anode (recall the definition of eV). The average binding energy of tungsten is 4 keV. Therefore, the maximum number of tungsten atoms that can be ionized on average is 120 keV/4 keV = 30. ■

4.3 Forms of Ionizing Radiation

Ionizing radiation can be divided into two broad categories: *particulate* and *electromagnetic*. We now describe these types of radiation in some detail.

4.3.1 Particulate Radiation

Any subatomic particle (i.e., a proton, neutron, or electron) can be considered to be ionizing radiation if it possesses enough kinetic energy to ionize an atom. In calculating the kinetic energy of these particles, relativistic effects cannot be ignored since the particles often travel at a significant fraction of the speed of light. As a particle approaches the speed of light, its relativistic mass increases, which significantly increases its kinetic energy above that which would be calculated using the usual low-speed approximation.

From Einstein's theory of relativity, we know that the (relativistic) mass of a particle is given by

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}}, \tag{4.1}$$

where m_0 is the rest mass of the particle, v is the speed of the particle, and c is the speed of light. Einstein's theory also tells us that there is an equivalence between the energy E of a particle and its mass, which is given by

$$E = mc^2 \,. \tag{4.2}$$

The kinetic energy of a particle is the difference in energy between the moving particle and the stationary particle:

$$\begin{aligned} \text{KE} &= E - E_0 \\ &= mc^2 - m_0c^2 \,. \end{aligned} \tag{4.3}$$

It is straightforward to show (see Problem 4.3) that when v is small relative to c, the kinetic energy reduces to

$$\text{KE} = \frac{1}{2}mv^2 \,, \qquad v \ll c \,, \tag{4.4}$$

which is the usual (nonrelativistic) equation for kinetic energy.

In nuclear medicine, we consider modes of radioactive decay that give rise to different kinds of particulate ionizing radiations (see Chapter 7). In projection radiography and computed tomography, however, the only particulate radiation we need to consider is due to electrons that are not bound in an atom and have gained kinetic energy. As we shall see, these electrons arise from a heated filament and a potential difference between a cathode and anode within an x-ray tube. In this case, we know the energies of these particles by the manner in which they were created. For example, electrons accelerated across an 80 kV potential will have energies equal to 80 keV. The speed and mass of these particles is interesting to know but largely irrelevant from an imaging point of view.

EXAMPLE 4.2

Consider an electron that has been accelerated between a cathode and anode held at a 120 kV potential difference. Assume the situation is *nonrelativistic*.

Question What is the speed of the electron when it "slams" into the anode?

Answer The rest mass of an electron is 9.11×10^{-31} kg. Because it was accelerated across a 120 kV potential, the electron must have kinetic energy given by KE = 120 keV. Therefore, assuming nonrelativistic speeds, we must have

$$\text{KE} = \frac{1}{2} \times 9.11 \times 10^{-31}\ \text{kg} \times v^2 = 120\ \text{keV}$$

Noting that 1 eV = $1.602\,177\,33 \times 10^{-19}$ J and recalling that 1 J = 1kg m^2/s^2, we find that

$$\begin{aligned} v &= \sqrt{\frac{2 \times 120 \times 10^3 \times 1.602\,177\,33 \times 10^{-19}\ \text{J}}{9.11 \times 10^{-31}\ \text{kg}}} \\ &= 2.054 \times 10^8\ \text{m/s} \,. \end{aligned}$$

This speed is a significant fraction of the speed of light, so our calculations are not accurate; they must be redone (see Problem 4.2). Looking ahead, this example tells us that the electrons in a typical x-ray tube are traveling at relativistic speeds by the time they strike the anode. ■

4.3.2 Electromagnetic Radiation

Electromagnetic radiation comprises an electric wave and a magnetic wave traveling together at right angles to each other. Radio waves, microwaves, infrared light, visible light, ultraviolet light, x-rays, and gamma rays are all examples of EM radiation. EM radiation has no rest mass and no charge, and can act like either a particle or a wave. When treated as a "particle," EM radiation is conceptualized as "packets" of energy termed *photons*. The energy of a photon is given by

$$E = h\nu \,, \tag{4.5}$$

where $h = 6.626 \times 10^{-34}$ joule-sec is Planck's constant and ν is the *frequency* of the radiation (in Hz). Viewed as a wave, EM radiation has a wavelength given by

$$\lambda = c/\nu \,, \tag{4.6}$$

where $c = 3.0 \times 10^8$ meters/sec is the speed of light.

Table 4.2 summarizes the frequencies, wavelengths, and photon energies of the EM spectrum relevant to medical imaging. Radio waves are very low frequency, very long wavelength EM radiations; their photon energy ranges from 10^{-10} to 10^{-2} eV. Since the photon energies of radio waves are below 13.6 eV, radio waves are not considered to be ionizing radiation. Visible light is moderate frequency, long wavelength EM radiation. The photon energy of light is about 2 eV, which is also not large enough to be considered ionizing. (See Problem 4.6 for consideration of ultraviolet light.) X-rays and gamma rays, on the other hand, are higher frequency, shorter wavelength EM radiation, having energies in the keVs to MeV range. Both x-rays and gamma rays are clearly ionizing radiations.

TABLE 4.2

The EM Spectrum

Frequency Range	Wavelengths	Photon Energies	Description
1.0×10^5–3.0×10^{10} Hz	3 km–0.01 m	413 peV–124 μeV	Radio waves
3.0×10^{12}–3.0×10^{14} Hz	100–1 μm	12.4 meV–1.24 eV	Infrared radiation
4.3×10^{14}–7.5×10^{14} Hz	700–400 nm	1.77–3.1 eV	Visible light
7.5×10^{14}–3.0×10^{16} Hz	400–10 nm	3.1–124 eV	Ultraviolet light
3.0×10^{16}–3.0×10^{18} Hz	10 nm–100 pm	124 eV–12.4 keV	Soft x-rays
3.0×10^{18}–3.0×10^{19} Hz	100–10 pm	12.4–124 keV	Diagnostic x-rays
3.0×10^{19}–3.0×10^{20} Hz	10–1 pm	124 keV–1.24 MeV	Gamma rays

Adapted from Johns and Cunningham, 1983.

Although it is suggested by Table 4.2, x-rays and gamma rays are *not* distinguished by their frequency or photon energies. Instead, they are distinguished by their point of origin. In particular, x-rays are created in the electron cloud of atoms while gamma rays are created in the nuclei of atoms, which in turn are undergoing reorganization due to radioactive decay. Thus, gamma rays are associated with radioactivity and x-rays are not. Although gamma rays tend to have higher frequencies (energies) than x-rays, there is a large overlap in the frequencies (energies) of x-rays and gamma rays used in medical imaging. Furthermore, once produced, x-rays and gamma rays behave the same in terms of their propagation properties and interaction with matter. Therefore, our understanding of the propagation and detection of EM waves in the x-ray–gamma ray photon energy range is essential background material for all the medical imaging modalities using ionizing radiation.

It turns out that both the production and detection of EM radiation over a large range of frequencies play major roles in medical imaging. Visible light is used in radiography to improve the efficiency of x-ray detection. The attenuation of x-rays is the primary mechanism used to create images in projection radiography and computed tomography. Finally, gamma rays are detected in order to locate radiotracers in nuclear medicine.

4.4 Nature and Properties of Ionizing Radiation

Particulate and EM ionizing radiations interact with the materials through which they are traveling, imparting energy to the material, losing energy from and redirecting their own radiation, and generating new types of particles and radiation. Generally speaking, the effects that we care about fall into two broad categories: (1) those that are used in imaging or that affect the imaging process and (2) those that are not used in imaging but contribute to *dose*—that is, have biological consequences. Specific concepts related to these two broad categories tend to apply separately to particulate radiation or EM radiation. Table 4.3 lists specific concepts related to ionizing radiation, positioned according to the categories for which they largely apply. We will now begin to discuss each of these concepts, starting with those important in forming and analyzing

TABLE 4.3

Radiation Concepts

	Imaging	Dose
Particulate	Bremsstrahlung Characteristic radiation *Positron annihilation** *Positron range*	Linear energy transfer Specific ionization
EM	Attenuation Photoelectric effect Compton scatter Characteristic radiation Polyenergetic	Air kerma Dose Dose equivalent Effective dose f-factor

*Italicized entries are discussed in Chapter 7.

medical images (column 1) and ending with those important in understanding the biological effects of radiation (column 2). Concepts that are relevant only in nuclear medicine are listed in italics and will be discussed in Chapter 7.

4.4.1 Primary Energetic Electron Interactions

The only particles of direct consequence to the *formation* of medical images are electrons and positrons.[1] Since positrons are used solely in nuclear medicine, we postpone discussion of the interaction of positrons with matter until Chapter 7. Here we consider only the interaction of *energetic electrons* with matter.

Energetic electrons interact and transfer energy to an absorbing medium by two modes: *collisional transfer* and *radiative transfer*. In collisional transfer, by far the most common type of interaction at the electron energies used in medical imaging, a (typically small) fraction of the electron's kinetic energy is transferred to another electron in the target medium with which it collides. As the affected atom returns to its original state, infrared radiation is generated, producing heat in the target medium. As shown in Figure 4.4(a), the incident electron's path may be redirected as a result of the collision, and many more interactions, both collisional and radiative, may subsequently take place, until the incident electron's kinetic energy is exhausted. Occasionally, a large amount of energy may be transferred to a struck electron, creating a new energetic electron, which forms a new path of ionization, called a *delta ray*, and a new set of collisional and radiative energy transfers.

In radiative transfer, the energetic electron's interaction with an atom produces x-rays. This can happen in two ways: *characteristic radiation* or *bremsstrahlung radiation*. In the generation of characteristic radiation, as shown in Figure 4.4(b), the incident electron collides with a K-shell electron, exciting

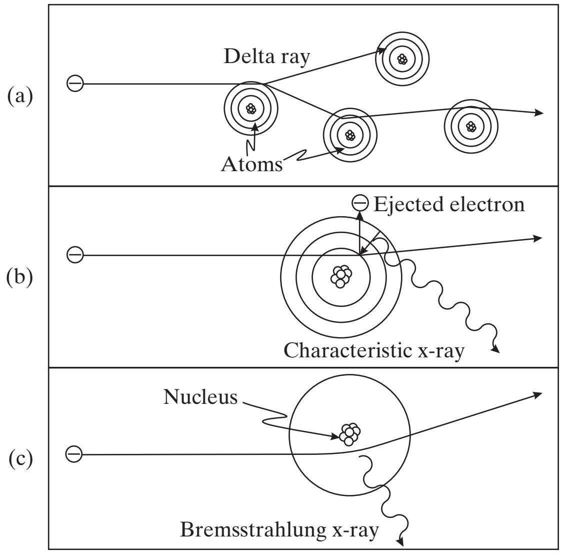


Figure 4.4
Energetic electrons can (a) collide with other electrons until they lose their energy; (b) eject a K-shell electron, generating a characteristic x-ray; or (c) be "braked" by a nucleus, generating bremsstrahlung radiation.

---

[1] Other particles, most notably alpha particles, are relevant in the *production* of radiotracers for use in nuclear medicine (see Chapter 7).

or ionizing the atom, and temporarily leaving a hole in that shell. (This can also happen to higher-orbit electrons but is not significant in imaging.) The K-shell hole is then filled by an electron from the L-shell, M-shell, or N-shell. Since the electron binding energy of the K-shell is higher than that of the L-, M-, or N-shell, there is a *loss* of energy as the electron fills the K-shell hole or vacancy. The loss of energy experienced by the outer-shell electron creates an EM photon. The x-ray is called a *characteristic x-ray*. The energy of the characteristic x-ray photon is exactly equal to the difference in electron binding energies between the two shells. In fact, the particular electron "subshells," which have slightly different energies according to quantum physics, are distinguished. Since these electron binding energies are determined by the particular atom and those transitions allowed by quantum physics, these x-ray energies are "characteristic" of particular atoms, which is the reason for the name *characteristic x-rays*. A given atom can be identified by the nature (energies and proportions) of the characteristic x-rays it generates.

Bremsstrahlung radiation is caused by the interaction of an energetic electron with the nucleus of an atom, as shown in Figure 4.4(c). As the electron approaches the nucleus, the positive charge of the nucleus attracts the electron, causing it to bend, as if to go into orbit around the nucleus. As the electron decelerates around the nucleus, it loses energy in the form of an EM photon, which comprises *bremsstrahlung radiation* (German for "braking radiation"). Unless the electron actually strikes the nucleus, the electron departs the atom with a kinetic energy reduced by the energy of the emitted photon. In the rare event of a collision with the nucleus, the electron is annihilated, and the photon emitted has energy equal to the kinetic energy of the incident electron. Otherwise, the emitted photon must have energy lower than that of the incident electron. The intensity of bremsstrahlung radiation increases with the energy of the incident electron and the atomic number of the atom with which it interacts.

As discussed in Chapter 5, bremsstrahlung radiation is the primary source of x-rays from an x-ray tube. This is because an x-ray tube works by accelerating electrons across a voltage potential in a vacuum. When the electrons strike the *target* or *anode*, which is usually tungsten, they lose energy by both collisional and radiative transfer; so heat, characteristic x-rays, and bremsstrahlung x-rays are produced. The x-ray energies produced by the characteristic and bremsstrahlung interactions differ in character, as shown in Figure 4.5. The bremsstrahlung spectrum, sometimes called the *continuous spectrum*, has its highest energy equal to the anode-to-cathode potential. In this figure, different potentials are considered: 45 kV, 61 kV, 80 kV, 100 kV, and 120 kV. For a given fixed potential, x-rays at the highest energies are produced only by the rare direct collisions between energetic electrons and anode nuclei. It is increasingly more likely for lower energy bremsstrahlung interactions to take place; it is approximately a linear increase with frequency decreasing from the maximum energy. At the lower energies, bremsstrahlung radiation is still produced, but is absorbed by the anode itself, so the spectrum goes to (effectively) zero.

Characteristic x-rays can be generated only if the incident electrons have sufficient energy to eject inner-shell electrons. For a tungsten anode, the K-shell binding energy is 69.5 keV, the L-shell binding energy is approximately 12 keV, and the M-shell binding energy is under 3 keV. Thus, the characteristic x-rays from tungsten are all under 70 keV, and include energies of approximately

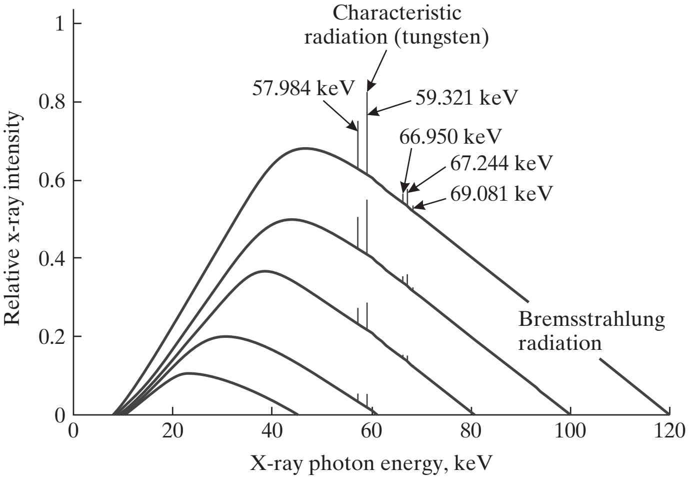


Figure 4.5
When energetic electrons bombard a target, two kinds of x-rays are produced: characteristic x-rays and bremsstrahlung x-rays. Low-energy x-rays of both types are absorbed by the medium.

58 keV (from an L-shell to K-shell transition), 67 keV (from an M-shell to K-shell transition), and 69 keV (from a very outer shell to K-shell transition), as shown in Figure 4.5. (Transitions to the L-shell or M-shell result in such low-energy characteristic x-rays that they are either absorbed within the tungsten anode or filtered out by the x-ray tube.)

4.4.2 Primary Electromagnetic Radiation Interactions

EM ionizing radiation interacts with matter through significantly different mechanisms than particulate ionizing radiation. The three main mechanisms by which EM ionizing radiation interacts with materials are (1) the photoelectric effect, (2) Compton scatter, and (3) pair production. Since a photon must have at least 1.02 MeV of energy for pair production to occur and since the range of photon energies in medical imaging is about 25–500 keV, we will consider only the photoelectric effect and Compton scattering here.

In both the photoelectric effect and Compton scattering, an incident x-ray photon interacts with the electron cloud of an atom. The primary difference between the two interactions is that in the photoelectric effect the photon is completely absorbed by the atom, while in Compton scattering the photon is not absorbed but instead loses energy and changes its direction. Both effects are important in imaging human tissues. The photoelectric effect is the primary mechanism that provides contrast between different types of tissues; Compton scattering is a primary mechanism for limiting the resolution of x-ray images, both in projection radiography and computed tomography. We now study both interactions in some detail.

Photoelectric Effect In the *photoelectric effect*, a photon with energy $h\nu$ interacts with the coulomb field of the nucleus of an atom, causing the ejection of an electron, usually a K-shell electron, from the atom. This process is illustrated in Figures 4.6(a) and (b). The incident photon is completely absorbed

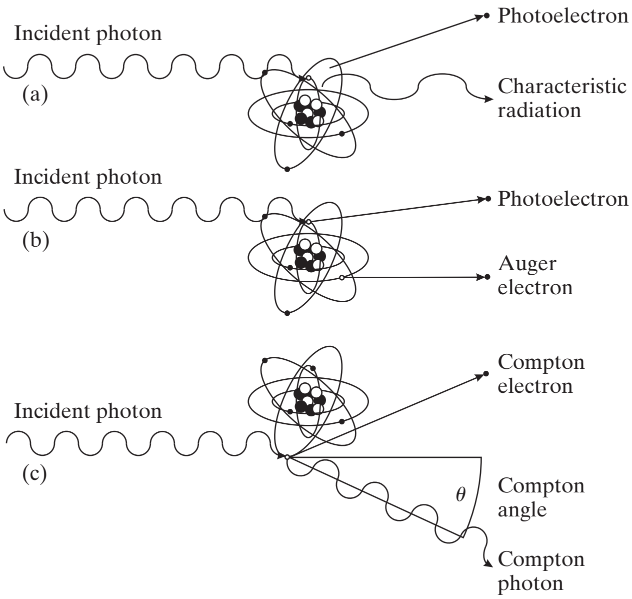


Figure 4.6
The photoelectric effect, shown in (a) and (b), and Compton scattering, shown in (c).

by the atom, and the ejected electron, called a *photoelectron*, propagates away with energy

$$E_{e^-} = h\nu - E_B \,, \tag{4.7}$$

where E_B is the binding energy of the ejected electron. The remaining atom is now an ion, having a hole, typically in the K-shell, which must be filled. The hole is filled by an electron transition from a higher orbit, which produces characteristic radiation, as discussed above and shown in Figure 4.6(a). Sometimes, the characteristic x-ray transfers its energy to an outer-orbit electron, called an *Auger electron*, which is ejected from the atom, accompanied by a readjustment of the remaining electron orbits, as shown in Figure 4.6(b). Photoelectrons and Auger electrons are energetic electrons that are free to interact with matter in the ways we described in the previous section. We will see below that these energetic electrons contribute to the detrimental biological effects of ionizing EM radiation by contributing radiation dose to the patient.

Compton Scattering In *Compton scattering*, as shown in Figure 4.6(c), a photon with energy $h\nu$ ejects a valence (outer-shell) electron, yielding a new energetic electron called a *Compton electron*. As a result of this interaction, the incident photon loses energy (to the Compton electron) and changes its direction. The energy of the scattered photon, the so-called *Compton photon*, is given by

$$h\nu' = \frac{h\nu}{1 + (1 - \cos\theta)h\nu/(m_0c^2)} \,, \tag{4.8}$$

where $m_0c^2 = 511$ keV is the energy equivalent to the rest mass m_0 of an electron, and θ is the angle through which the photon is scattered [see Figure 4.6(c)]. We see that the amount of energy remaining with the Compton photon depends on the scatter angle. The more the photon is deflected, the more energy it loses.

This implies that the maximum energy loss occurs when the photon is deflected 180° back towards the source, that is, a 180° *backscatter*. The kinetic energy of the Compton electron is given by

$$E_{e^-} = h\nu - h\nu' \,. \tag{4.9}$$

EXAMPLE 4.3

Compton scattering is usually undesirable in medical imaging. In planar scintigraphy (Chapter 8), the energy of a photon is used to determine whether it has been scattered prior to arrival at the detector.

Question Suppose a photon with energy $h\nu = 100$ keV is incident to some material and exits with energy $h\nu'$. A detector decides that the photon has not been scattered if $h\nu' > 98$ keV. What is the maximum angle by which the photon is scattered but is still being treated as a photon traveling along a straight path?

Answer The energy of the photon after Compton scattering is

$$\begin{aligned}
h\nu' &= \frac{h\nu}{1 + (1 - \cos\theta)h\nu/(m_0c^2)} \\
&= \frac{100\ \text{keV}}{1 + (1 - \cos\theta)100\ \text{keV}/(511\ \text{keV})} \\
&= 98\ \text{keV} \\
\theta &= \cos^{-1}\left(1 - \frac{511}{100 * 49}\right) \\
&= 26.4^\circ
\end{aligned}$$

So, a photon scattered by an angle up to 26.4° is still considered as traveling along a straight path by the detector. ■

Probability of EM Interactions We now consider the factors that make photoelectric and Compton events more likely to occur. There are many reasons why this subject is important. For example, the formation of a useful projection radiograph depends on the *differential attenuation* of an x-ray beam passing through a patient, which is caused by the photoelectric effect and Compton scattering, as we shall see in the next chapter. As another example, we observed above that the photoelectric effect causes complete absorption of an incident photon. Therefore, if we find a way to increase the likelihood of photoelectric events, we can *block* or *shield* objects (or patients, staff, and physicians) from the source of ionizing EM radiation. We should therefore be able to understand what makes lead a good x-ray shield while plastic is not; we should also be able to predict whether lead will shield high-energy photons better than low-energy photons. As another example, the reason why Compton scattering is detrimental to imaging is that it causes photons to deviate from straight-line paths. We would like to know the conditions under which Compton scattering is large so that instrumentation can be designed to compensate for its effects. We might choose to use x-ray energies that minimize the Compton scattering effect to image patients.

Consider a thin "slab" of material having a fixed thickness. Suppose that an incident photon can either pass through the material or experience a photoelectric or Compton event. What are the factors that make a photoelectric effect more likely? Since a photoelectric event occurs as a result of an interaction with the coulomb field of the nucleus of an atom and if there are more (positively charged) protons in the nucleus, the likelihood should increase. The probability of a photoelectric event is related to the atom's atomic number Z. In a compound composed of a variety of different atoms, we define an *effective atomic number* Z_{eff}, which characterizes the average atomic number in a certain sense. The probability of a photoelectric event is (approximately) proportional to Z_{eff}^4.[2]

The other factor affecting the probability of a photoelectric event is the incident photon energy. Intuitively, we might expect that high-energy photons are more penetrating—that is, they pass farther into an object before being absorbed. This intuition is correct; in fact, the probability of a photoelectric event (given all else held constant) is proportional to $1/(h\nu)^3$. Putting this together, we find

$$\text{Prob[photoelectric event]} \propto \frac{Z_{\text{eff}}^4}{(h\nu)^3}\,. \tag{4.10}$$

Finally, we note that when the photon energy rises above the binding energy of L-shell or K-shell electrons, experiments reveal that the probability of a photoelectric interaction increases abruptly, then begins to diminish as the energy is further increased. This sudden increase occurs because a large number of electrons has suddenly become available to be ejected from their host atoms. This property will be important in the use of *contrast agents* in radiography and computed tomography.

Now we ask another question: What are the factors that make a Compton scattering event more likely? Since Compton events occur with very loosely bound (or "free") electrons in the outer shells, what matters is the number of electrons per kilogram of material—the *electron density* (ED). The electron density is given by

$$\text{ED} = \frac{N_A Z}{W_m}\,, \tag{4.11}$$

where N_A is Avogadro's number (atoms/mole), Z is the atomic number (electrons/atom), and W_m is the molecular weight of the atom (grams/mole). As shown in Table 4.4, except for hydrogen, the electron density for various biological materials is nearly the same, 3×10^{26} electrons/kg (see Problem 4.7). Therefore, the probability of Compton scattering is nearly independent of the atomic number (actual or effective).

The other factor that might influence the probability of Compton scatter is the photon energy. There is a fairly complicated relationship described by the *Klein-Nishina formula*, which predicts that the probability of a Compton event generally decreases with increasing energy. However, this decrease is a very gradual one, taking place over the very highest energies of interest in diagnostic imaging. Over the range of energies of importance in x-ray imaging, the probability is reasonably constant. In summary, we conclude that

$$\text{Prob[Compton event]} \propto \text{ED}\,. \tag{4.12}$$

---

[2] For high Z materials (as opposed to human tissues), it is more nearly proportional to Z_{eff}^3.

TABLE 4.4

Physical Properties of Some Materials

Material	Density (kg/m^3)	Z_{eff}	Electron Density (electrons/kg)
Hydrogen	0.0899	1	5.97×10^{26}
Carbon	2,250	6	3.01×10^{26}
Air	1.293	7.8	3.01×10^{26}
Water	1,000	7.5	3.34×10^{26}
Muscle	1,040	7.6	3.31×10^{26}
Fat	916	6.5	3.34×10^{26}
Bone	1,650	12.3	3.19×10^{26}

From A. B. Wolbarst, p. 119, 1993.

TABLE 4.5

Photoelectric Versus Compton Interactions in Water

Photon Energy (keV)	Percentage of Compton Interactions	Percentage of Compton Energy
10.0	3.2	0.1
15.0	11.8	0.4
20.0	26.4	1.3
30.0	58.3	6.8
40.0	77.9	19.3
50.0	88.0	37.2
60.0	93.0	55.0
80.0	97.0	78.8
100.0	98.4	89.6
150.0	99.5	97.4

Adapted from Johns and Cunningham, p. 163, 1983.

Since photoelectric and Compton events affect diagnostic images differently, it is also important to understand the *relative* frequency of occurrence of these two events in body tissues. There are two ways to look at this: by a relative frequency or percentage of occurrence or by the percentage of energy deposited by a given type of event. Table 4.5 summarizes these two characterizations over the diagnostic range of energies. We see that the relative frequency of Compton interactions increases as diagnostic x-ray energy increases, becoming the dominant type of interaction above 30 keV. However, even at 60 keV, when Compton events occur over 90 percent of the time (whenever either event occurs), these events account for only 55 percent of the energy deposited in the tissue. This apparent discrepancy occurs because photoelectric events deposit *all* of their incident photon energy, while Compton events deposit only a fraction of their incident photon energy.

4.5 Attenuation of Electromagnetic Radiation

Attenuation is the process describing the loss of strength of a beam of EM radiation. Tissue-dependent attenuation is the primary mechanism by which contrast is created in radiography modalities. Although this is actually a statistical process, we study attenuation here as though it were deterministic. In Chapter 5, we will include the statistical aspects in a very natural way, as we study noise in projection radiographs.

4.5.1 Measures of X-Ray Beam Strength

In radiography, we are generally concerned with a brief burst of x-rays, also called an x-ray *beam*. For many reasons, we are also concerned with the "strength" of the x-ray burst. For example, in the design of x-ray detection systems used to form images, it is critical to characterize the strength of the x-ray beam in order to characterize the inherent noise in the system and to adjust the dynamic range of the detection system. Measures of x-ray beam strength are also used in estimating the (adverse) biological effects of ionizing radiation. This is critical because ionizing radiation is potentially harmful. At high doses, it can cause burns or cataract formation, for example. Even at low doses, it can increase the risk of cancer. It turns out that there are many ways to define and measure the strength of x-rays. In this section, we consider several of the most common and important measures of x-ray beam strength.

The conceptually simplest measure of the strength of an x-ray burst is simply the number of photons N in the burst. It is also important to consider the *area* over which these photons are spread. Accordingly, we define the *photon fluence* Φ as the number of photons N per unit area A,

$$\Phi = \frac{N}{A}\,, \tag{4.13}$$

where the area is oriented at a right angle to the direction of the radiation beam propagation. It is sometimes important to account for *time*, since measurements may take place over a fixed interval Δt. Accordingly, *photon fluence rate* ϕ is defined as

$$\phi = \frac{N}{A\Delta t}\,. \tag{4.14}$$

It may be of interest to know how much energy the burst is carrying and could deposit into a material if totally absorbed. If each photon in the burst has the same energy $h\nu$—that is, the beam is monoenergetic—then the total energy of the burst is simply $Nh\nu$. We then define the *energy fluence* Ψ and *energy fluence rate* ψ as

$$\Psi = \frac{Nh\nu}{A}\,, \tag{4.15}$$

$$\psi = \frac{Nh\nu}{A\Delta t}\,. \tag{4.16}$$

The energy fluence rate is also known as the *intensity* of an x-ray beam, and it is often given the symbol I. From (4.14) and (4.16), we see that

$$I = E\phi\,, \tag{4.17}$$

where $E = h\nu$. Intensity has units of energy per unit area per unit time.

In x-ray imaging, a photon burst is always polyenergetic, since it arises mainly from bremsstrahlung. In principle, if only the photon count were important, photon fluence rate could still be used as a valid measure of radiation strength. However, these measures are not very important in practice, since photons are not counted in x-ray imaging modalities, but instead their total energy is measured. It is more useful, therefore, to capture the idea of intensity for polyenergetic sources.

Since each x-ray photon carries its own discrete energy, a plot of N as a function of E for polyenergetic sources is a *line spectrum*. Because of the random nature of bremsstrahlung, the fine details of a line spectrum change with each photon burst; however, the line density—for example, number of photons per unit energy as a function of E—remains constant for a given source. The *x-ray spectrum* $S(E)$ is this line density per unit area per unit time. Examples of x-ray spectra were shown in Figure 4.5. The energies at which $S(E)$ is large have larger numbers of x-ray photons per unit energy.

Since the integral of a line density yields the number of lines, from the definition of $S(E)$ and (4.14), the integral of the spectrum yields the photon fluence rate

$$\phi = \int_0^\infty S(E')\,dE' , \tag{4.18}$$

a quantity that does not depend on energy. Furthermore, by analogy to (4.17), it is natural to define the intensity of a polyenergetic source as

$$I = \int_0^\infty E' S(E')\,dE' . \tag{4.19}$$

EXAMPLE 4.4

It is often desirable to model a polyenergetic x-ray beam as a monoenergetic source.[3]

Question What energy would a (hypothetical) monoenergetic source have to be in order to produce the same intensity as the (true) polyenergetic source using the same number of photons?

Answer For a given area A and time interval Δt, the number of photons in the polyenergetic source is

$$N_p = A\Delta t \int_0^\infty S(E')\,dE' .$$

The intensity of the polyenergetic source is

$$I_p = \int_0^\infty E' S(E')\,dE' .$$

The equivalent monoenergetic beam has intensity

$$I_p = E \frac{N_p}{A\Delta t} .$$

---

[3] In a monoenergetic beam, all photons have the same energy, $h\nu$.

Therefore, the energy of this equivalent beam must be

$$E = \frac{\int_0^\infty E' S(E')\, dE'}{\int_0^\infty S(E')\, dE'} \,.$$

This expression should look vaguely familiar; it is the expression for center of mass of a sheetlike object having mass density $S(E)$. Alternatively, the numerator is the expression for the mean of a random variable if $S(E)$ were its pdf. The denominator normalizes $S(E)$ so that the ratio effectively does compute the mean. The energy computed this way is sometimes called the *effective energy* of a polyenergetic source. We will see an alternate definition in Chapter 6. ■

4.5.2 Narrow Beam, Monoenergetic Photons

Consider a beam of N monoenergetic photons incident upon a thin slab of homogeneous material, as shown in Figure 4.7(a). The depicted geometry is called *narrow beam geometry* because the photon beam is no wider than the detector used to detect or count the photons. In this case, if the slab were not present (and ignoring statistical effects and problems of detector efficiency), the detector would record N photons. With the slab present, however, some photons will be absorbed within the slab by the photoelectric effect, and other photons will be deflected away from the detector by Compton events. Overall, the detector count N' will be smaller than N, which is the primary characteristic of *attenuation*, which we now explore in more detail.

Suppose n photons are "lost" in the above experiment. Provided that Δx is small, we would expect that doubling Δx should double n, and doubling N should also double n. In other words, n is proportional to both N and Δx, or

$$n = \mu N \Delta x \,, \tag{4.20}$$

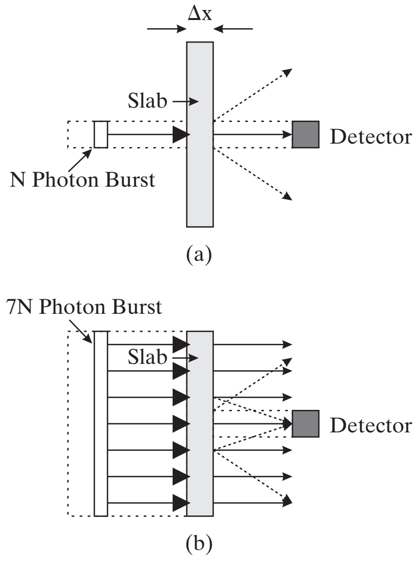


Figure 4.7
Two geometries by which to study attenuation: (a) narrow beam geometry and (b) broad beam geometry.

where μ is a constant of proportionality called the *linear attenuation coefficient*. Rearranging (4.20) yields a definition of the linear attenuation coefficient,

$$\mu = \frac{n/N}{\Delta x}, \tag{4.21}$$

which can be interpreted as the fraction of photons that are lost per unit length.

The change in the number of photons upon interaction with the slab is

$$\begin{aligned} \Delta N &= N' - N \\ &= -n \\ &= -\mu N \Delta x . \end{aligned} \tag{4.22}$$

Letting the slab become incrementally thin and treating N as a continuous quantity leads to the differential equation

$$\frac{dN}{N} = -\mu \, dx, \tag{4.23}$$

which can be integrated (see Problem 4.9) to yield

$$N = N_0 e^{-\mu \Delta x}, \tag{4.24}$$

where N_0 is the number of photons at $x = 0$. This equation can be called the *fundamental photon attenuation law*; it is of central importance throughout all of x-ray radiography. It should be noted that in the monoenergetic case, this law can also be written in terms of intensity as

$$I = I_0 e^{-\mu \Delta x}, \tag{4.25}$$

where I_0 is the intensity of the incident beam.

Using (4.24), we can compute the fraction of (monoenergetic) photons that will be stopped or transmitted by a noninfinitesimal layer of homogeneous material. One question can be asked: What thickness of a given material will attenuate half of the incident photons? This thickness, called the *half value layer* (HVL), satisfies

$$\frac{N}{N_0} = \frac{1}{2} = e^{-\mu \mathrm{HVL}} . \tag{4.26}$$

A simple series of manipulations of (4.26) yields

$$\mathrm{HVL} = \frac{\ln 2}{\mu} = \frac{0.693}{\mu} . \tag{4.27}$$

EXAMPLE 4.5

We will see in Chapter 7 that 140 keV gamma rays are generated by the radioactive decay of technetium-99m, and that sodium iodide crystals are used to detect such gamma rays. Assume that the HVL of sodium iodide at 140 keV is 0.3 cm.

Question What percentage of gamma rays will pass right through a 1.2 cm sodium iodide crystal?

Answer We need to determine

$$\begin{aligned}\text{Percent transmitted} &= 100\% \times \frac{N}{N_0} \\ &= 100\% \times e^{-\mu 1.2\ \text{cm}} .\end{aligned}$$

Since $\mu = \ln 2/\text{HVL}$, we have

$$\begin{aligned}\text{Percent transmitted} &= 100\% \times e^{-\ln 2(1.2\ \text{cm})/(0.3\ \text{cm})} \\ &= 100\% \times e^{-4\ln 2} \\ &= 6.25\% .\end{aligned}$$

■

Suppose the slab is not homogeneous. In this case, the linear attenuation coefficient depends on the position x, and we are faced with solving

$$\frac{dN}{N} = -\mu(x)\,dx . \tag{4.28}$$

Straightforward integration yields the number of photons at position x,

$$N(x) = N_0 \exp\left\{-\int_0^x \mu(x')\,dx'\right\} , \tag{4.29}$$

where x' is a dummy variable for integration. The analogous relationship holds for intensity as well,

$$I(x) = I_0 \exp\left\{-\int_0^x \mu(x')\,dx'\right\} . \tag{4.30}$$

This form can be thought of as the *integral form* of the fundamental x-ray attenuation law. This equation is the most important physical model for projection radiography and computed tomography. It is also a valid model for the attenuation of gamma rays in nuclear medicine modalities.

4.5.3 Narrow Beam, Polyenergetic Photons

In the narrow beam case, photons (of any energy) are either completely absorbed or scattered away from the detector. Therefore, we can consider the attenuation principles in the narrow beam, polyenergetic case to be the same as in the monoenergetic case, except that they take place at each energy independently.

In general, the linear attenuation coefficient is different for different materials, and it also varies as a function of energy for the same material. A plot showing these relationships for bone, muscle, and fat is given in Figure 4.8. This plot shows that bone is more attenuating than muscle, which is in turn

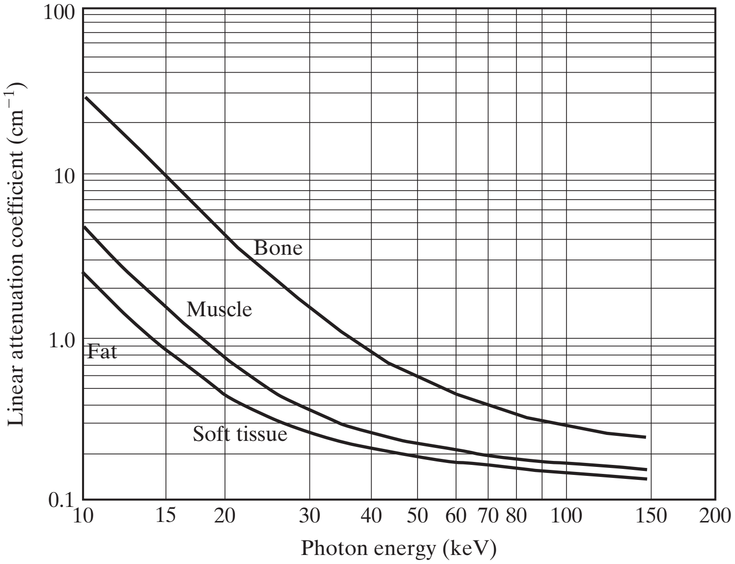


Figure 4.8
Linear attenuation coefficient for bone, muscle, and fat as a function of incident x-ray photon energy.

more attenuating than fat. This plot shows as well that x-rays experience less attenuation at higher energies—that is, they are more penetrating at higher energies. For a polyenergetic x-ray beam, we must treat the linear attenuation coefficient of a given material as a function of E—that is, $\mu(E)$.

Consider a homogeneous slab of material of thickness Δx with an incident x-ray beam having spectrum $S_0(E)$. The spectrum leaving the slab is attenuated in a fashion identical with that of the monoenergetic case [see (4.25)], except that the linear attenuation coefficient depends on E,

$$S(E) = S_0(E)e^{-\mu(E)\Delta x}\,. \tag{4.31}$$

For a heterogeneous slab, the linear attenuation coefficient is dependent on both position along the line and energy. In this case, the integral form of the attenuation law obeys

$$S(x; E) = S_0(E) \exp\left\{-\int_0^x \mu(x'; E)\, dx'\right\}\,. \tag{4.32}$$

If the overall intensity of the beam is of interest, then we can use (4.19) together with (4.31) and (4.32) to yield

$$I = \int_0^\infty S_0(E')E' \exp\left\{-\mu(E')\Delta x\right\} dE' \tag{4.33}$$

and

$$I(x) = \int_0^\infty S_0(E')E' \exp\left\{-\int_0^x \mu(x'; E')\, dx'\right\} dE'\,. \tag{4.34}$$

Although (4.34) is a valid model for heterogeneous materials, including the body, we will not find it useful in developing imaging equations for x-ray modalities. The energy dependence of μ, while useful in helping to understanding the basic imaging properties, is intractable from a mathematical standpoint. The concept of effective energy (see Example 4.4) is more useful, as we shall see.

4.5.4 Broad Beam Case

Now consider the *broad beam geometry* in Figure 4.7(b), which should be compared with the narrow beam geometry in Figure 4.7(a). In (b), the x-ray beam is seven times wider and has $7N$ as many photons (the number 7 being used here for illustration purposes). When the slab is removed, the detector, having the same width as in (a), will detect only N photons, since the remaining photons miss the detector. When the slab is in place, photons previously directed at the detector might be absorbed by photoelectric interactions or scattered by Compton events—just as they are in the narrow beam case. However, there is an additional possibility here: photons from outside the detector's line-of-sight geometry might get scattered toward the detector by Compton interactions.

Thus, in the broad beam case, more photons are generally detected than is predicted by a monoenergetic, narrow beam analysis. The general attenuation law [(4.24) and (4.25)] does not hold in this case, and the assumption of straight propagation of x-rays is also violated. Furthermore, the photons that comprise the detected burst are no longer monoenergetic, since the Compton scattering process reduces photon energy. In such a situation, the average or effective energy of the x-ray beam is reduced, a process called *beam softening*.

Fortunately, most x-ray imaging modalities use *detector collimation*, which reduces the number of x-rays from nonnormal directions that can hit the detectors. Therefore, from an imaging standpoint, the narrow beam geometry assumption can be viewed as fairly accurate. We make the assumption of narrow beam geometry throughout this book. From the point of view of calculating the dose to the patient or determining how much lead is needed to shield a radiologist or technician, however, broad beam geometry must be considered. After all, there is no detector collimation on the "human detectors" involved in the imaging process.

4.6 Radiation Dosimetry

There are many ways to report the presence and amount of radiation. We have already described the ideas related to photon and energy flow via *fluence* and *intensity* concepts. We did so with a focus on the production of medical images. However, radiation also produces adverse biological effects. Here, we are now more concerned with what EM radiation *does* than what it *is*. That is the essence of the concept of *dose*.

4.6.1 Exposure

By definition, ionizing radiation is capable of ionizing the hydrogen atom. Therefore, it is capable of producing ions in air. It is straightforward to build and calibrate an *ionization chamber* that measures the current produced between two plates held at a fixed potential due to radiation producing ions in the air between the two plates.

The term *exposure*, which is given the symbol X, refers to the number of ion pairs produced in a specific volume of air by EM radiation. The International System unit for exposure is coulombs per kilogram of air (C/kg); however, a more useful unit in medical imaging is the roentgen (R), which equals 2.58×10^{-4} C/kg. The conversion between the two is 1 C/kg $=$ 3,876 R.

EXAMPLE 4.6

For a point source of radiation, the exposure at a distance d from the source follows an inverse square law.

Question If the exposure at $d = 30$ cm from a point source is 1 R, what is the exposure at $d = 5$ cm from the source?

Answer The exposure at $d = 5$ cm is 36 ($(30/5)^2 = 36$) times that at $d = 30$ cm. So the exposure at $d = 5$ cm is $X_5 = 36X_{30} = 36$ R. ■

4.6.2 Dose and Kerma

As ionizing EM radiation passes through a material, it deposits energy into the material by both the photoelectric effect and Compton scattering. This is the concept of *dose*. The unit of *absorbed dose* is the *rad*, which is defined as the absorption of 100 ergs per gram of material. It is given the symbol D. Notice that this unit refers to an energy-deposition concentration rather than a total amount of energy. The SI unit for absorbed dose is the *gray* (Gy); 1 Gy $=$ 1 J/kg $=$ 100 rads. For the degree of accuracy required in biomedical dosimetry, one roentgen of exposure yields one rad of absorbed dose in soft tissue.

There is a another quantity closely related to dose, called *kerma*, which is given the symbol K. Kerma is defined as the amount of energy per unit mass imparted directly to the electrons in a given material. It is also measured in units of gray (Gy). As a practical matter, at diagnostic x-ray energies, kerma and dose are essentially equivalent. When kerma is being used in air for calibration purposes, it is referred to as *air kerma* and is given the symbol K_{air}.

4.6.3 Linear Energy Transfer (LET)

The *linear energy transfer* is a measure of the energy transferred by radiation to the material through which it is passing per unit length; higher LET radiation tends to produce greater adverse biological consequences. *Specific ionization* (SI) is the number of ion pairs formed per unit length. LET and SI are related to each other by the average amount of energy required to form one ion pair (often referred to as W), which is a characteristic of the material.

4.6.4 The *f*-Factor

It is often useful to be able to measure exposure but to express the results as dose to an individual in that radiation field. Because of the way in which the roentgen and rad are defined, a relationship between exposure in air and dose in air may be easily derived: 1 R = 0.87 rad. To compute the dose to a material other than air, the f-factor is used as a conversion,

$$D = fX\,. \tag{4.35}$$

The f-factor is defined as

$$f = 0.87\frac{(\mu/\rho)_{\text{material}}}{(\mu/\rho)_{\text{air}}}\,, \tag{4.36}$$

where μ is the linear attenuation coefficient and ρ is the mass density of the material (numerator) and air (denominator). The quantity (μ/ρ) is called the *mass attenuation coefficient*.

4.6.5 Dose Equivalent

We are all exposed to ionizing radiation, from cosmic rays, radioactivity in building materials and soil, and even from radioactivity in our bodies. Although dose itself is a well-defined quantity, different types of radiation, when delivering the same dose, can actually have different effects on the body. In order to account for this, the concept of *dose equivalent*, given the symbol H, is used. Dose equivalent is defined as

$$H = DQ\,, \tag{4.37}$$

where Q is the so-called *quality factor*, a property of the type of radiation used. For example, the quality factor of x-rays, gamma rays, electrons, and beta particles is $Q \approx 1$, whereas that for neutrons and protons is $Q \approx 10$, and for alpha particles is $Q \approx 20$.

Since $Q \approx 1$ for radiation used in medical imaging, dose equivalent is equal to dose. When D is measured in rads H is considered to have the units *rems*. In SI units, the dose is measured in grays. For a dose of 1 gray and $Q = 1$ the dose equivalent H is 1 sievert (Sv).

EXAMPLE 4.7

Consider a chest x-ray at an energy of 20 keV. For simplicity ignore all tissues except the lung.

Question If we want to keep the absorbed dose below 10 mrads (i.e., 0.01 rads), what should be the limit on the exposure?

Answer From Table 4.6, we find that at 20 keV,

$$(\mu/\rho)_{\text{air}} = 0.78\ \text{cm}^2/\text{g}\,,$$

and

$$(\mu/\rho)_{\text{lung}} = 0.83\ \text{cm}^2/\text{g}\,.$$

TABLE 4.6

X-Ray Mass Attenuation Coefficients for Some Materials

Material	20 keV (cm^2/g)	100 keV (cm^2/g)
Air	0.7779	0.1541
Water	0.8096	0.1707
Bone	0.4001	0.1855
Muscle	0.8205	0.1693
Lung	0.8316	0.1695
Brain	0.8281	0.1701

Source: Hubbell and Seltzer, NIST online tables.

So the f-factor is

$$f = 0.87 \frac{(\mu/\rho)_{\text{lung}}}{(\mu/\rho)_{\text{air}}} = 0.93\,.$$

The absorbed dose is related to the exposure by (4.35). If we want to limit the dose to be under 10 mrads, the exposure must be under $10/0.93 = 10.8$ mR. ■

4.6.6 Effective Dose

For the purpose of relating dose of ionizing radiation to risk, an extension of the dose equivalent is used to express dose as that which would have been received if the whole body had been irradiated uniformly. The *effective dose* is obtained as the sum of dose equivalents to different organs or body tissues weighted in such a fashion as to provide a value proportional to radiation-induced somatic and genetic risk even when the body is not uniformly irradiated. This can be expressed as

$$D_{\text{effective}} = \sum_{\text{organs}} H_j w_j \tag{4.38}$$

where $D_{\text{effective}}$ is effective dose, H_j is the dose equivalent for organ j, and w_j is the weighting factor for organ j. The sum is over all organs; note that

$$\sum_{\text{organs}} w_j = 1. \tag{4.39}$$

In this fashion, risks may be compared for different radiations and different target tissues.

On average our annual individual effective dose (exclusive of medical imaging procedures) is about 300 mrem. Medical imaging procedures produce doses over a wide range, depending on the specific procedure. At the low end, a typical chest x-ray might result in a dose of 3–4 mrem; at the high end, a fluoroscopic study might yield several rem.

The main risk from ionizing radiation at the doses involved in medical imaging is cancer production (*radiogenic carcinogenesis*). In practice, we assume that any dose of radiation increases the risk of getting cancer, with larger doses proportionately increasing risk. Of importance, the physician and patient together usually make the decision that the medical benefits of the imaging procedure outweigh any potential risks.

4.7 Summary and Key Concepts

Ionizing radiation is used in several imaging modalities, including projection radiography and CT. The interaction of particulate radiation creates x-rays used in these modalities; the interaction of the x-rays in the patient creates the radiograph. In this chapter, we presented the following key concepts that you should now understand:

1. *Ionization* is the ejection of an orbiting electron from an atom; *ionizing radiation* has sufficient energy to produce ionization.
2. Ionizing radiation may be *particulate* or *electromagnetic*; the main ionizing radiations of interest in medical imaging are *x-rays*, *gamma rays*, *energetic electrons*, and *positrons*.
3. Particulate ionizing radiation transfers energy via *collisional transfer* and *radiative transfer* (which results in *bremsstrahlung x-rays*).
4. The *probability of radiative transfer* increases with increasing effective atomic number of the material through which the particulate radiation passes.
5. EM ionizing radiation transfers energy in medical imaging applications via the *photoelectric effect* or *Compton scattering*.
6. The *probability of electromagnetic interactions* depends on the photon's energy and the effective atomic number, density, and molecular weight of the material through which the EM radiation passes.
7. *Attenuation* of the intensity of a beam of EM radiation by some material is described by a monoexponential relation and is a function of the thickness of the material and the *linear attenuation coefficient*, which itself depends on characteristics of the material and photon energy.
8. *Radiation dose* to tissues is characterized by *effective dose*, which itself includes considerations of deposited energy, the biological effectiveness of a given ionizing radiation, and the relative radiosensitivity of different tissues.

Further Reading

Attix, F.H. *Introduction to Radiological Physics and Radiation Dosimetry*. New York, NY: Wiley, 1986.

Bushberg, J.T., Seibert, J.A., Leidholdt, E.M., and Boone, J.M. *The Essential Physics of Medical Imaging*, 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins, 2012.

Carlton, R.R. and Adler, A.M. *Principles of Radiographic Imaging: An Art and a Science*, 5th ed. Clifton Park, NY: Delmar Cengage Learning, 2012.

Hubbell, J.H. and Seltzer, S.M. *Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients*, version 1.4.

Online at *http://physics.nist.gov/xaamdi*, Gaithersburg, MD: National Institute of Standards and Technology, 2004.

Johns, E.J. and Cunningham, J.R. *The Physics of Radiology*. 4th ed. Springfield, IL: Charles C Thomas Publisher, 1983.

Wolbarst, A.B. *Physics of Radiology*, 2nd ed. Norwalk, CT: Appleton and Lange, 2005.

Problems

Physics of Atoms

4.1 The rest masses of a proton, neutron, and electron are $1.67262171 \times 10^{-27}$kg, $1.67492728 \times 10^{-27}$kg, and $9.1093826 \times 10^{-31}$kg, respectively.

- **(a)** The atomic mass unit (amu) is defined as one twelfth of the actual mass of a carbon-12 atom (as a neutral atom in its nuclear and electronic ground state). The mass of a carbon-12 atom is less than the sum of the masses of its constituents; the difference is the mass defect. Calculate the mass defect (in kg or amu).
- **(b)** The binding energy is the energy equivalent to the mass defect. What is the binding energy of carbon-12?

4.2 **(a)** Calculate the mass-equivalent energy of an electron at rest.

- **(b)** Determine the voltage potential difference between a cathode and anode in which electrons accelerated between the two will obtain speeds equal to 1/10 the speed of light.
- **(c)** Using relativistic equations, determine the speed of an electron that is accelerated across a 120 kV potential.

4.3 Show that when a particle's speed v is much smaller than the speed of light, its kinetic energy can be computed using the formula $\mathrm{KE} = mv^2/2$.

Ionizing Radiation

4.4 Compare characteristic radiation and bremsstrahlung radiation.

4.5 **(a)** Explain why radiation with energy smaller than 13.6 eV is not considered ionizing.

- **(b)** What are the differences between ionization and excitation?

4.6 Ultraviolet light is defined as EM waves having wavelengths in the range 4–400 nanometers.

- **(a)** Determine the frequency range of ultraviolet light.
- **(b)** Determine the photon energy range of ultraviolet light.
- **(c)** Determine whether ultraviolet light is ionizing radiation or not.

4.7 **(a)** Explain why the electron density of hydrogen is very nearly 6×10^{26} electrons/kg.

- **(b)** Explain why the electron density of all other low atomic number materials is nearly 3×10^{26} electrons/kg.
- **(c)** Speculate why the electron density of materials might differ slightly from 3×10^{26} electrons/kg.

Attenuation of EM Radiation

4.8 (**a**) Calculate the thickness of shielding material needed to block out 99.5% of incident radiation for a material with linear attenuation coefficient μ.

(**b**) Range is defined as the reciprocal of μ, the linear attenuation coefficient. It represents the distance at which the intensity of a beam has been reduced to $1/e$ of its original value. What approximate order of magnitude would you want the range of an ionizing beam in tissue to be? Microns? Millimeters? Centimeters? Meters? Kilometers? Why?

4.9 Prove Equation (4.24).

4.10 A bar phantom of thickness 0.4 cm is uniformly irradiated by monoenergetic x-ray photons, and a screen is placed behind the phantom to detect the x-ray photons. The bars of the phantom are made from a material that has an HVL of 0.1 cm. In the space between the bars, x-ray photons pass without attenuation. Assuming the intensity of the image is proportional to the x-ray photons that hit the screen in a unit area, what is the contrast of the resulting image?

4.11 Compton scattering of photons causes image blurring since the secondary x-ray photon diverges from the primary beam. You are designing an x-ray detector system and wish to eliminate all photons that have scattered more than 25 degrees in an attempt to improve the resulting image quality. You are using a mono-energetic x-ray source that emits photons with wavelength 8.9×10^{-2} angstroms. Your detector is capable of discriminating the energy of the incoming photon by examining the pulse height. What range of photon energies will your system accept?

4.12 Suppose photons with energy $h\nu = 100\,\text{keV}$ are incident upon some material and some are scattered.

(**a**) A detector decides that a photon has not been scattered if its energy satisfies $h\nu' > 99\,\text{keV}$. What is the maximum scattering angle such that a photon is still treated as traveling along a straight path?

(**b**) You are designing a detector system to eliminate all photons that have been scattered more than 25°. What range of photon energies will your system accept?

4.13 Sodium iodide (NaI) crystals can be used to detect x-rays. Suppose the linear attenuation coefficient of an NaI crystal is $2.31\ \text{cm}^{-1}$ at 140 keV.

(**a**) What is the HVL of this NaI crystal at 140 keV?

(**b**) Suppose an x-ray photon with energy $E = 140\,\text{keV}$ hits the NaI crystal and scatters by 90°. What is the energy E' of the scattered x-ray photon?

(**c**) Is the scattered photon more or less likely to be absorbed than the incident photon? Explain. (Note that iodine has a single K-edge at 33 keV.)

4.14 To block out 90% of the incident radiation, a 1.5 cm thickness of a particular shielding material is needed.

(**a**) What is the HVL of this material?

(**b**) Suppose a photon with energy $E = 102.2$ keV is scattered by 90° within the shielding material. What is the energy E' of the scattered photon?

4.15 Consider a narrow beam geometry. X-rays with intensity I_0 are incident upon a 1 cm thick material for which $\mu = 0.3\ \text{cm}^{-1}$.

(**a**) What intensity will be detected?

(**b**) What thickness of the same material will block 1/2 of the incident x-rays?

(**c**) If the broad beam geometry is used instead, will the detected intensity be larger or smaller? Explain.

Radiation Dosimetry

4.16 Consider a chest x-ray given by a point source operating at an energy of 20 keV. The exposure at $d = 1$ cm is 10 R. If we want to keep the dose equivalent to be under 10 mrems, how far should the patient be away from the x-ray source? (Ignore the tissues other than the lung, and ignore the size of the lung.)

4.17 Consider the case of hand x-ray imaging. For simplicity, assume that the hand consists of only bone and muscle, and the weighting factors for the bone and the muscle in the hand are $w_{\text{bone}} = w_{\text{muscle}} = 0.002$. What is the effective dose for an exposure of $X = 1$ R for x-rays at 20 keV?

Projection Radiography

CHAPTER 5

5.1 Introduction

Projection radiography, sometimes called *conventional radiography*, refers to the most commonly used method of medical imaging utilizing x-rays. A conventional *radiograph* represents a projection of the three-dimensional volume of the body onto a two-dimensional imaging surface. The term *projection* will be defined more rigorously shortly, but conceptually, the projection radiograph represents the transmission of the x-ray beam through the patient, weighted by the integrated loss of beam energy due to scattering and absorption in the body. The primary radiographic image is a two-dimensional display of these transmitted intensities. It is useful to think of a projection radiograph as a *shadow* cast by a semitransparent body illuminated by x-rays.

There are many types of projection radiography systems in common use. While their overall system designs have common elements, there are different clinical requirements, which make the details differ. Three examples, representative of the breadth of possibilities required by different clinical requirements, are shown in Figure 5.1.

Overall, the advantages of projection radiographic systems include short exposure time (0.1 second), the production of a large area image (e.g., 14 × 17 in), low cost, low radiation exposure (3–4 mR for a chest radiograph, roughly equivalent to about one-hundredth of the annual background dose exclusive of medical imaging procedures), and excellent contrast and spatial resolution. Projection radiography is used to screen for pneumonia, heart disease, lung disease, bone fractures, cancer, and vascular disease. Typically, the *chest x-ray* is the most common imaging examination performed in a hospital. Its major limitation is the lack of depth resolution—superimpositions of shadows from overlying and underlying tissues sometimes "hide" important details and limit contrast.

In this chapter, we describe the components of standard projection radiography systems, develop mathematical models to characterize image formation, describe the factors affecting image quality, and look at several applications and special techniques for projection radiography systems.

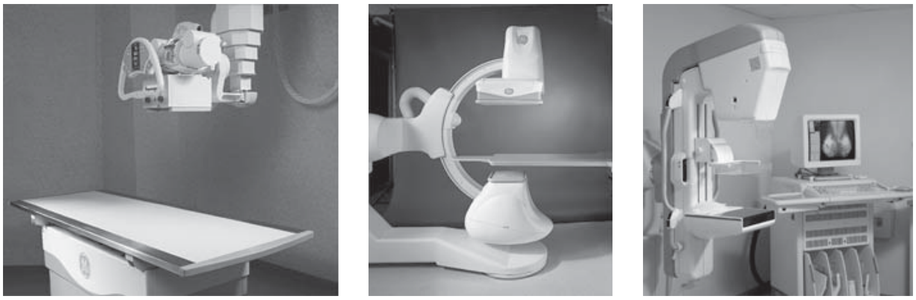

(a) (b) (c)

Figure 5.1
(a) A general projection radiography system; (b) a fluoroscopy system; and (c) a mammography system. Courtesy of GE Healthcare.

5.2 Instrumentation

A diagram of a conventional projection radiographic system is shown in Figure 5.2. The x-ray tube generates a short pulse of x-rays as a *beam* that travels through the patient. X-ray photons that are not absorbed within the patient or scattered outside the region of the detector impinge upon the large area detector, ultimately creating an image on a sheet of film or electronic detector. We now examine each element of this imaging system in detail.

5.2.1 X-Ray Tubes

X-rays are used in projection radiography and computed tomography because they penetrate the body well and their wavelengths are small enough for high-resolution imaging. X-rays are generated using an *x-ray tube*, as shown in Figure 5.3. The operation of an x-ray tube is shown schematically in Figure 5.4. A current, typically 3–5 amperes at 6–12 volts, is passed through a thin thoriated tungsten wire, called the *filament*, contained within the *cathode assembly*. Electrical resistance causes the filament to heat up and discharge electrons in a cloud around the filament through a process called *thermionic emission*. These electrons are now available to flow (i.e., be accelerated) toward the anode when the anode voltage is applied, producing the *tube current*, which is referred to as the *mA*. The filament current directly controls the tube current

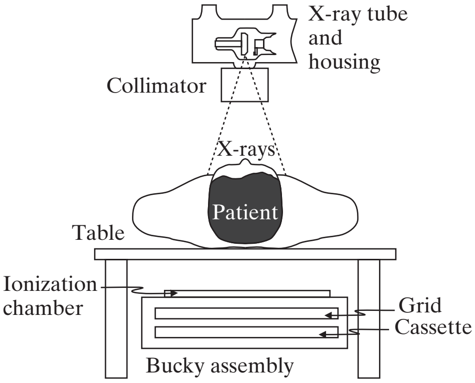


Figure 5.2
A conventional projection radiographic system.

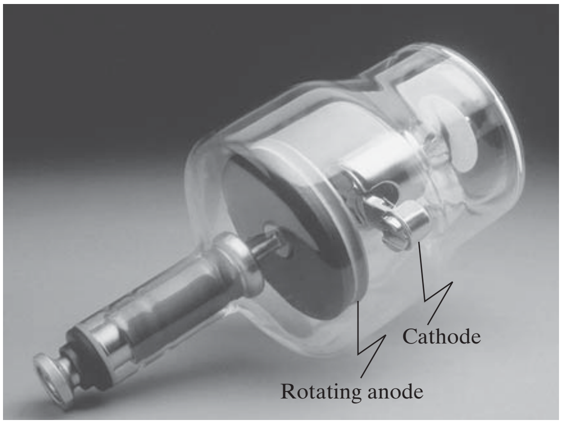


Figure 5.3
An x-ray tube.

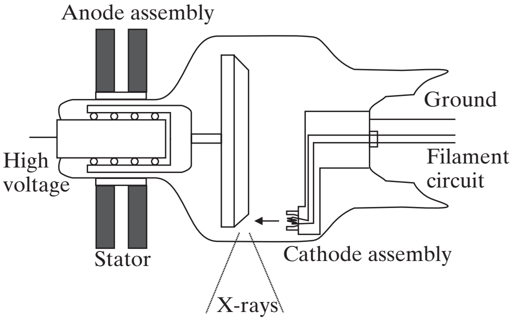


Figure 5.4
Schematic diagram of an x-ray tube.

because the current controls filament heat, which in turn determines the number of discharged electrons. The x-ray control console is calibrated according to the tube current, which typically ranges between 50 and 1,200 mA.

Once the filament current is applied, the x-ray tube is primed to produce x-rays. This is accomplished by applying a high voltage, the *tube voltage* or *kVp*, between the anode and cathode for a brief period of time. The tube voltage is typically generated by transforming the alternating current line voltage to a higher voltage and then rectifying this voltage. The abbreviation kVp refers to the *peak* kilovoltage applied to the anode; the voltage ripple (temporal variation) below the peak value depends on the specific type of high-voltage generator in use. Typical values for the tube voltage lie in the range 30–150 kVp.

While the tube voltage is being applied, electrons within or near the cathode are accelerated toward the anode. The *focusing cup*, a small depression in the cathode containing the filament, is shaped to help focus the electron beam toward a particular spot on the anode. This target, or focal point, of the electron beam is a beveled edge of the anode disk, which is coated with a rhenium-alloyed tungsten. The anode disk itself is made from molybdenum, and in the case of x-ray tubes used in mammography, molybdenum is used in the target area as well. These energetic electrons bombard the target transferring energy by both collisional and radiative transfer. As discussed in Chapter 4, this results in both characteristic and bremsstrahlung x-rays. As can be seen in Figure 5.5, the vast majority of the x-rays produced by an x-ray tube are from bremsstrahlung.

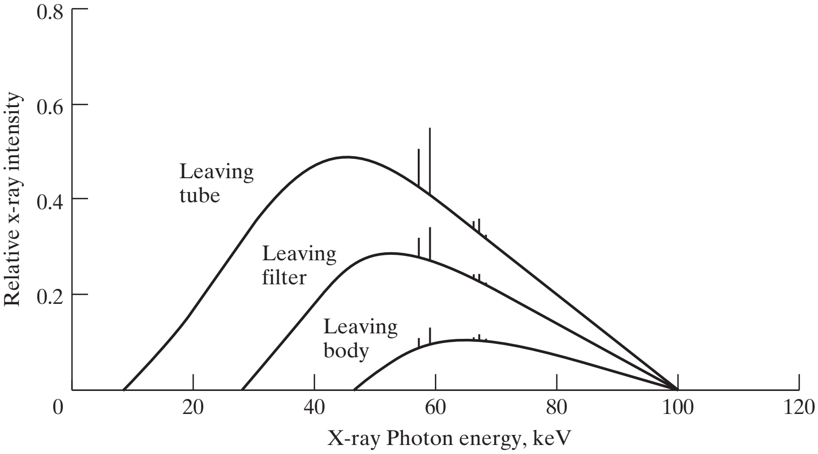


Figure 5.5
X-ray spectra leaving the tube, the filter, and the body.

Besides x-rays, heat is also produced at the anode by the bombardment of electrons. In fact, only about 1 percent of the energy deposited by the electron beam is turned into x-rays; the remaining 99 percent turns into heat. So, at the same time the filament current is applied, the anode is set into rotation by applying alternating current to the stator electromagnets located outside the tube's glass housing. The rotors of this induction motor are connected to the anode inside the glass housing and arranged around bearings so that the anode can rotate freely. Anode rotation is necessary for nearly all diagnostic radiography equipment in order to avoid melting the anode target area because of high-energy transfer from the accelerated electrons to the target. In most x-ray tubes, the anode rotates at 3,200–3,600 revolutions per minute (rpm), but higher rpm tubes are also available for even better heat dissipation.

In order to maintain a constant tube current (mA) while the tube voltage (kVp) is being applied, the filament circuit is very carefully controlled to keep a consistent electron cloud around the filament (i.e., a consistent amount of thermionic emission). The overall exposure is then determined by the duration of the applied kVp, which is controlled by either a fixed timer circuit or an automatic exposure control (AEC) timer. A fixed timer is generally a silicon-controlled rectifier (SCR) switch timed by a microprocessor. Timing accuracy for these circuits is approximately 0.001 seconds. AEC timers have 5-mm-thick parallel-plate ionization chambers placed between the patient table and the film-screen cassette (see Figure 5.2). An ionization chamber is a radiation detector that generates a current when air molecules are ionized by x-rays passing through the chamber and the free electrons and ions are attracted to anode and cathode plates, respectively, which are held at a constant voltage. The voltage achieved across the plates is used to trigger the SCR, which shuts off the tube voltage and terminates the exposure.

The product of the tube current and the exposure time yields the milliampere-second (mAs) value for the exposure. When a fixed timer is used, the radiologist controls both the mA and the exposure time directly and thereby determines the mAs for the exposure. In AEC timers, the mAs is set by the radiologist and the exposure time is determined automatically by the AEC circuitry. A maximum time is set to prevent accidental overdose in the event the AEC circuit malfunctions or the ionization chamber is missing or incorrectly positioned.

5.2.2 Filtration and Restriction

The bremsstrahlung x-rays that are generated within the anode do not all enter the patient, and not all that enter the patient end up leaving the patient. In this section, we discuss modifications to the x-ray beam that take place before the x-rays enter the body. *Filtration* is the process of absorbing low-energy x-ray photons before they enter the patient. *Restriction* is the process of absorbing the x-rays outside a certain field of view. We now discuss these processes in detail.

Filtration The maximum energy of the emitted x-ray photons is determined by the tube voltage. For example, if the tube voltage is 100 kVp, then the maximum photon energy is 100 keV (recall the definition of eV). However, there is a *spectrum* of lower energy x-ray photons that represent bremsstrahlung, as shown in Figure 5.5. The top line depicts the theoretical spectral distribution of bremsstrahlung radiation (normalized to 0.5 at 50 keV). The intensity "spikes" also shown in Figure 5.5 are *characteristic radiations*, and they are due to electrons making transitions between specific electron shells in the anode atoms (as discussed in Chapter 4). Because the x-ray photons emitted from an x-ray tube have a distribution of energies, the x-ray sources used in medical imaging systems are *polyenergetic*.

In practice, it is very undesirable for low-energy photons to enter the body, as these photons are almost entirely absorbed within the body, thus contributing to patient radiation dose but not to the image. There are three important filtering processes in a projection x-ray machine that reduce the number of low-energy x-ray photons that enter the body (see Figure 5.5). First, the tungsten anode itself absorbs a large fraction of low-energy x-ray photons before they even leave the anode. Second, the glass housing of the x-ray tube and the dielectric oil that surrounds it filter out more low-energy photons. This effect might be accentuated over time since aging x-ray tubes tend to accumulate a tungsten film on the inside of the housing due to vaporization of the filament during repeated heating. These first two processes are referred to as *inherent filtering* since they are caused by the x-ray tube itself.

The third x-ray filtering process is called *added filtering*, since it arises from metal placed in the x-ray beam outside of the tube. The most common material used to filter x-rays is aluminum (1–3 mm thick), which is considered the standard x-ray filter material. Other filter materials are often rated in terms of the equivalent aluminum (Al/Eq) that would have to be used to achieve the same attenuation. For higher energy systems, copper might be used because it yields more attenuation for an equivalent actual thickness of aluminum. Note that copper must be followed by aluminum in order to attenuate the 8 keV characteristic x-ray photons created within the copper. Another filtering effect comes from the silvered mirror placed within the collimator (as discussed in the next section). This might provide an additional 1.0 mm Al/Eq of filtering.

Figure 5.5 shows a progressive shift in the position of the spectrum "to the right" (i.e., to higher average energies) as the beam passes through successive materials. This increase in the beam's "effective energy" (see Example 4.4) is called *beam hardening*. Beam hardening is caused by the preferential absorption of lower energy photons, for which the attenuation is higher in most materials (see Figure 4.8). These materials include the filters discussed here, as well as the tissues of the body, which further "harden" the beam.

EXAMPLE 5.1

For radiography systems operating above 70 kVp, the National Council on Radiation Protection and Measurements (NCRP) recommends a minimum total filtration of 2.5 mm Al/Eq at the exit port of the x-ray tube. Although such filtration reduces high-energy as well as low-energy x-rays, thus requiring longer exposure times to properly expose the x-ray detector, the overall dose to the patient is reduced because of the reduction in low-energy x-rays that are absorbed almost entirely by the patient.

Question At 80 kVp, what thickness of copper would provide 2.5 mm Al/Eq of filtration?

Answer The mass attenuation coefficient of aluminum at 80 kVp is $\mu/\rho = 0.02015$ m^2/kg. The density of aluminum is $\rho = 2,699$ kg/m^3. Therefore,

$$\begin{aligned}\mu(\text{Al}) &= 0.02015 \text{ m}^2/\text{kg} \times 2,699 \text{ kg/m}^3 \\ &= 54.38 \text{ m}^{-1}\,.\end{aligned}$$

For copper at 80 kVp: $\mu/\rho = 0.07519$ m^2/kg, $\rho = 8,960$ kg/m^3. So,

$$\begin{aligned}\mu(\text{Cu}) &= 0.07519 \text{ m}^2/\text{kg} \times 8,960 \text{ kg/m}^3 \\ &= 673.7 \text{ m}^{-1}\,.\end{aligned}$$

Since attenuation is determined by the exponential $e^{-\mu\Delta x}$, the x-ray attenuation is equal if the exponents are equal. Hence, the following relation must be satisfied:

$$\mu(\text{Al})x(\text{Al}) = \mu(\text{Cu})x(\text{Cu}).$$

The copper thickness equivalent to 2.5 mm of aluminum at 80 kVp is therefore given by

$$\begin{aligned}x(\text{Cu}) &= \frac{54.38 \text{ m}^{-1} \times 2.5 \text{ mm}}{673.7 \text{ m}^{-1}} \\ &= 0.2 \text{ mm}.\end{aligned}$$

(Remember that, as stated above, copper must be followed by a thin layer of aluminum (typically 1 mm thick) to stop the 8 keV characteristic photons generated by the copper.) ■

Restriction Beam X-ray tubes generate x-rays in all directions. Many photons are absorbed by the anode, others by the x-ray tube housing itself. By design, the x-rays that exit through the tube window form a cone that is ordinarily much larger than the desired body region to be imaged. The exiting beam must be further restricted both to avoid exposing parts of the patient that need not be imaged and to help reduce the deleterious effects of Compton scatter (see below). *Beam restriction* is used for this purpose.

There are three basic kinds of beam restrictors: diaphragms, cones or cylinders, and collimators. *Diaphragms* are flat pieces of lead with holes cut into them, centered on the x-ray beam, and usually placed close to the tube

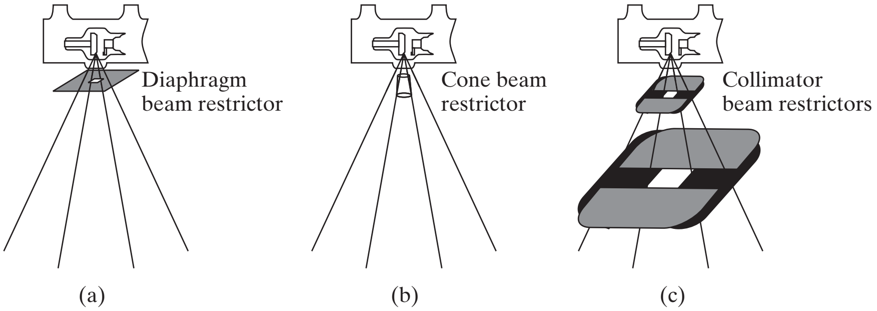


Figure 5.6
Different beam restrictors.

window, as shown in Figure 5.6(a). They are simple and inexpensive but result in a fixed geometry that can usually be used only in dedicated systems that have only one purpose (such as chest radiography). *Cones* or *cylinders*, as shown in Figure 5.6(b), are fixed in their geometry but can have somewhat better performance. *Collimators* are more expensive but are so much more flexible and better performing that they are used nearly all the time in projection x-ray systems. As shown in Figure 5.6(c), collimators have variable diaphragms comprising movable pieces of lead. Most often, there are two collimators, one near the tube and one farther away from the tube. Typically, there is a scored mirror in between these two collimators so that a light coming from the side will shine through the second collimator, illuminating the field of view with an alignment grid.

5.2.3 Compensation Filters and Contrast Agents

Attenuation is the process by which x-rays are absorbed or redirected (scattered) within the body or other objects in the field of view. Body tissues attenuate x-rays in different amounts depending on their linear attenuation coefficients and the x-ray energies. It is this differential attenuation that gives rise to contrast—and therefore the ability to differentiate tissues—in an x-ray image. In certain circumstances, we may want to artificially change the natural attenuation of the body prior to detecting the x-rays. We now discuss compensation filters and contrast agents, both used for this purpose.

Compensation Filters Thick body parts or dense materials (such as metal or bone) stop more x-rays than thinner body parts and normal soft tissues. For example, the torso attenuates more x-rays in the middle than at its edges, simply because it is thicker in the middle. Imaging both locations—that is, a location having large numbers of transmitted photons and a location having low numbers of photons—within the same exposure is difficult because of the limited dynamic range of x-ray detectors. One region would be overexposed while the other is underexposed and neither would yield a usable diagnostic image.

In these cases, a *compensation filter*, comprising a specially shaped aluminum or a leaded plastic object, can be placed between the x-ray source and the patient or in some cases between the patient and the detector. Several types of compensation filters are shown in Figure 5.7. Notice that the compensation

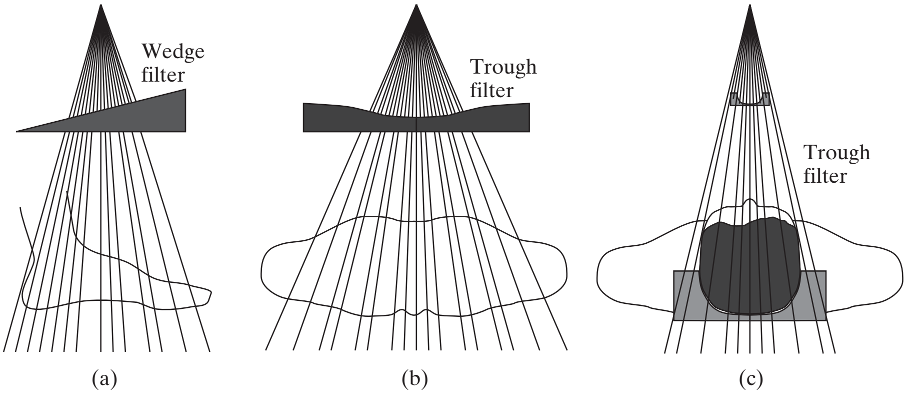


Figure 5.7
Various compensation filters. (Adapted from Carlton and Adler, 2001, and Wolbarst, 1993.)

filter is thicker where the body part is thinner and vice versa so that the x-ray detector requires a smaller dynamic range.

Contrast Agents Differential attenuation in the body gives rise to contrast in the x-ray image. Put another way, in order to see the difference between anatomical structures on a radiograph, these structures must have different x-ray attenuations. Often, however, different soft tissue structures are difficult to visualize because of insufficient intrinsic contrast. This situation can be improved by using *contrast agents*—chemical compounds that are introduced into the body in order to increase x-ray absorption within the anatomical regions into which they are introduced, thereby enhancing x-ray contrast (compared with neighboring regions without such agents).

The two most common contrast agents used in x-ray diagnosis are iodine ($Z = 53$) and barium ($Z = 56$). In addition to their relatively high atomic numbers, iodine and barium also have K-shell electrons whose binding energy falls within the diagnostic x-ray energy range; $E_k = 33.2$ keV for iodine and $E_k = 37.4$ keV for barium. As we have already discussed, when the energy of x-ray photons slightly exceeds the K-shell binding energy of a material, the probability of photoelectric interaction significantly rises. Such a photoelectric interaction will cause electrons from the K-shell to be ejected, and the x-ray photons will be completely absorbed. This effect, called *K-edge absorption*, significantly increases the attenuation coefficient of the material in x-ray energies slightly higher than the K-shell energy. The linear attenuation coefficients of bone, muscle, fat, and two contrast agents, one made using iodine and one using barium, are shown in Figure 5.8. This stepwise increase in attenuation apparent in these plots is called the *K absorption edge*; the location of these edges on the energy scale is directly related to the energies of the characteristic radiations of the contrast agents themselves. There is a dramatically higher absorption of x-rays that are slightly above the K-edge, leading to a very high differential absorption between the contrast agent and the surrounding tissues. These physical properties make iodine and barium radiologically ideal compounds for enhancing contrast.

Because iodine is naturally abundant in the thyroid gland, the thyroid itself is readily imaged in x-ray studies. Fortunately, iodine can also be synthesized into

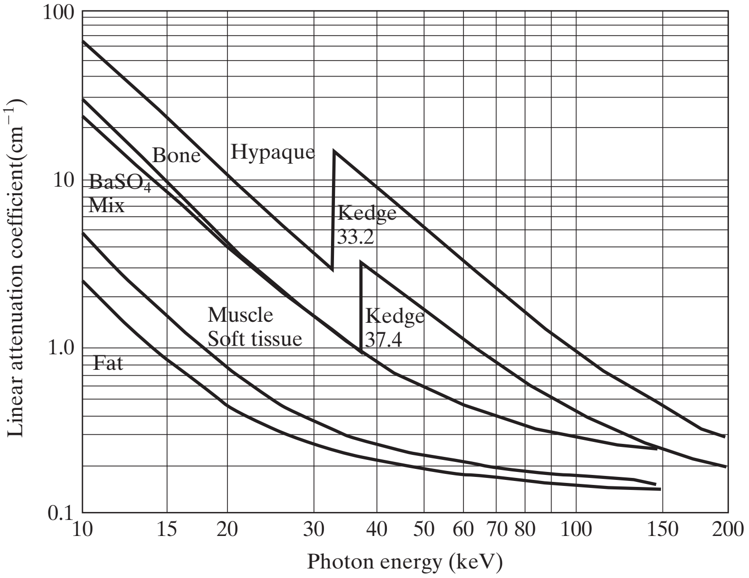


Figure 5.8
Linear attenuation coefficients of bone, muscle, fat, and two contrast agents. (Adapted from Johns and Cunningham, 1983.)

soluble compounds that are safely introduced into the body through intravascular injection or ingestion. Iodinated contrast compounds are used for enhancing the appearance of blood vessels, heart chambers, tumors, infections, and other abnormalities on x-ray images (projection radiography as well as computed tomography (CT) scans). Because these iodine compounds are concentrated by the kidney and excreted into the urine, they are also used for diagnosing problems of the kidneys and bladder using x-ray imaging.

Barium is used exclusively to enhance contrast within the gastrointestinal tract. It is typically administered as a chalky "milkshake," which passes through the gastrointestinal tract without being absorbed or altering the basic function of the stomach and bowel. Barium is the standard agent for enhancing the contrast of the stomach and small and large bowel in conjunction with projection radiography studies.

Air itself can also be used as a contrast agent. Because it does not readily absorb x-rays, it is used as an "opposite" type of contrast to that of iodine and barium. For example, by inflating the lungs, air provides contrast for the lung tissue. Air is also used with barium to provide a double contrast between regions of the bowel containing barium and those containing air.

5.2.4 Grids, Airgaps, and Scanning Slits

X-ray photons that are not absorbed or scattered arrive at the detector from along a line segment originating at the x-ray source (anode of the x-ray tube). However, if an x-ray photon is scattered and still manages to hit the detector, its point of origin will not intersect the x-ray source (unless a highly unlikely multiple scattering sequence occurs). Since scattering is a random phenomenon, this process will cause a random "fog" throughout the image if left uncorrected, thereby reducing the contrast of the direct image (see Section 5.4.3).

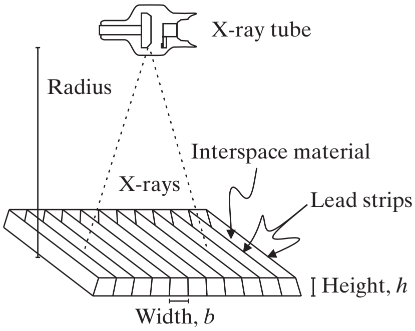


Figure 5.9
A typical x-ray grid.

In nuclear medicine, where inherently monoenergetic radiation is used, energy-sensitive detectors can be used to discriminate between the primary radiation and the scattered radiation, which has lower energy photons. However, this concept cannot be employed in conventional radiography because the radiation that emerges from an x-ray tube is inherently polyenergetic and because an integrated measurement of beam intensity is made. Instead, there are three other methods used in conventional radiography to reduce the effects of scatter: (1) grids, (2) airgaps, and (3) scanning slits. We now describe these in detail.

Grids Scatter-reducing *grids* use thin strips of lead alternating with highly transmissive interspace material, typically aluminum and sometimes plastic. A typical x-ray grid is shown in Figure 5.9. This *linear, focused grid* has lead strips that are arranged in lines and angled toward the x-ray tube. By careful placement of the grid, photons that travel from the source to the detector pass through the interspace material and are not severely attenuated by the grid. Photons that arise from Compton scattering within the patient, however, travel off-axis, intercept the lead strips, and are absorbed in the grid.

Other combinations of grid geometries are sometimes used as well. For example, “crosshatch” grids have lead strips arranged in an actual grid pattern (as in the original design of Gustav Bucky), and “parallel” grids have lead strips that are parallel to each other (so that the focal radius is infinite). Linear grids, however, have the advantage that the x-ray beam can be angulated along the grid line direction, allowing adjustments of the viewing angle. Furthermore, focused grids have the advantage that very few x-rays are actually blocked by the geometry of the grid.

The effectiveness of the grid for reducing scatter is a function of the *grid ratio*, defined as

$$\text{grid ratio} = \frac{h}{b}, \tag{5.1}$$

where h is the height of the lead strips and b is the spacing between the lead strips (see Figure 5.9). This fraction is generally reported as a normalized ratio in which the denominator is unity. Grid ratios ranging from 6:1 to 16:1 are used in conventional radiographic systems and can be as low as 2:1 in mammography systems. Grid spacing is generally reported using its reciprocal, which is known as *grid frequency*, which ranges from 60 cm^{-1} for conventional radiographic systems to as much as 80 cm^{-1} for mammography systems.

Grids that have tall lead strips or fine lead strip spacing—that is, higher grid ratios—are more capable of stopping off-axis radiation. However, there is a price to be paid for using grids. As the lead strips become wider or are packed more closely together in an effort to stop more scattered photons from reaching the detector, they will also stop some photons in the primary beam. This loss of beam intensity can increase image noise. To maintain a high-quality image, it then becomes necessary to increase the amount of radiation arriving at the grid, which increases the radiation dose to the patient. The *grid conversion factor* (GCF) characterizes the amount of additional exposure required for a particular grid. GCF is defined as

$$\mathrm{GCF} = \frac{\text{mAs with the grid}}{\text{mAs without the grid}} \,. \tag{5.2}$$

Typical GCF values range from 3 to 8.

Having GCFs in the range 3–8 implies that some grids can absorb a large fraction of the incident radiation, so the impact of a grid should not be ignored. Since Compton scattering events are more isotropic at lower energies, it may not be necessary to use a grid at lower energies. A rule of thumb in projection radiography is to use a grid when the tube voltage is above 60 kVp. Compton scattering also increases with the thickness of the body part. Another rule of thumb in projection radiography is to use a grid when imaging a body part thicker than 10 cm.

Grids can be mounted as *stationary grids* or used with a *Potter-Bucky diaphragm*, which moves the grid during exposure. The reason that Potter-Bucky grids are desirable is that stationary grids—especially those with low grid frequency—can introduce visible artifacts—lines or grid lines—into the image. The Potter-Bucky diaphragm causes the grid to move 2–3 cm during exposure in a linear or circular path. Since the lead strips within the grid move during exposure, their x-ray shadow also moves across the image plane, and its image is blurred out while the image of the patient is not blurred at all.

Grids are used for nearly all x-ray examinations. The type of grid used varies with the examination. For example, thin body parts (e.g., extremities) generate less scatter than thick body parts (e.g., abdomen or chest), so a lower grid ratio might be employed for extremity exams than for abdominal exams.

Airgaps Physical separation of the detector and the object—that is, leaving an *airgap* between the patient and the detector—is an effective means of scatter rejection. Because the scattered photons diverge radially from their point of origin inside the object, while the primary beam photons diverge from the source, separating the source and detector reduces the percentage of scattered photons that reach the plane of the detector. This can be demonstrated by simple geometry. The price that is paid for this airgap method of scatter reduction is increased geometric magnification and increased blurring or unsharpness due to x-ray focal spot size effects (see Section 5.4).

Scanning Slits The use of mechanical lead slits that are placed in front of and/or in back of the patient has been evaluated and are used in some systems. These slits are moved together during the x-ray exposure, effectively providing a linear "scan" of the patient. Because the x-ray beam is collimated to a thin line, the amount of scatter generated is small, and because the detector is also collimated to a thin line, the amount of scatter accepted is small. Hence, these systems can provide greater than 95% scatter reduction, at the expense of a more complex and costly system and longer exposure times.

5.2.5 Film-Screen Detectors

In 1895, Roentgen discovered x-rays and made the first radiograph by allowing the x-rays to directly expose a photographic plate. X-rays can directly expose today's modern photographic film, but this is a very inefficient way to create a radiograph. In fact, only about 1–2 percent of the x-rays are stopped by the film, so creating radiographs by direct film exposure requires an unnecessarily large x-ray dose to the patient.

To greatly improve their efficiency, film-based diagnostic x-ray units always have *intensifying screens* on both sides of the radiographic film. The intensifying screens stop most of the x-rays and convert them to light, which then exposes the film. This is a very efficient process, and the screens cause only a small amount of additional image blurring.

In this section, we first describe the process of luminescence, and the composition and dimensions of modern intensifying screens. We then describe the optical properties of radiographic film. Finally, we conclude with a discussion of the radiographic cassette, which holds the intensifying screens and the film.

Intensifying Screens A cross section of a typical radiographic intensifying screen is shown in Figure 5.10. Since screens are used on both the front and back of the radiographic film, all parts of the screen except the phosphor must be uniformly radiolucent. The phosphor is the active part of an intensifying screen; its purpose is to transform x-ray photons into light photons. The light photons then travel into the film, causing it to be exposed and to form a latent image. The *latent image* is the "virtual" image resident in the film, but it is not yet viewable by an observer.

The base of the screen is provided for mechanical stability, but it must be somewhat flexible so that it can be pushed tightly against the film. It is typically made of polyester plastic. The reflective layer is provided so that light from the phosphor is reflected back into the film rather than getting lost in the base. It is typically about 25 μm thick and made of magnesium oxide or titanium dioxide. A plastic protective coating is applied to the film side of the screen to protect it from repeated film loading and unloading.

Phosphors are examples of materials that are *luminescent*; that is, they convert one form of energy, in this case x-rays, into light. Two types of luminescence are distinguished: *fluorescence*, in which the emission of light takes

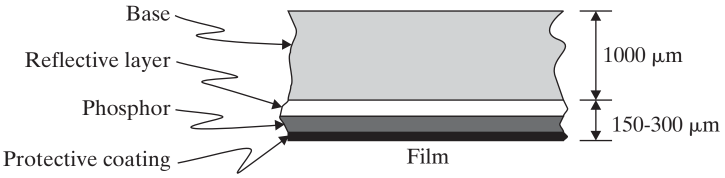


Figure 5.10
Geometry of an intensifying screen.

place entirely within 1×10^{-8} second of the excitation, and *phosphorescence*, in which light emission can be delayed and extended over a longer period of time. For screens, it is desirable to use a luminescent material that is much more fluorescent than phosphorescent. This way, there is little chance of an afterglow that might spoil the exposure by either motion after exposure or by light from a previous exposure.

Among the many luminescent materials, good intensifying screen phosphors should also be highly x-ray attenuating and should emit many light photons for every x-ray photon that is stopped. Thus, the best phosphors have high atomic numbers (so the linear attenuation coefficient is large) and high conversion efficiencies. The *conversion efficiency* is a measure of the number of light photons emitted per incident x-ray photon. Typical conversion efficiencies are between 5 and 20 percent depending on the type of phosphor used and its thickness.

A typical conversion efficiency for modern screens yields 1×10^3 light photons per incident 50 keV x-ray photon. This number depends on the energy of the incident x-ray photon as well since the conversion of a higher energy x-ray photon will produce more light. Total light output depends on the combination of attenuation of the material at a given x-ray energy and the conversion efficiency at that energy. The *speed* of a screen is really just a measure of its conversion efficiency. If the conversion efficiency is higher, then the screen is "faster," because the larger numbers of light photons emitted by the phosphor will expose the film faster.

Thomas Edison explored many phosphors in the early 1900s for their possible use in radiography. He discovered that calcium tungstate ($CaWO_4$) is an excellent fluorescent phosphor for radiological applications. Until relatively recently, nearly all intensifying screens were made from calcium tungstate. In the late 1970s, rare earth phosphors were introduced and are more commonly used today because they have higher conversion efficiencies.

Radiographic Film After Roentgen discovered x-rays by observation of a mysteriously glowing phosphor, he quickly learned that x-rays could also expose a photographic emulsion. Although digital detectors (see Section 5.2.7) are undergoing rapid development and widespread adoption, film is still very commonly used in radiography both as an x-ray receptor and as an archival storage mechanism for radiographic images (even for those acquired digitally). Because of the nearly exclusive use of scintillation screens, radiographic film is really just optical film, designed to capture the optical image created within the screens that sandwich the film. Common sizes in the United States include 14 × 17, 14 × 14, 11 × 14, 10 × 12, 8 × 10, and 7 × 17 inches.

Many different film speeds are available, depending on the application. The detailed optical properties of film greatly influence the resultant quality of the developed image. Also, the precise details of chemical development play a large role in the final appearance of the film as well. We will only consider these issues from the highest level, in enough detail so that we can understand the broad issues relating x-ray exposure to film appearance (see Section 5.3.4).

Radiographic Cassette A radiographic cassette is really just a holder for two intensifying screens and the film "sandwiched" between. One side of the cassette is radiolucent, while the other usually includes a sheet of lead foil. Therefore, the cassette can be loaded only one way into the x-ray machine. At least one side of the interior of the cassette contains a spongy, foam material that applies a uniform pressure against the screen so that a uniform contact is created between the film and all parts of each of the screens. Optical mirrors are located outside the screens, pointed inward so that nearly all the light produced in the screens ends up exposing the film. Some systems, particularly those with high resolution such as mammography systems, contain only one screen, but the principle of holding the film tightly against the single screen within the cassette still applies.

Film must be loaded in a darkroom in order to avoid premature exposure. The cassette is opened like a suitcase or briefcase only as much as is required to slip in the radiographic film. This minimizes the amount of dust and other contaminants that might end up inside the cassette. When the cassette is closed, pressure is applied so that the foam inside compresses, creating a uniform contact between the screens and the film. Cassettes and screens must be cleaned regularly in order to avoid contaminants that might appear in the images.

5.2.6 X-Ray Image Intensifiers

X-ray image intensifiers (XRIIs) are used in fluoroscopy, where low-dose, real-time projection radiography is required, whereas a conventional x-ray system might use a filament current of 50–1,200 mA, a fluoroscopy unit would typically use only 0.5–5.0 mA so that the x-ray tube can operate continuously. The light levels produced by these low-intensity x-rays striking a conventional phosphor are too low for human vision in room lighting, so the XRII was developed to amplify the light. A diagram depicting an XRII is given in Figure 5.11.

X-rays pass through an *input window* made of aluminum or titanium, with a thickness ranging between 0.25 and 0.5 mm. This type of material creates minimal loss of x-ray photons, yet it is capable of holding a vacuum. The x-rays then strike a 0.5-mm-thick *input phosphor* [typically CsI(Na)], which is circular with a diameter ranging between 15 and 40 cm. The x-ray photons that are absorbed in the input phosphor generate flashes of light that are channeled toward the photocathode. Photons that travel backward toward the input window are reflected by a 0.5-mm-thick aluminum sheet that supports the input phosphor.

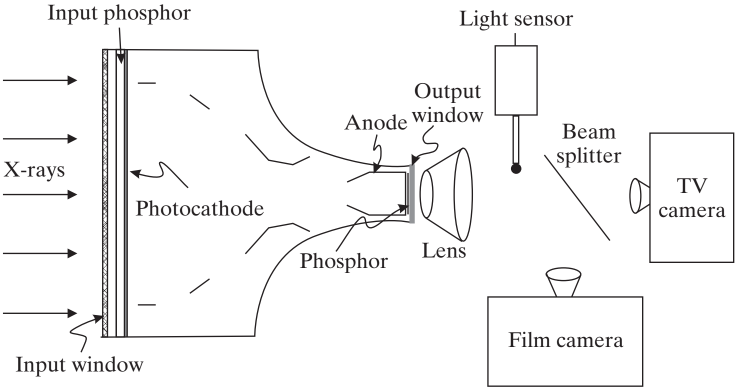


Figure 5.11
Block diagram of an x-ray image intensifier (XRII).

The remainder of the XRII resembles a standard image intensifier, such as that found in night vision devices. The light photons generated in the input phosphor strike a photocathode, which in turn generates free electrons within the vacuum tube. The electrons are accelerated through a series of electrodes (called *dynodes*) toward the anode, which is kept at a voltage of 25–35 kV relative to the cathode. The free electrons do not strike the dynodes as in photomultiplier tubes (see Chapter 8). Instead, the electrons are shaped by the dynodes into an (inverted) electron intensity image that represents the intensity of x-rays entering the XRII. The voltage profile of the dynodes can be altered—continuously in some systems and in discrete steps in other systems—to provide variable image magnification. For example, a 40-cm XRII could be used to image a 20 cm field of view.

The electrons accelerate through the XRII until they are absorbed by either the anode or the *output phosphor screen*. The output phosphor is a P20-type phosphor deposited on the *output window*, typically a 15-mm piece of glass 25–35 mm in diameter. A thin aluminum film placed on the inner side of the phosphor acts as the anode and a reflector, keeping light from passing back into the XRII, which might otherwise cause secondary excitation of the photocathode. The light passing through the output window typically encounters a *lens*, which magnifies the image.

The magnified light image leaving the XRII can be used in several ways, three of which are depicted in Figure 5.11. First, a light sensor can be put into the field to provide an automatic gain control—that is, feedback to the x-ray tube current—to provide a relatively constant image brightness throughout an exam. Second, the main function of a fluoroscope is achieved by feeding the light image into a TV camera, which provides real-time viewing capability by sending the camera's signal into a standard TV monitor. Finally, using a splitter, the image can also be fed into a film camera, which can capture selected still images onto film.

5.2.7 Digital Radiography

Digital radiographic systems are rapidly replacing film-screen combinations. Although the first version of such systems was introduced in the 1980s, it took over twenty years for the cost to be driven down and for the quality to improve sufficiently for these systems (and subsequent versions of these systems) to compete successfully with film-based systems. There are four types of digital systems, which we describe next, all of which are in clinical use today.

Computed Radiography Systems *Computed radiography* (CR) systems store latent images in *photostimulable imaging plates* (PSPs), which are removed from the x-ray system and scanned using a special laser scanner to form a digital image. The PSP is a phosphor typically made from barium fluorohalide bromides and iodides with europium activators. Detected x-ray photons are absorbed in the phospor by the photoelectric effect, which causes electrons to be ejected from atoms in the detector. About half of these electrons will recombine with the resulting ions causing light to be emitted (as in a typical image intensifying screen). The remaining electrons will be temporarily trapped in the fluorohalides, leaving “holes” in the europium; it is these holes that represent the latent image that is stored in the PSP. Latent images last for hours, but as electrons find ways to recombine with the holes, they will be unrecoverable after a few days of storage.

PSPs must be processed in order to recover the latent image. From the viewpoint of work flow in a radiology clinic, this aspect of CR bears great similarity to that of film-based radiography, which is one reason (once price dropped and quality improved) why CR became a natural way to transition to digital radiography from film-based systems. To make it even easier to transition, the CR cassette, as shown in Figure 5.12, can often be used directly in place of the film-screen cassette in existing systems. But unlike film, there are no chemicals involved in recovering the radiographic image in CR. Instead, reading a PSP involves use of a focused laser, which stimulates the trapped electrons so that they will recombine with the holes and produce light. To do this, a mechanical device, much like the paper roller in a laser printer, is used to slide the PSP to a position where a focused laser is scanned across a row of the PSP; the PSP is then slid to a new position and a new row is scanned, and so on. At each position of the laser, which is approximately 700 nm in wavelength and therefore red in color, blue light is emitted from the PSP in proportion to the number of trapped electrons that are liberated by the laser excitation. This emitted light is directed by mirrors toward a photodetector that is sensitive only to blue light, and the resultant electrical signal is digitized using an analog to digital converter (ADC). This produces a digital value—a pixel value—that is proportional to the x-ray intensity at the spot on the PSP being probed by the laser.

Since the PSP is by design “raster scanned” row-by-row, the resulting data comprise rectilinear samples of a digital image, which can be directly manipulated and stored on a computer and displayed on a digital display. The density of spatial samples during scanning affects the resolution of the system, and the number of bits per pixel in the ADC affects the overall dynamic range of the system. For example, high-resolution systems might sample at 10 pixels/mm and each pixel value might be represented by a 16-bit digital word. The dynamic

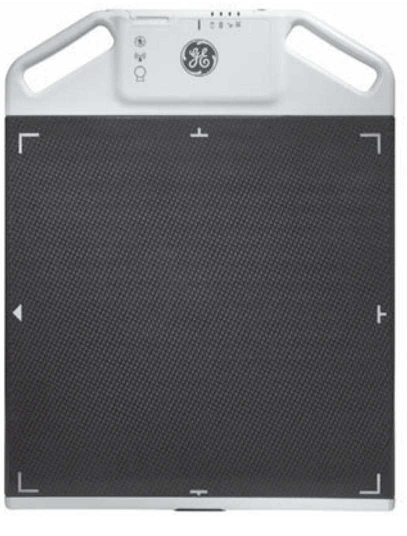


Figure 5.12
A computed radiography (CR) cassette. Image courtesy of GE Healthcare.

ranges of digital image created by scanning a PSP can be dramatically different depending on the anatomy being imaged as well as the kVp and mAs selected by the operator.

Unlike film, which saturates at high x-ray exposures, PSPs can store perfectly viable latent images over a wide range of incident x-ray exposures. This means that fewer images are "spoiled" by having poor latent images through improper x-ray tube parameter selection, but the downside is that very different dynamic ranges must be anticipated. Most systems use an air ionization photocell to perform AEC, as previously described for film-based radiography. The technician must specify an appropriate exposure based on the patient size and anatomy being imaged. Many systems also permit some preprocessing of the scanned images prior to finalizing them for the patient record, which permits additional specification of the image values in accordance with expected norms. Another somewhat unfortunate downside of the fact that few images are spoiled in CR acquisition is that, since lower exposures cause the usual quantum noise and higher exposures simply improve signal-to-noise ratio (SNR), there is a natural tendency to use higher x-ray exposures in CR. It is important to gaurd against this tendency since higher exposures yield higher doses to the patients.

It turns out that the process of reading a PSP does not liberate all the electrons that are stored in the plate; in fact, a PSP can be scanned more than once without much degradation. In order to reuse the plate—that is, to create a new x-ray image—it must first be erased, which is accomplished very simply by exposing it to very bright light. Usually this is done in the same mechanical unit that reads the plate.

CCD-Based Digital Radiography Systems Charge coupled device (CCD) detectors combine a collection of light-sensitive capacitors, which store charge in proportion to incident light, with a CCD, which uses a linear charge transfer to permit a high-fidelity readout of the stored charges. Two-dimensional CCD detector arrays capture 2-D light images and are used in high-quality photographic equipment today. Their use in x-ray imaging is made possible simply by using a scintillator, typically Tl-doped CsI, to convert the x-rays to visible light.

The main obstacle to the use of CCD detectors in projection radiography is that they are small devices, while the field of view of most x-ray examinations is large. The most straightforward approach to solve this problem is the use of a lens to *demagnify* the scintillator image so that it fits within the optical window of the CCD array. With a typical field-of-view and typical CCD detector size, demagnification by a factor of roughly 10:1 is necessary. For example, to match a 44 cm × 44 cm field of view (FOV) to a 4 cm × 4 cm CCD with 3,000 × 3,000 pixels, each 0.013 mm × 0.013 mm, a demagnification factor of just over 11 is required. However, it turns out that there is a significant loss of light—as much as 99%—with such demagnification, and this yields significantly higher quantum noise in the resultant images. Use of higher x-ray doses can compensate for this loss, and because of the very low cost of CCD detectors, this is one approach that has been used to produce very low-cost digital x-ray systems.

Another way to use CCD detectors is in a so-called *slot-scan* geometry. In this geometry, multiple CCD detector arrays are positioned in a line, long enough to cover the entire field of view (or a significant fraction of the FOV) in one dimension (termed horizontal)—for example, the expected maximum width of a human chest. If demagnification is required, it is relatively small (on the order of 2:1 or 3:1), and it is carried out using fiber optic "light pipes" so that only a small proportion of light photons are lost. For example, a typical complete-coverage geometry would use eight detector arrays, each 5.5 cm by 1.1 cm, tiled together to yield a single linear detector of dimension 44 cm by 1.1 cm. The CCD detector might have 342 elements in the horizontal direction and 68 elements in the vertical direction. The x-rays are then collimated into a fan beam and directed at the linear CCD detector, which can collect data simultaneously for a small collection of horizontal rows—for example, 68 rows—of the digital image in parallel. To collect additional rows, either the entire apparatus—x-ray source, collimator, and detector—are translated (vertically) together or just the source collimator and detector are translated (vertically) together. Translation can be carried out, depending on the field of view, over a time period between 8 and 15 seconds. Patients having a chest x-ray using this type of system would be required to hold their breath for the duration of the scan in order to minimize motion artifacts due to breathing.

Thin-Film-Transistor-Based Digital Radiography Systems CCD detectors have pixel sizes that are inherently small and cannot be scaled up due to noise that is added in the charge transfer mechanism with larger detector areas. Therefore, although constructing large CCD detectors—for example, large enough to be full-field detectors such as that achieved by film and PSPs—is technically possible, cost, sensitivity to x-ray damage, complexity, low-speed, and calibration issues currently make this impractical. However, the advantage of having a direct digital read-out afforded in CCD cameras is clear, and thin-film-transistor-based digital radiography systems provide that capability.

Thin film transistors (TFTs) are made by depositing thin films of a semiconductor material, a dielectric, and metal conductors onto glass. In flat panel displays, which are familiar to us because they are used in modern LCD and LED television systems, the transistors are laid out in a 2-D grid so that each pixel can be activated by addressing a horizontal and vertical wire simultaneously—this is the so-called *active matrix* technology. In x-ray flat panel detectors, the TFT active matrix design is used to read out a charge that has accumulated at each pixel due to x-ray flux.

To completely specify the TFT-based flat panel detector, we must specify how the x-rays will be detected and charge will be accumulated. There are two methods, so-called *indirect* and *direct*, both of which lead to charge stored on a capacitor associated with each image pixel. Indirect detection is very familiar to us: A scintillator, typically CsI:Tl, is placed between the x-ray source and the TFT matrix, which causes detected x-rays to generate light photons. Amorphous silicon (a-Si) is then used as a photodiode to convert the light emitted from the scintillator to charge on the storage capacitors. Direct detection, on the other hand, uses amorphous selenium (a-Se) to directly convert x-rays to electrons, which are then moved directly to the storage capacitors using an applied voltage across the selenium plate. Currently, flat panel detectors based on the indirect detection method are more common, in part because these detectors are faster (i.e., high light output and shorter afterglow, which is important in temporal acquisitions such as fluoroscopy) and they have lower noise. However, direct detection has the potential for higher spatial resolution and are being rapidly improved in materials that will offer lower noise designs.

CMOS-Based Digital Radiography Systems Complementary metal oxide semiconductor (CMOS) active-pixel sensor (APS) imagers are the most common optical sensor for consumer cameras. Their use in x-ray detection is relatively new and the advantages and disadvantages are still being discovered even as the technology continues to improve. Current technology allows for pixels in CMOS sensors to be made larger than CCD elements, they have low cost and low power consumption, they are tolerant to x-ray irradiation, and they have fast pixel readout capability.

Each pixel in a CMOS APS comprises a photodiode and three or more MOSFET transistors. Because light is used as an input, these detectors are indirect and must be coupled to a scintillator as in a-Si TFT detectors. Like TFTs, the pixels within a CMOS APS are accessed by activating an access enable conductor, which identifies the row and then reads a particular output conductor, which identifies the column. As opposed to the CMOS *passive-pixel sensor* (an earlier design that has only has one transister per pixel), the CMOS APS uses extra transistors to amplify the signal (accumulated charge) prior to readout, yielding much lower noise and faster readouts.

The pixel size in CMOS sensors can be much larger than that of CCD detectors, which is an advantage. A typical pixel size is 0.075 mm × 0.075 mm, and they can be grouped together to make larger "superpixels"—for example, up to 4 × 4 to make pixels of size 0.3 mm × 0.3 mm—which enables faster readout rates for real-time applications. Like CCD imaging arrays, CMOS imaging arrays must be manufactured on a single semiconductor substrate and this is a limiting factor in comparison to TFT-based detectors. Second generation sizes support up to 5 cm × 10 cm substrates and these can be tiled (with greater ease than tiling CCD arrays) to make up larger panels, currently up to 290 mm × 230 mm, which is suitable for mammography applications.

Wireless and Rechargable DR Detectors In addition to its capability for high resolution, convenience of workflow is a key reason that CR got an early headstart in the digital x-ray area. In particular, existing x-ray systems could be retrofitted to support the CR cartridge instead of film screen cartridges, and only a separate processing unit needed to be purchased to replace the chemical film developers. Early DR systems were sold to replace existing x-ray systems—not just their detectors—and therefore represented a considerable capital investment.

Newer DR detectors are completely portable plates or cartridges that have their own rechargable battery power sources and wireless ethernet connectivity. They are completely encased in plastic so that they can be easily cleaned and will survive minor bumps in the course of normal use. These detectors can be shared between multiple rooms or x-ray systems and can be retrofit into existing table and wall-mounted systems. Data from these detectors can be transmitted directly to a viewing station or departmental picture archiving and communication system. Because of these conveniences, the improving resolution and SNR, and the lack of a requirement for separate reading and erasing hardware, DR systems are beginning to overtake CR in modern radiology departments.

5.2.8 Mammography

Mammography is a sufficiently specialized and widely used modality within projection radiography that we consider it separately here. Mammography is used for the early detection of breast cancer either by direct detection of tumors in the images or by detection of microcalcifications (tiny deposits of calcium) that sometimes indicate the presence of breast cancer. A conventional mammography unit is based on a traditional projection radiography system, whose size is optimized for the breast. Low-energy x-rays (around 30 kVp) are typically used. A significant feature of such systems is the integration of *compression* via a *paddle* to produce a reduced and more uniform (and somewhat more standardized across patients) thickness organ for imaging. In this approach, the breast is compressed between the paddle and the detector; compression thus also reduces motion during imaging. Another important feature of most conventional mammography systems is an automated intensity control system that optimizes detector exposure.

With the advent of *full field digital mammography*, more than half of all mammography systems are now direct digital radiography systems. Early digital systems had 70 micron pixels, leading to spatial resolution around 3.5 line pairs/mm (compared with 12 or so line pairs/mm to as much as 20 line pairs/mm for film-screen-based systems); current systems have resolution around 6–8 line pairs/mm. However, digital systems have much greater dynamic range than film, avoiding problems of under- or over-exposure. The digital display of images provides control over brightness and contrast. Furthermore, digital systems have advanced processing algorithms for contrast enhancement and small-structure enhancement, as well as computed-aided detection and diagnosis (CAD) algorithms that help the radiologist to interpret the images.

The most advanced mammography systems use tomosynthesis or stereo digital mammography. In the latter, two mammograms shifted by 5–10 degrees are captured and digitally fused and displayed. In the former, multiple (typically, 7–12) angular views are captured over a 50-degree arc, and then digitally reconstructed to yield a volumetric dataset, from which individual tomographic slices can be extracted and displayed. Most tomosynthesis systems are based on the concept of linear tomography and are only implemented in state-of-the-art full-field digital mammography systems. Now that the technology is becoming more widely accepted, more mammography systems can be equipped for tomosynthesis, and some have iterative reconstruction methods that permit dose reduction and improved visualization of microcalcifications.

5.3 Image Formation

5.3.1 Basic Imaging Equation

The x-ray tube emits a burst of x-rays that, after filtration and restriction, are incident upon the patient. These x-rays are then attenuated as they pass through the body in a spatial pattern that depends on the linear attenuation coefficient distribution in the body. Consider a particular line segment through the object starting at the x-ray origin and ending on the detector plane at point (x, y), as shown in Figure 5.13. The linear attenuation is a function of x, y, and z, in general; on the line, it can be considered to be a function of its distance s from

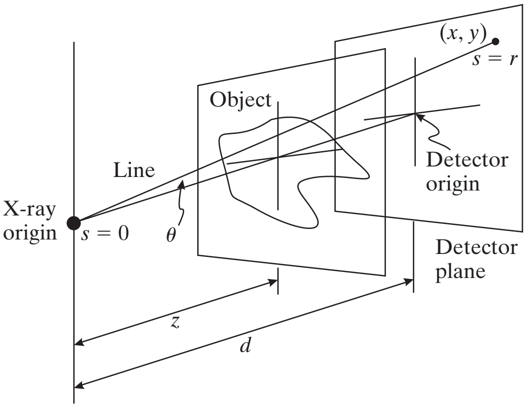


Figure 5.13
The geometry of a conventional projection radiographic system.

the origin. Suppose that the length of the line segment is $r = r(x, y)$, where it is made explicit that the length of the line depends on the position (x, y). Then, referring to (4.34), the intensity of x-rays incident on the detector at (x, y) is given by

$$I(x, y) = \int_0^{E_{\max}} S_0(E')E' \exp\left\{-\int_0^{r(x,y)} \mu(s; E', x, y)ds\right\} dE' , \qquad (5.3)$$

where $S_0(E)$ is the spectrum of the incident x-rays. Given a different line, the linear attenuation function is usually different since it passes through different materials; that is why μ is made explicitly dependent on x and y here. Therefore, if we were to actually evaluate this double integral, we would have to first determine the function $\mu(s; E, x, y)$ for the particular line of interest and then proceed with the integration. Any photon "surviving" its passage through the body then hits the detector apparatus (radiographic cassette or digital detector), and is either absorbed by the detector itself, is absorbed by the detector packaging including its lead backing, is scattered in an abitrary direction, or exits out the back of the detector.

5.3.2 Geometric Effects

X-ray images are created from a diverging beam of x-rays (see again Figure 5.13); this divergence produces a number of undesirable effects that arise from geometric considerations. These effects are multiplicative; we now study them in sequence.

Inverse Square Law The *inverse square law* states that the net flux of photons (i.e., photons per unit area) decreases as $1/r^2$, where r is the distance from the x-ray origin. Assume that the beam intensity integrated over a small sphere surrounding the source is given by I_S. Let the source-to-detector distance be d, as shown in Figure 5.13. Let the detector origin ($x = 0$ and $y = 0$) be defined by the point on the detector whose normal direction passes through the x-ray source. Then, assuming there is no object causing beam attenuation between the source and detector, the intensity at the origin of the detector

$$I_0 = \frac{I_S}{4\pi d^2} . \qquad (5.4)$$

The intensity at an arbitrary point (x, y) on the detector is smaller than that at the detector origin simply because it is farther away from the x-ray source. Let $r = r(x, y)$ be the distance between the x-ray origin and the detector point (x, y). Then the intensity at (x, y) is

$$I_r = \frac{I_S}{4\pi r^2}\,, \tag{5.5}$$

where, again, we have assumed that there is no object attenuation. These relations, and the fact that $\cos\theta = d/r$ (see Figure 5.13), yield

$$I_r = I_0 \frac{d^2}{r^2} = I_0 \cos^2\theta\,. \tag{5.6}$$

Thus, the inverse square law causes a $\cos^2\theta$ drop-off of x-ray intensity away from the detector origin, even without object attenuation. Without compensation, this effect could be falsely interpreted as object attenuation in a circular pattern around the detector origin.

EXAMPLE 5.2

The inverse square law has a very practical use in radiography. Suppose an acceptable chest radiograph was taken using 30 mAs at 80 kVp from 1 m. Suppose that it was now requested that one be taken at 1.5 m at 80 kVp.

Question What mAs setting should be used to yield the same exposure?

Answer The intensity at the detector should remain constant in order to have the same exposure. We determine that

$$I_{\text{old}} = \frac{I_S(\text{old})}{4\pi d_{\text{old}}^2} = I_{\text{new}} = \frac{I_S(\text{new})}{4\pi d_{\text{new}}^2}$$

or

$$I_S(\text{new}) = I_S(\text{old})\frac{d_{\text{new}}^2}{d_{\text{old}}^2},$$

where I_S refers to the intensity at the x-ray source. X-ray intensity at the source is directly proportional to mAs, the product of the tube current and the exposure time. Therefore,

$$\begin{aligned} \text{mAs(new)} &= \text{mAs(old)}\frac{d_{\text{new}}^2}{d_{\text{old}}^2} \qquad (5.7)\\ &= 30\ \text{mAs} \times \frac{(1.5\ \text{m})^2}{1\ \text{m}^2}\\ &= 67.5\ \text{mAs}\,. \end{aligned}$$

Equation (5.7) is called the *density maintenance formula*; it is of significant practical utility in maintaining equivalent x-ray exposures while varying patient distance. ■

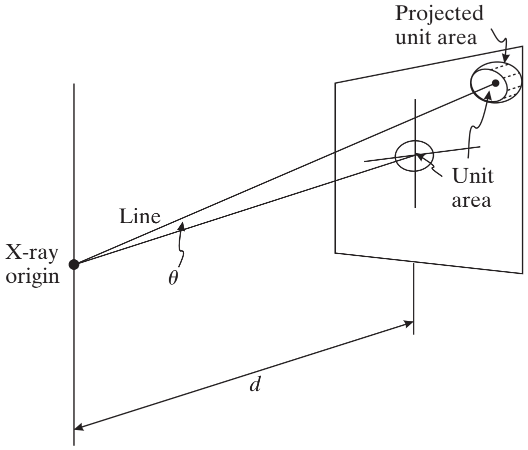


Figure 5.14 The effect of obliquity on spot size.

Obliquity *Obliquity* is a second factor that acts to decrease the beam intensity away from the detector origin. The obliquity effect is caused by the detector surface not being orthogonal to the direction of x-ray propagation (except at the detector origin). As shown in Figure 5.14, this fact implies that x-rays passing through a unit area orthogonal to the direction of x-ray propagation actually pass through a larger area on the detector. Thus, the x-ray flux is lower, which directly results in a lower measured x-ray intensity on the detector surface.

Given an area A orthogonal to the direction of x-ray propagation, the projected area on the detector is $A_d = A/\cos\theta$. Therefore, the measured intensity due to obliquity alone is

$$I_d = I_0 \cos\theta \,, \tag{5.8}$$

where we have again assumed that there is no object attenuation.

Beam Divergence and Flat Detector The effects of beam divergence and the flat detector act together to reduce intensity at the detector plane in two ways: (1) reduction in beam intensity due to the inverse square law effect [(5.6)] and (2) reduction in beam intensity due to obliquity [(5.8)]. The combination of these two effects is multiplicative. So, the overall beam intensity relative to the intensity I_0 at the detector origin, assuming no object attenuation, is given by

$$I_d(x, y) = I_0 \cos^3\theta \,. \tag{5.9}$$

If θ is small, then $\cos^3\theta \approx 1$, and both effects can be ignored. This is a good approximation for most projection radiography systems, where either the tube is very far away from the detector or the detector is small enough that θ is small even at the edges of the detector.

EXAMPLE 5.3

Suppose a chest x-ray is taken at 2 yards using 14 inch by 17 inch film.

Question What will be the smallest ratio I_d/I_0 across the film (assuming no object attenuation)?

Answer The corner of the film is distance r_d away from the detector origin, where

$$r_d = \sqrt{7^2 + 8.5^2} = 11.0 \text{ in.}$$

The cosine of the largest angle θ made by the beam is therefore

$$\begin{aligned}\cos\theta &= \frac{d}{\sqrt{d^2 + r_d^2}} \\ &= \frac{72}{\sqrt{72^2 + 11^2}} \\ &= 0.989\,.\end{aligned}$$

The smallest intensity ratio is therefore

$$\frac{I_d}{I_0} = \cos^3\theta = 0.966\,,$$

which represents about a 3 percent variation. Given the other effects we will be exploring—anode heel effect, path length, depth-dependent magnification, noise, and scattering—this effect is a negligible intensity perturbation across the film. ■

Anode Heel Effect We have heretofore modeled the beam coming from an x-ray tube as having uniform intensity within a cone. The *anode heel effect*, however, predicts a stronger intensity beam in the cathode direction and no variation orthogonal to the cathode–anode direction. The reason for this effect is due to the geometry of the anode, as shown in Figure 5.4. Although the generation of x-rays within the anode is essentially isotropic at the atomic level, those x-rays traveling out of the anode in the forward (cathode-to-anode) direction have more anode material through which to propagate before leaving the anode itself than those leaving the anode in the backward direction. Thus, the beam intensity is higher in the cathode direction; the variation can be as much as 45 percent intensity variation in the cathode to anode direction. Thus, the anode heel effect far outweighs the effects of obliquity and inverse square law in its overall effect on the uniformity of intensity across the detector surface.

The anode heel effect can be compensated by using an x-ray filter that is thicker in the cathode direction than in the anode direction. If it is not compensated by filtration, then it must be recognized when positioning patients relative to the x-ray tube. In particular, the anode heel effect should be used as a kind of inherent compensation for natural gradients in the body's x-ray attenuation profile. For example, when imaging the foot, the toes should be positioned toward the anode since that part of the foot is thinner; the cathode end will have more x-rays to penetrate the thicker (proximal) part of the foot. This principle can be applied throughout the body. In our development of x-ray imaging equations, we will assume that the anode heel effect is compensated by filtration.

Path Length Consider imaging a slab of material with a constant linear attenuation coefficient μ and thickness L arranged parallel to the plane of the

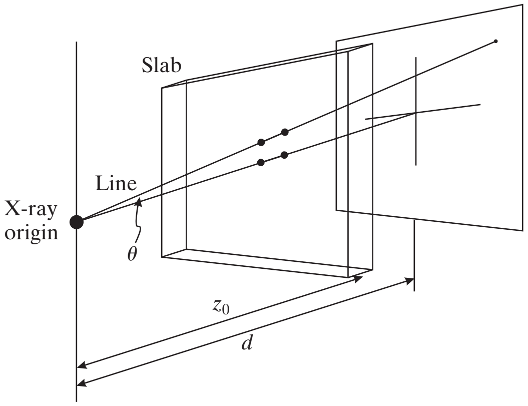


Figure 5.15
Imaging a uniform slab.

detector, as shown in Figure 5.15. Assume a monoenergetic x-ray spectrum for the present analysis. The central ray of the x-ray beam—that is, the ray that goes through the center of the object and hits the detector origin—encounters a net path length L through the object. Thus, the x-ray intensity at $(x, y) = (0, 0)$ will be $I_0 \exp(-\mu L)$, where I_0 is the intensity of the beam that would be present at the detector origin if the slab were not present.

The x-rays propagating toward the detector point (x, y), however, experience a different path length through the slab. The path length through the slab along the line between the x-ray source and (x, y) is given by

$$L' = \frac{L}{\cos\theta} \,. \tag{5.10}$$

Since $L' > L$, more x-rays will be attenuated along this line than along the central path, and the x-ray intensity at the detector will be smaller. Ignoring the inverse square law, obliquity, and the anode heel effect, the intensity is given by

$$I_d(x, y) = I_0 e^{-\mu L/\cos\theta} \,. \tag{5.11}$$

If the inverse square law and obliquity are included, then the intensity is given by

$$I_d(x, y) = I_0 \cos^3\theta \, e^{-\mu L/\cos\theta} \,. \tag{5.12}$$

If the anode heel effect is compensated by filtration, then (5.12) represents a valid imaging equation for a homogeneous slab parallel to the image plane within the field of view.

The interpretation of radiographs, both visually and through automatic image analysis techniques, should be informed by (5.12) and by the anode heel effect (if not compensated). Since materials with larger attenuation are brighter when viewed on a light box (or a computer screen if digitally displayed), these effects will cause a *shading artifact* within a homogeneous object. This could be interpreted as representing a different attenuation within the object or different object thickness. It is truly an ambiguous situation. Radiologists are trained to study radiographic intensities locally, and to not compare absolute image values across the field of view. This is not a desirable situation for automatic (computer-based) image analysis methods, in general, which requires such methods to be more complicated than would otherwise be the case.

Depth-Dependent Magnification Another consequence of divergent x-rays is object magnification, which is specifically called *depth-dependent magnification* in radiography. Consider the "stick" object of height w shown in Figure 5.16. It is clear from the geometry that the object will always appear larger on the detector than it is in reality. Furthermore, the height of the object on the detector is different depending on the position of the object in the field-of-view.

Using the concept of similar triangles, it is easy to show that when the object is at position z, its height w_z on the detector will be

$$w_z = w\frac{d}{z} . \tag{5.13}$$

From this, we see that the *magnification* $M(z)$ is given by

$$M(z) = \frac{d}{z} . \tag{5.14}$$

The magnification is a function of z, which is why this magnification is referred to as *depth-dependent*.

There are three important consequences of depth-dependent magnification. First, two objects within the body of the same size may appear to have different sizes on the radiograph. Thus, judgments about relative sizes of anatomical features must be made with caution and knowledge. Second, anatomical features studied longitudinally—that is, using radiographs of the same patient taken over several months or years—can only be compared in size if the same radiographic conditions and patient positioning is used. A change in the patient's weight, for example, may position a physiological landmark differently relative to the x-ray source causing an apparent change in size which could be falsely associated with the progression of disease. Third, the boundaries of a single object can be blurred simply because the "front" of the object is closer to the radiograph than its "back." This effect, which might be called *depth-dependent blurring*, is studied in the following example.

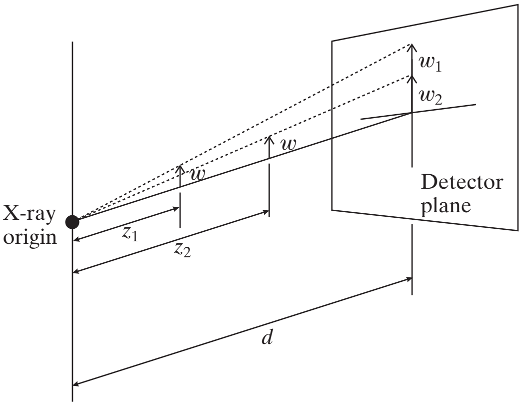


Figure 5.16
Depth-dependent magnification of a "stick" object.

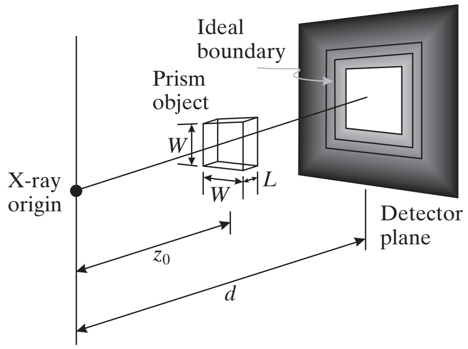


Figure 5.17
Imaging a rectangular prism.

EXAMPLE 5.4

Consider imaging the rectangular prism shown in Figure 5.17, defined by

$$\mu(x, y, z) = \mu_a \operatorname{rect}(y/W) \operatorname{rect}(x/W) \operatorname{rect}([z - z_0]/L)\,, \tag{5.15}$$

where [see (2.16)]

$$\operatorname{rect}(x) = \begin{cases} 1 & |x| \le 1/2 \\ 0 & \text{otherwise} \end{cases}\,.$$

Question How is this rectangular prism portrayed in a radiograph?

Answer In the center of the figure, the appearance of the image is governed by the inverse square law, obliquity, and path length variations, so (5.12) applies,

$$I_d(x, y) = I_0 \cos^3\theta\, e^{-\mu_a L/\cos\theta}\,.$$

However, when the x-rays are passing through the edges of the object, there is a loss of object path length and corresponding reduction in attenuation. Suppose a ray passes through the edge of the prism at range z'. Then, using simple trigonometry, we find that the path length through the prism is $[z' - (z_0 - L/2)]/\cos\theta$. (Notice that if $z' = z_0 + L/2$, this yields the path length $L/\cos\theta$, as above.) In this case,

$$I_d(x, y) = I_0 \cos^3\theta\, e^{-\mu_a(z'-(z_0-L/2))/\cos\theta}\,.$$

The third possibility is that the ray misses the prism entirely. In this case,

$$I_d(x, y) = I_0 \cos^3\theta\,.$$

Given (x, y) on the detector, what is required to complete this analysis is to determine which of the three ranges apply and if at the edge, the value of z'. This analysis is left as an exercise for the reader in Problem 5.9(a). ■

This example shows that divergent rays lead to edge blurring, an additional distortion to understand when viewing radiographs. This effect can be reduced by keeping the x-ray source as far from the detector plane as possible and keeping the object (the patient) as close to the detector as possible. For example, for routine chest radiography, the distance between the x-ray source and the detector (film) is typically six feet. However, for other examinations (such as imaging extremities), the source-to-detector distance may be much smaller.

Imaging Equation with Geometric Effects It is useful to develop an imaging equation that incorporates the geometric effects we have presented in this section. To do this, we utilize an idealized object $t_z(x, y)$ that is infinitesimally thin and located in a single plane given by the coordinate z and is capable of differentially attenuating x-rays as a function of x and y. This object can be thought of as a sheet of paper impregnated with a distribution of lead; a pictorial representation is given in Figure 5.13.

We should think of $t_z(x, y)$ as a *transmittivity*, rather than an attenuation; it replaces the entire exponential factor, rather than μ itself. Accordingly, if the object is located at the detector face—so that there is no magnification—the recorded intensity would be

$$I_d(x, y) = I_0 \cos^3\theta \, t_d(x, y)\,, \tag{5.16}$$

where

$$\cos\theta = \frac{d}{\sqrt{d^2 + x^2 + y^2}}\,. \tag{5.17}$$

In general, when the object is located at arbitrary z, where $0 < z \leq d$, the magnification must be taken into account. This effect is just a scaling of the size of the object, and it is

$$I_d(x, y) = I_0 \cos^3\theta \, t_z(x/M(z), y/M(z))\,. \tag{5.18}$$

Putting (5.18) together with (5.14) and (5.17) yields

$$I_d(x, y) = I_0 \left(\frac{d}{\sqrt{d^2 + x^2 + y^2}} \right)^3 t_z(xz/d, yz/d)\,. \tag{5.19}$$

This equation is a reasonable approximation for relatively thin objects that have nearly the same magnification and no variation in their attenuation in the z direction. This equation does not take into account the effect of integration along ray paths through thick objects (such as the human body).

5.3.3 Blurring Effects

It was shown above that divergent rays will blur edges of objects that are thick in the range direction. There are two effects, however, that will blur objects even if they do not have a z extent: extended sources and the detector thickness—that is, the intensifying screen in film-based systems, the PSP phosphor in CR systems, and the scintillator or s-Se plate in CCD, TFT, and CMOS digital radiography systems. Both of these processes can be modeled as convolutional effects that degrade image resolution. They are not present in (5.18) and (5.19) because in these equations there is only magnification and intensity changes made to the infinitesimally thin object $t_z(x, y)$. We now explore these two effects in detail.

Extended Sources Another characteristic of projection radiographs is blurriness arising from the nonzero extent of the x-ray source—the *spot size* of the site of x-ray emissions from the anode. As shown in Figure 5.18, this blurring

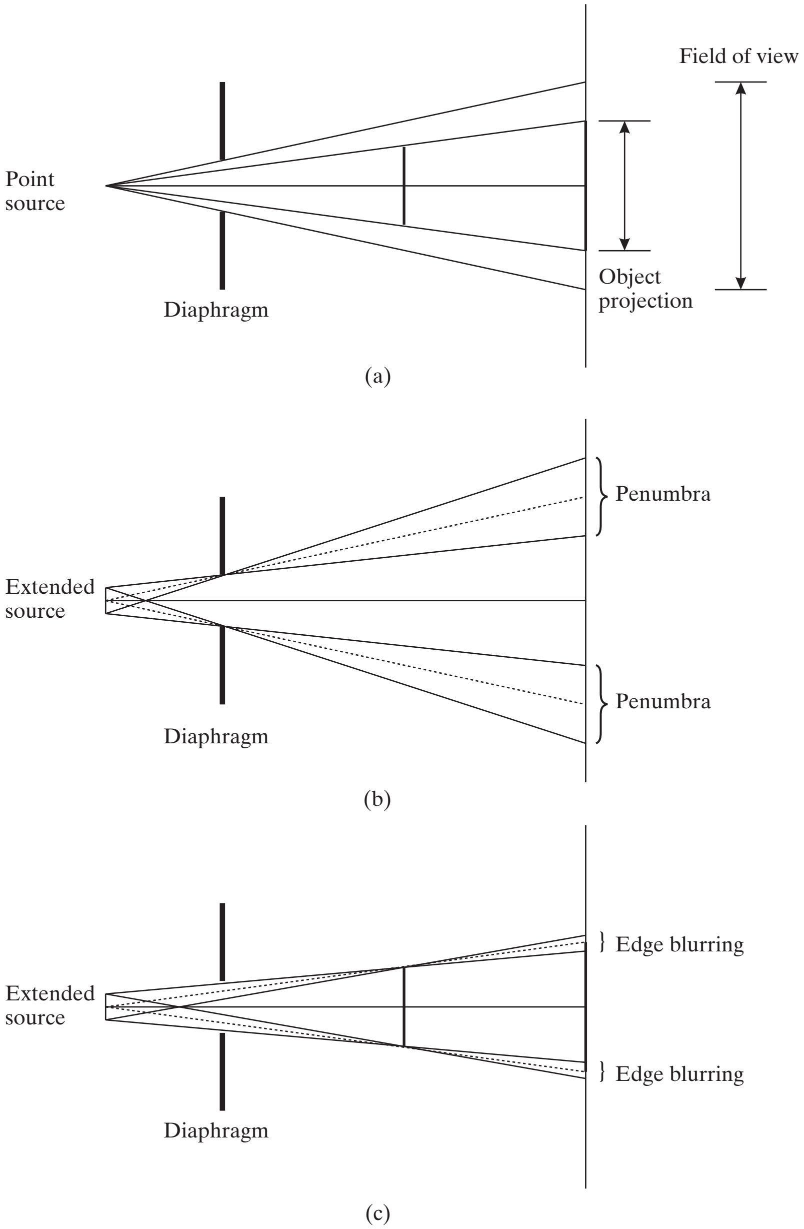


Figure 5.18
Effects of extended source. (a) Ideal field of view and object projection (with magnification). (b) Penumbra at edges of field of view due to extended source. (c) Blurred object edges due to extended source.

contributes to both a fuzziness at the edge of the field-of-view and a fuzziness of an object boundary. Here, we care about the blurring caused to the object. We will show that the extended source effect is a convolution of the source shape with the object shape (with magnification effects also included). Because of this, extended sources produce a significant loss of resolution of x-ray images. (This problem cannot be solved by shrinking the spot size on the anode because of heat dissipation requirements in the anode material.)

The physical extent of blurring caused by an extended source depends on the size of the source spot and the location of the object. Consider, for example,

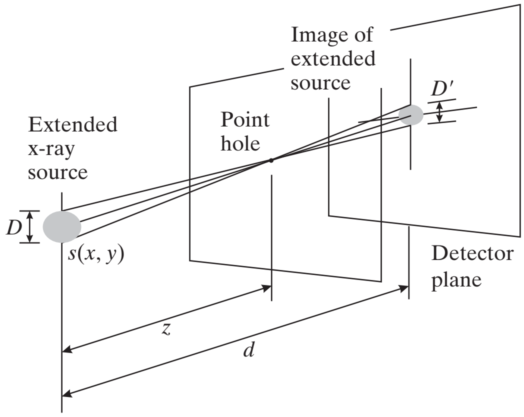


Figure 5.19
Principle of source magnification.

the image of the point hole shown in Figure 5.19. From the geometry, if the source has diameter D, then the image of the point hole located at range z will have diameter D', given by

$$D' = \frac{d-z}{z} D\,. \tag{5.20}$$

The factor multiplying D in (5.20) is the absolute value of the *source magnification* $m(z)$, which is given by

$$m(z) = -\frac{d-z}{z}\,. \tag{5.21}$$

The source magnification is negative because it inverts the image of the source, as can be appreciated by studying Figure 5.19. From (5.14), it can be seen that the depth-dependent (object) magnification and the source magnification are related as

$$m(z) = 1 - M(z)\,. \tag{5.22}$$

To develop an imaging equation incorporating source magnification, consider the image of the point hole on the z-axis, as shown in Figure 5.19. Let the x-ray tube source spot be represented by the source intensity distribution $s(x, y)$, which we can think of as a disk emanating x-rays from all points within the disk. Ignoring geometric effects, the image will be an inverted and spatially scaled version of the source intensity distribution. In fact, the intensity distribution is

$$I_d(x, y) = ks(x/m, y/m)\,, \tag{5.23}$$

where m is the source magnification for a point at range z, as given by (5.21). In particular, since m is negative, I_d will be the symmetric image of s.

The amplitude scaling term k in (5.23) must be found. It is determined using the fact that the integrated intensity on the detector plane must remain constant regardless of the position of the point hole on the z-axis. In other words,

$$\iint ks(x/m(z), y/m(z))\, dx\, dy = \text{ constant}\,, \tag{5.24}$$

regardless of z, but k might depend on z. Using the Fourier transform,

$$km^2(z)S(0,0) = \text{ constant}, \tag{5.25}$$

which implies that

$$k \propto \frac{1}{m^2(z)}. \tag{5.26}$$

This implies that the amplitude of the blurring caused by the extended source depends on the location of the object relative to the detector plane. In particular, if the object approaches the detector, then $m(z) \to 0$, and

$$\frac{s(x/m, y/m)}{m^2} \to S(0,0)\delta(x,y). \tag{5.27}$$

In this case, there is no loss of resolution or amplitude; the image of the extended source is actually perfect.

The above analysis holds for a point hole located anywhere in the same z-plane. We can then use superposition [see (2.33)] to understand the behavior of a spatial distribution of attenuating objects within a given z plane, such as the transmittivity function $t_z(x,y)$ used before in describing imaging under object magnification [see (5.19)]. This analysis leads to an imaging equation that is the convolution of the magnified object by the magnified and scaled source function,

$$I_d(x,y) = \frac{\cos^3\theta}{4\pi d^2 m^2} s(x/m, y/m) * t_z(x/M, y/M). \tag{5.28}$$

Notice that the term $4\pi d^2$ is included to account for the fact that there is loss of source intensity due to the inverse square law. For objects close to the detector, $M \approx 1$ and $m \approx 0$; the object will have unity magnification and will not be blurred by the source focal spot, regardless of the size of that focal spot. This is just one more motivation for putting the patient directly against the detector.

Film-Screen and Digital Detector Blurring The detailed geometry of a film-screen detector is shown in Figure 5.20. The film is sandwiched between two

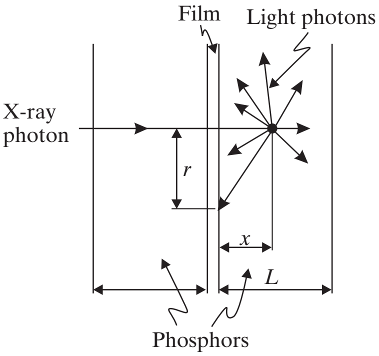


Figure 5.20
A film-screen detector.

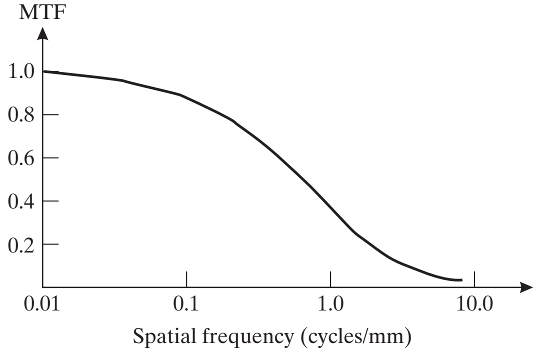


Figure 5.21
A typical MTF for a film-screen detector.

phosphors—that is, the active parts of intensifying screens. For each x-ray photon that is absorbed by a phosphor (by photoelectric absorption), a large number of lower-energy light photons are produced. Unlike x-ray photons, these light photons are efficiently detected by film emulsions (see Section 5.2.5). The light photons travel (approximately) isotropically from the point where the x-ray was absorbed, as shown in Figure 5.20. They can then be absorbed in the film at locations far from the x-ray's path. The union of detected light photons form a "spot" on the film, which is effectively an impulse response to the x-ray "impulse." Using superposition, an imaging equation incorporating film-screen blurring is readily derived. To good approximation, it is simply (5.28) with an additional convolution with the film-screen impulse response function $h(x, y)$,

$$I_d(x,y) = \frac{\cos^3\theta}{4\pi d^2 m^2} s(x/m, y/m) * t_z(x/M, y/M) * h(x,y)\,. \tag{5.29}$$

Typically, the impulse response of a film-screen detector system is circularly symmetric. The modulation transfer function (MTF) of a typical film-screen detector is shown in Figure 5.21.

Use of a thinner phosphor yields less film-screen blurring. However, thin screens do not stop as many x-rays, reducing the efficiency of the detector system. Detector *efficiency* η is defined as the fraction of photons captured by the detector on average. Calcium tungstate is often used for intensifying screens because it has reasonably good efficiency for stopping x-ray photons ($\eta \approx 0.30$). In general, the thicker the screen, the higher the efficiency and the worse the resolution. We will see in Section 5.4.1 that maintaining high efficiency is key to keeping image noise low. Thus, while thinner screens are desirable for good resolution, there is a tradeoff below which the improvements in resolution will be offset by increased noise.

Digital detectors have different mechanisms that produce blurring at the detector. Like film-screen blurring, there are effects derived from the thickness of the scintillators (when used). In addition, there is the effect in some systems of discrete detector "bins" that accumulate charge over specific regions within the detector. To a first approximation, all digital systems can be modeled by the convolution model in (5.29) but each having their own unique PSF $h(x, y)$.

5.3.4 Film Characteristics

As noted in Section 5.2.7, computed radiography and digital radiography systems are replacing conventional film-screen approaches. Even so, it is important to understand the basic terminology and characteristics of film, as similar concepts can be applied to CR and DR.

The intrinsic spatial resolution of film is generally significantly better than that of the intensifying screen; so, for most situations, we can ignore the spatial frequency characteristics of the film. If necessary or desirable, the spatial resolution of film can be modeled by a point spread function analogous to that of the detector blur above; the overall performance would be a cascade of subsystems, as depicted in (2.46), (3.29), and (3.30). Rather than a film's spatial resolution properties, it is its intensity transformation properties that are of primary interest here.

Because film is a very poor direct detector of x-rays, we will ignore the image created directly on the film by the x-rays themselves and only consider the image created by the light produced in the phosphors adjacent to the film. The parameters of interest are the contrast or *gamma* of the film, and its dynamic range or *latitude*; we will define these terms in a moment. We use many different types of film with varying contrast and latitude depending upon the type of x-ray examination desired. Those of you who are amateur photographers may have a practical appreciation of these film characteristics.

We consider the relationship between the amount of light that is incident upon the film and the final image that is viewed when the film is exposed, developed, and placed on a light source ("light box"). The light photons that are captured by the film produce a *latent image* (i.e., it is not visible). When the film is developed, that latent image produces a blackening of the film. We therefore characterize film as a transformation between exposure to light and the degree of blackening of the film, which is characterized by the *optical density* of the film.

When a developed film is viewed on a light box, the blacker parts of the film absorb more of the light coming from the box, creating the impression of a gray level image when viewed by the eye. The *optical transmissivity* is defined as the fraction of light transmitted through the exposed film or

$$T = \frac{I_t}{I_i}, \tag{5.30}$$

where I_i is the irradiance of the incident light and I_t is the irradiance of the transmitted light, both in units of energy per unit area per second. For example, the darker parts of the film might have a transmissivity of 0.1, while the more transparent parts might have a transmissivity of 0.9. The *optical opacity* is the inverse of the transmissivity, and the *optical density* is defined to be the common logarithm of the optical opacity,

$$D = \log_{10} \frac{I_i}{I_t}. \tag{5.31}$$

We see that the optical density characterizes how black the film is on a logarithmic scale. Optical densities in the range of 0.25–2.25 are usable, but the human eye discriminates shades of gray best when $1.0 < D < 1.5$.

We now appreciate that the appearance of the x-ray film after exposure and development is determined by the spatial distribution of optical density, $D(x, y)$. How does x-ray exposure relate to optical density? We first recall from Chapter 4 that x-ray exposure X is measured in units of roentgens (R); in particular, 3,876 R of exposure is equivalent to 1 C of charge per kilogram of air; the charge comes from the ions produced by the x-ray interactions. For

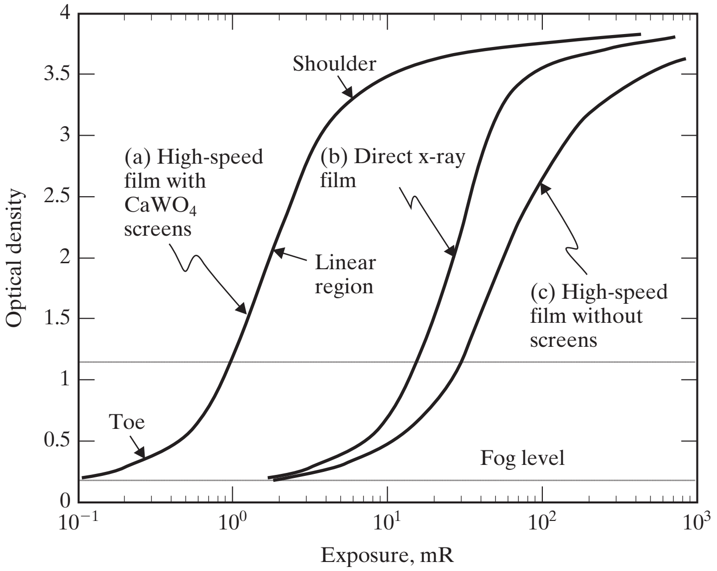


Figure 5.22
An H&D curve.

direct exposure (no screens) and $D < 2$, D is directly proportional to X. When screens are used and $D > 2$, however, the relationship between X and D is nonlinear, and it is given by the classic *H&D curve* (named after Hurter and Driffield, who first described this function), as shown in Figure 5.22. A typical H&D curve is S-shaped, with a low exposure–low D *toe*, a linear portion, and a high exposure–high D *shoulder*. Notice that D never goes to zero; there is always a *base fog* optical density, even in the absence of exposure.

In the linear region of an H&D curve, the response between the common logarithm of exposure and optical density is approximately linear,

$$D = \Gamma \log_{10}(X/X_0)\,, \tag{5.32}$$

where Γ is the slope of the H&D curve in the linear region and X_0 is the exposure at which the linear region would hit the horizontal axis ($D = 0$) of the H&D curve.

The quantity Γ is called the *film gamma*. It is unique to a particular film and a particular method of developing that film (that is, it is a function of the developing chemicals, temperature, time of development, etc.), and it is typically in the range of 0.5–3. Increasing the gamma of the film increases contrast but reduces the range of exposures over which the optical density response is linear. Correct exposure of a film occurs only in the linear range of its H&D curve. The *latitude* of a film is the range of exposures over which the H&D curve is linear. The *speed* of a film is the inverse of the exposure at which $D = 1 +$ fog level. Consider the radiographic films shown in Figure 5.22. Clearly, the film/screen combination (a) is the fastest film. The direct x-ray exposure film (b) is next fastest and has the largest gamma and lowest latitude. Finally, standard high-speed optical film (c) without screens is the slowest; it would require the largest x-ray exposure of all three films in order to achieve optical densities in the typical range of 1–1.5.

5.4 Noise and Scattering

Until now, we have only considered the effects of the x-ray source and the composition of the object. We have implicitly assumed that the intensity of the x-ray beam at the detector plane is faithfully reproduced by the detector. In reality, of course, the detector does not faithfully reproduce the incident intensity distribution. Furthermore, x-rays arrive in discrete packets of energy (*quanta* or *photons*, as discussed in Section 4.3.2). The discrete nature of their arrival can lead to fluctuations in the image. In the following sections, we discuss some of the additional factors introduced by x-ray physics and the image recorder.

5.4.1 Signal-to-Noise Ratio

Assuming unity magnification and infinitesimal source size, a rectangular object, like that depicted in Figure 5.17, will cast a rectangular shadow on the detector with dimensions equal to the object dimensions. A 1-D slice through the intensities on the detector will look like a rect function, as defined by (2.19) and shown in Figure 5.23. Let I_b be the background intensity and I_t be the object (target) intensity on the detector plane, with $I_t > I_b$ in this example. As discussed in Section 3.2, the *local contrast* of this object is given by

$$C = \frac{I_t - I_b}{I_b} \,. \tag{5.33}$$

Because the x-rays arrive in discrete quanta (photons), there will be random fluctuations in the number of photons arriving in each small area of the detector, leading to *noise*, as discussed in Section 3.4. This effect is called *quantum mottle*, and it is responsible for the imprecision of detector measurements of x-ray intensity.

We can quantify the effect of noise on image formation using the concept of SNR, as presented in Section 3.5. The higher the SNR, the less evident will be the granularity in the image resulting from this quantum effect. In this scenario, the "signal" is $I_t - I_b$, the difference between the target intensity and the background intensity. The noise is caused by quantum-related fluctuations in the background, and it can be characterized by a standard deviation σ_b. Therefore, the basic SNR in this scenario is given by

$$\text{SNR} = \frac{I_t - I_b}{\sigma_b} = \frac{C I_b}{\sigma_b} \,. \tag{5.34}$$

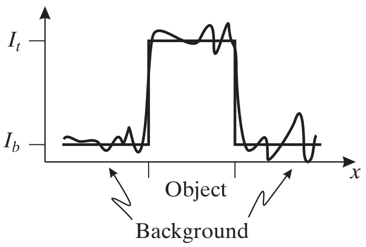


Figure 5.23
Detector intensities from a rectangular object.

Suppose that all photons arriving at the detector have the same energy, $h\nu$, which might be called the *effective energy* in the actual case of a polyenergetic x-ray beam (see Example 4.4). In this case [combining (4.14) and (4.17)], the intensity is related to the number of photons by

$$I = \frac{Nh\nu}{A\Delta t}\,, \tag{5.35}$$

where A is a small area on the detector—for example, a pixel—and Δt is the duration of the x-ray burst. The number of photons N is a Poisson random variable, and in the background, we can denote the average number of photons per burst per area A as N_b. Then, the average background intensity is given by

$$I_b = \frac{N_b h\nu}{A\Delta t}\,. \tag{5.36}$$

Also, the variance of the number of photons per burst per area A in the background is given by

$$\sigma_b^2 = N_b \left(\frac{h\nu}{A\Delta t}\right)^2\,. \tag{5.37}$$

From (5.34), (5.35), (5.36), and (5.37), the local SNR is given by

$$\mathrm{SNR} = C\sqrt{N_b}\,. \tag{5.38}$$

Equation (5.38) reveals a fundamental tradeoff in x-ray imaging. In order to improve the visibility of a particular structure in a radiograph, it is necessary to either increase the contrast of the structure or to increase the number of photons used in the visualization or analysis (or both). The contrast of the structure might be improved by changing the energy of the x-rays or by using a contrast agent or dual-energy techniques. (See Problem 5.20.) The number of photons might be increased by several methods: by increasing the filament current, the duration of the x-ray pulse, or the energy of the x-rays (which would then penetrate the body better), or by using larger area elements (pixels) or a more efficient detector.

Recall that the x-ray tube provides a spectrum of x-ray energies ranging from low energies to the maximum energy determined by the peak kilovoltage applied to the x-ray tube anode. By changing the kVp, we modify the x-ray energy. Contrast, SNR, and patient radiation exposure (dose) are markedly affected by changes in the kVp. For example, at low kVp and therefore low x-ray energies, image contrast is high because the difference between the attenuation of different body tissues increases as energy decreases. However, the body is less transparent to x-rays at lower energies, so that fewer x-ray photons penetrate the body. Hence, although C is high, N is low, and it turns out that SNR is low. Because the body is highly absorbent to these low energy x-ray photons, the patient radiation dose is high.

At high energies, on the other hand, contrast is low, as the attenuation coefficients for different types of body tissues become similar. While the body is highly transparent at high energies, the number of x-ray photons per Roentgen decreases, and therefore for a given radiation exposure, the SNR will be low at very high energies. Between high and low x-ray energies, the SNR achieves a maximum at a point where there is reasonably good tissue contrast, the body is still relatively transparent to x-rays, and the number of photons per Roentgen is high.

We can express the SNR in more detail by adding several additional concepts. Since we indicated above that SNR is proportional to the number of photons per unit area of detector, if we increase the unit area we also increase the number of photons. Therefore, a detailed expression for SNR is given by

$$\mathrm{SNR} = C\sqrt{\Phi A R t \eta}\,, \tag{5.39}$$

where Φ is the number of photons per Roentgen per cm^2, A is the unit area, R is the body's radiation exposure in Roentgens, t is the fraction of photons transmitted through the body, and η is the detector efficiency.

EXAMPLE 5.5

Consider the following parameters, which are from a typical chest x-ray:

$$\Phi = 637 \times 10^6 \text{ photons R}^{-1}\text{ cm}^{-2}$$

$$R = 50 \text{ mR}$$

$$t = 0.05$$

$$\eta = 0.25 \text{ (25\% efficiency)}$$

$$A = 1 \text{ mm}^2$$

Question What is the SNR of a lesion having 10% contrast, that is, $C = 0.1$?

Answer Using (5.39), we compute that SNR = 16 dB. ■

5.4.2 Quantum Efficiency and Detective Quantum Efficiency

In projection radiography, x-rays that are not stopped by the body should be detected. Ideally, all photons incident upon a detector would be detected and characterized by their location, energy, and time-of-arrival. However, no detector can achieve these ideals. Instead, detector designs must emphasize certain performance measures at the expense of others, and these tradeoffs influence the quality of the measured radiographs. In this section, we develop the concept of *detective quantum efficiency* (DQE), which represents one way to characterize the performance of a detector that has a direct relation to the SNR of the images it produces.

Quantum Efficiency Ideally, all photons incident upon a detector should be detected. What does this really mean? In order to be "detected," an incident photon must interact with the detector—for example, it must be stopped by a photoelectric interaction—and that interaction must produce some measurable output—for example, a flash of light, a collection of ionized atoms, or an electrical current. *Quantum efficiency* (QE) is the probability that a single photon incident upon the detector will be detected; it is a basic property of any x-ray detector.

A detector with a higher QE does not necessarily mean that the detector or the image it produces is better, however. Consider two detectors having the same thickness and linear attenuation coefficients. They are capable of stopping the same fraction of incident photons; their basic "stopping power" is the same. Suppose that the output of a detected event is highly predictable in the first detector and highly variable in the second. We intuitively understand that the first detector is better. In order to more accurately characterize the quality of a detector, we must account for these variations in detector output given a single event. That is the goal of the parameter DQE.

Detective Quantum Efficiency To better characterize detector performance, detective quantum efficiency (DQE) considers the transformation of SNR from a detector's input to its output. In this way, DQE moves away from simply counting photons as in QE. The *detective quantum efficiency* is defined as

$$\mathrm{DQE} = \left(\frac{\mathrm{SNR}_{\mathrm{out}}}{\mathrm{SNR}_{\mathrm{in}}}\right)^2 , \tag{5.40}$$

where $\mathrm{SNR}_{\mathrm{in}}$ is the intrinsic SNR of the incident radiation and $\mathrm{SNR}_{\mathrm{out}}$ is the SNR of the measured quantity—for example, flash of light, detector voltage, film density, etc. DQE can be viewed as a measure of the degradation in the SNR due to the detection process. It can also be viewed as the fraction of photons that are detected "correctly." In general $\mathrm{DQE} \le \mathrm{QE} \le 1$.

EXAMPLE 5.6

Consider a hypothetical detector having $\mathrm{QE} = 0.5$ and the ability to perfectly localize every photon that is stopped by the detector.

Question What is the DQE of this detector?

Answer Suppose a photon burst with an average of $\overline{N}$ is incident upon the detector. The SNR of the input is just the intrinsic SNR of the photon burst, which is given by

$$\mathrm{SNR}_{\mathrm{in}} = \sqrt{\overline{N}}\,.$$

Since $\mathrm{QE} = 0.5$, we know that $0.5 \times \overline{N}$ photons (on average) are stopped by the detector. These photons are perfectly localized, meaning that the detector produces an accurate count of the number and position of all detected photons no matter what pixel resolution is desired. Therefore, the output SNR is

$$\mathrm{SNR}_{\mathrm{out}} = \sqrt{0.5\overline{N}}\,,$$

and the DQE is

$$\begin{aligned}\mathrm{DQE} &= \left(\frac{\mathrm{SNR}_{\mathrm{out}}}{\mathrm{SNR}_{\mathrm{in}}}\right)^2 \\ &= \left(\frac{\sqrt{0.5\overline{N}}}{\sqrt{\overline{N}}}\right)^2 \\ &= 0.5\,.\end{aligned}$$

In this case, $\mathrm{DQE} = QE$, which occurs because of the hypothetical perfect localization of all detected photons. ■

EXAMPLE 5.7

Suppose that an x-ray tube is set up to fire n 10,000-photon bursts at a detector and the detector's output x is recorded as x_i, $i = 1, \ldots, n$. Suppose that the mean and variance of $\{x_i\}$ is found to be $\overline{x} = 8,000$ and $\sigma_x^2 = 40,000$, respectively.

Question What is the DQE of this detector?

Answer Think of $\overline{x} = 8,000$ as an estimate of the average number of photons in the output of the detector. If the detector were simply counting photons, then the variance should also be 8,000, since photon count is a Poisson random variable. Yet, the measured variance is 40,000, five times the Poisson variance. This is conclusive evidence that there is variation in the detector's response to detected photons. The SNR of the input is simply the intrinsic SNR of the photon burst,

$$\mathrm{SNR}_{\mathrm{in}} = \frac{10{,}000}{\sqrt{10{,}000}} = 100\,.$$

The SNR of the output is

$$\mathrm{SNR}_{\mathrm{out}} = \frac{8{,}000}{\sqrt{40{,}000}} = 40\,.$$

Therefore, the DQE for this detector is

$$\mathrm{DQE} = \left(\frac{40}{100}\right)^2 = 0.16\,.$$

Only about 16 percent of the incident photons are detected "correctly." ■

5.4.3 Compton Scattering

Earlier, we noted that Compton scattering degrades image quality. The reason for this is that Compton photons are deflected from their ideal straight-line path, and some are detected in locations away from the correct, straight-line location. This produces two unwanted results: a decrease in image contrast and a decrease in SNR.

Effect on Image Contrast Compton scatter has an negative effect on image contrast. Consider a target with local contrast [see (3.12)] given by

$$C = \frac{I_t - I_b}{I_b}\,.$$

Compton scatter adds a constant intensity I_s to both target and background intensity, yielding a new contrast of

$$\begin{aligned} C' &= \frac{(I_t + I_s) - (I_b + I_s)}{I_b + I_s} \\ &= C\frac{I_b}{I_b + I_s} \\ &= \frac{C}{1 + I_s/I_b}\,. \end{aligned} \tag{5.41}$$

Therefore, the effect of scatter is to reduce contrast by the factor $1/(1 + I_s/I_b)$. The ratio I_s/I_b is called *scatter-to-primary ratio*; clearly, it should be kept as small as possible in order to preserve contrast.

Signal-to-Noise Ratio with Scatter The derivation of SNR in the presence of Compton scattering follows the Compton-free derivation very closely:

$$\begin{aligned} \mathrm{SNR}' &= \frac{I_t - I_b}{\sigma_b} \\ &= C\frac{I_b}{\sigma_b} \\ &= C\frac{N_b}{\sqrt{N_b + N_s}} \\ &= \frac{C\sqrt{N_b}}{\sqrt{1 + N_s/N_b}} . \end{aligned} \tag{5.42}$$

Here, the symbol N_s stands for the number of Compton scattered photons per burst per area A on the detector, and the symbol C is the underlying contrast (not the Compton scatter-reduced contrast).

Equation (5.42) shows the reduction of SNR due to Compton scattering. The SNR with Compton scattering is related to the SNR without Compton scattering by

$$\mathrm{SNR}' = \mathrm{SNR}\frac{1}{\sqrt{1 + I_s/I_b}} \tag{5.43}$$

Thus, in addition to the loss in contrast, there is an additional loss of $1/\sqrt{1 + I_s/I_b}$ in SNR due to the effects of Compton scattering.

EXAMPLE 5.8

Suppose 20 percent of the incident x-ray photons have been scattered in a certain material before they arrive at detectors.

Question What is the scatter-to-primary ratio? By what factor is the SNR degraded?

Answer The x-ray photons that contribute to the background intensity are those that hit the detectors without Compton scattering. The number of these photons is $0.8N$, where N is the number of incident x-ray photons. The number of scattered photons is $0.2N$. Because the intensity of the image is proportional to the number of photons detected, we have

$$I_b \propto 0.8N, \quad I_s \propto 0.2N .$$

The scatter-to-primary ratio is

$$\frac{I_s}{I_b} = \frac{0.2N}{0.8N} = \frac{1}{4} .$$

From (5.42), we can see that the loss of SNR is $1 - 1/\sqrt{1 + I_s/I_b} = 0.11$. So the Compton scattering introduces 11 percent loss of SNR. ■

5.5 Summary and Key Concepts

Projection radiography is the oldest and most fundamental medical imaging modality. This modality uses x-ray tubes and either film or digital detectors to record images of the "shadow" created by body tissues. In this chapter, we presented the following key concepts that you should now understand:

1. *Projection radiography* produces *radiographs*, which are 2-D projections of 3-D objects.
2. A *projection radiography system* consists of an x-ray tube, devices for beam filtration and restriction, compensation filters, grids, and a film-screen or digital detector.
3. The *basic imaging equation* describes the energy- and material-dependent attenuation of the x-ray beam produced by the system as it passes through the patient.
4. The basic imaging equation must be modified by several geometric effects, including the *inverse square law*, *obliquity*, *beam divergence*, *anode heel effect*, *path length*, and *depth-dependent magnification*.
5. The film-screen detector produces an *optical image on film*; the degree of film blackening—the *optical density*—depends on film exposure in a nonlinear way characterized by the H&D curve.
6. *Digital detectors* convert x-rays by direct or indirect means to electronic signals that can be read directly into a computer and viewed on a computer display.
7. *Noise* arising from the random nature of x-ray production and transmission reduces an image's SNR, and thus the *detective quantum efficiency* of the system.
8. Acceptance of *Compton scattered photons* reduces both image contrast and SNR.

Further Reading

Bushberg, J.T., Seibert, J.A., Leidholdt, E.M., and Boone, J.M. *The Essential Physics of Medical Imaging*, 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins, 2012.

Carlton, R.R. and Adler, A.M. *Principles of Radiographic Imaging: An Art and a Science*, 5th ed. Clifton Park, NY: Delmar Cengage Learning, 2012.

Johns, E.J. and Cunningham, J.R. *The Physics of Radiology*, 4th ed. Springfield, IL: Charles C Thomas Publisher, 1983.

Macovski, A. *Medical Imaging Systems*, Englewood Cliffs, NJ: Prentice Hall, 1983.

Webster, J.G. *Medical Instrumentation: Application and Design*, 4th ed. New York, NY: Wiley, 2009.

Wolbarst, A.B. *Physics of Radiology*, 2nd ed. Norwalk, CT: Appleton and Lange, 2005.

Problems

Instrumentation

5.1 Examine the projection radiography system in Figure 5.13. Is this system linear? Is it shift invariant?

5.2 (a) What determines the highest energy of x-ray photons emitted from an x-ray tube? What determines the energy spectrum of the x-ray photons?

(b) In radiographic imaging, why are low-energy photons undesired? What measures can be taken to reduce the number of the low-energy photons entering the human body?

(c) What is beam hardening? What are the causes of beam hardening?

5.3 We are designing an x-ray tube and want to use a lighter filter. Suppose we have, in addition to copper, a new material with a density of 5,000 kg/m^3 and a mass attenuation coefficient of 0.08 m^2/kg at 80 kVp. In order to satisfy the NCRP recommendation of 2.5 mm Al/Eq at 80kVp, what should we use, copper or the new material? Explain.

5.4 (a) Why are iodine and barium commonly used as contrast agents?

(b) Explain with a diagram why an airgap can be used to reject scattered photons. What is the main problem with airgaps?

5.5 (a) Why is Compton scattering bad for the images produced using projection radiography?

(b) For a film with a given H&D curve, why is it better to have a range of x-ray exposures that are in the linear region of the curve?

(c) Why do we have to filter x-rays from the x-ray source?

(d) Lead strips can be used (e.g., in a grid) to reduce scatter in radiography systems. If the height of the grid is eight times its width, what is the maximum scatter angle of the photon after which the photon cannot pass through the grid? (Assume an orientation that is orthogonal to the grid lines.)

Image Formation

5.6 Suppose a chest x-ray imaging system with a perfect point source can provide images with acceptable quality when the intensity variation is smaller than 5% when imaging a slab of material with uniform attenuation coefficient. Suppose the source-to-detector distance is 2 m, what is the maximal size of the image?

5.7 (a) Derive a simple expression for the magnification of a thin object on a projection radiograph. Assume x-ray source-to-object and source-to-detector distances.

(b) In practice, what simple strategies can an x-ray technician use to reduce the magnification and distortion effects of the projection radiography system?

5.8 You are designing a digital x-ray detector panel with an array of discrete detectors. You decide that you are able to compensate for the beam divergence effect by weighting the outputs of your detectors appropriately. This way, the effective beam intensity should be uniform throughout the entire image.

(a) Assume your detectors are very small and very close together and that the detector panel is a distance d away from the x-ray source. Find an appropriate weighting (as a function of r_d) that will compensate for the beam divergence variation.

(b) Do these adjustments improve the image quality? Explain.

5.9 Consider the radiographic image of the prism, as developed in Example 5.4.

(a) Determine an expression for the intensity image $I_d(x, y)$ of the prism.

(b) Plot the intensity profile along the line $y = 0$ on the detector plane.

(c) Write an expression for the image on the developed radiographic film of the prism.

5.10 Derive (5.28) from first principles.

Image Quality

5.11 Assume that each photon incident upon an x-ray image intensifier is detected with probability (QE) p, independently from other photons. If the number $N(t)$ of photons that arrive at the detector within the time interval $[0, t)$ follows a Poisson distribution with mean μt, find the probability mass function (PMF) of the number $D(t)$ of photons detected in the time interval $[0, t)$.

5.12 Consider two 1-D functions

$$h_1(x) = e^{-x^2/5} \quad \text{and} \quad h_1(x) = e^{-x^2/10}.$$

(a) Suppose $h_1(x)$ is the PSF of the extended source in a projection radiography system for an object at range $z = 3d/4$, where d is the source-to-detector distance. What is the PSF of the extended source as a function of z?

(b) In this same projection radiography system, the intensifier screen causes a blurring that can be modeled as $h_2(x)$. Find the MTF of the overall blurring as a function of source magnification, m.

(c) What is the FWHM of the system as a function of m?

5.13 We examine the effect on the SNR of an increase in the recorded scatter fraction from 0.35 to 0.8. Suppose that a 1 mm-thick bone presents a scatter-free subject contrast of 0.08, and suppose that 1,000 photons/mm^2 were collected. What would happen if the detector absorption efficiency were halved?

5.14 Show that the quantum efficiency QE of a detector is at least as large as its DQE; that is,

$$\text{DQE} \leq \text{QE}.$$

5.15 Suppose that an x-ray tube fires, on average, 10,000 photons per second at an x-ray image intensifier. If the detector provides an average count of 10,000 photons per second, calculate and plot the variance of the detector's output as a function of DQE. What DQE value is required for an output variance of 2,000?

5.16 Calculate the DQE as a function of (u, v) for a nonideal detector with PSF

$$h(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2},$$

by assuming that the input to the detector is white Poisson noise.

5.17 Figure P5.1 shows a projection radiograph of an off-axis, hollow, plastic cylinder. Make a simple sketch of the cylinder's position and orientation relative to the source and detector, and explain the distortion in the image.

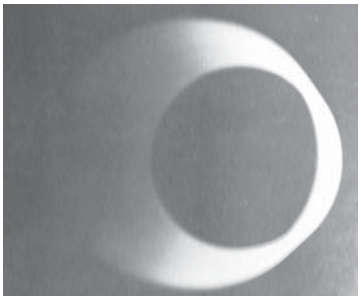


Figure P5.1
A projection radiograph of a plastic cylinder. (From A. Macovski, Medical Imaging Systems, 1983.)

5.18 Consider the projection radiography system depicted in Figure P5.2. Express the magnification ratio $m_{12} = s_1/s_2$ of the projections of the two objects O_1 and O_2 in terms of s, ϕ, d, d_1, and d_2. If $s = 5$ cm, $\phi = 45^\circ$, $d = 120$ cm, $d_1 = 40$ cm, and $d_2 = 80$ cm, what is the value of m_{12}?

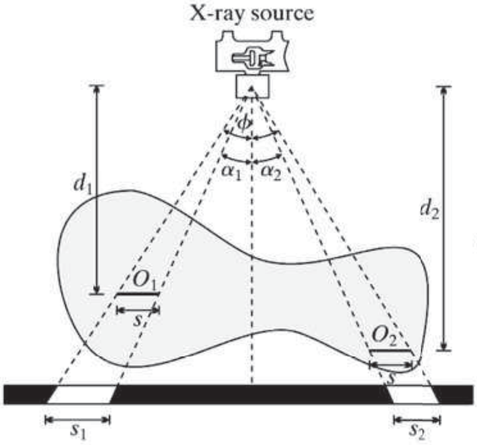


Figure P5.2
An illustration of depth-dependent magnification.

5.19 Consider the x-ray projection radiography system shown in Figure P5.3. The linear attenuation coefficients are μ_A and μ_B for A and B, respectively.

The detector is large enough to image object A and B. Suppose the geometric effects can be ignored and the x-rays are parallel and horizontal and are wide enough to image the objects.

(**a**) Assume the intensity of the incoming x-ray beam is I_p. Give the intensity distribution $I(d, y)$ on the detector for $|y| \leq a$.

(**b**) What is the local contrast of object B on the detector? (Use the intensity at $y = 0$ as the target intensity.)

(**c**) Under what condition is the local contrast of object B positive?

Now let us consider the geometric effects. A point x-ray source lies at the origin $(0, 0)$.

(**d**) Consider the inverse square law, obliquity, and path length. Ignore scattering. Suppose without object A and B, at the center of the detector (on the x-axis), the intensity is I_0. Give the intensity at $(d, \frac{r}{\sqrt{z^2+2rz}}d)$ on the detector.

(**e**) For a point on the detector (d, y_0), what point y on the centerline (i.e., $x = z + r$) of object A and B, contributes to the intensity $I(d, y_0)$?

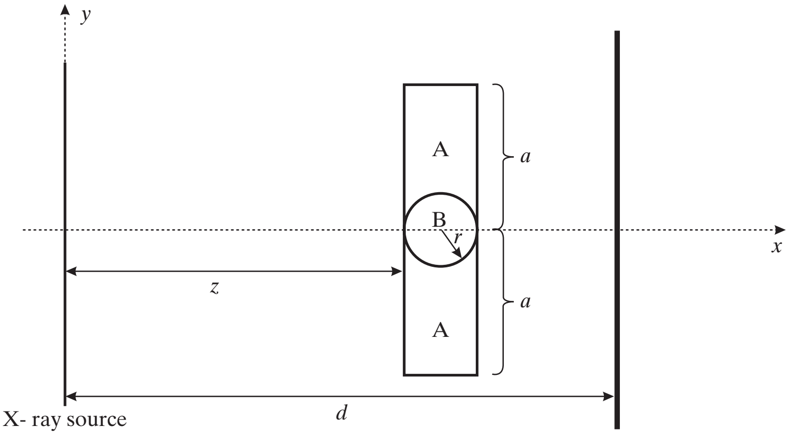


Figure P5.3
System geometry for Problem 5.19.

Applications

5.20 Interesting properties of tissues can sometimes be revealed by imaging them at two x-ray energies. Suppose you are going to use your conventional chest x-ray setup to make two films, one where the highest x-ray energies are at 30 keV and the other where the highest energies are at 100 keV.

(**a**) In order to make these two images, what physical parameter will you change and to what values will you set it?

(**b**) Assuming nothing else changes, which film would you expect to be more exposed than the other? Explain.

(**c**) At which energy is Compton scattering more of a problem? Describe how it will affect that image.

(**d**) At which energy is the absorbed dose to the patient higher?

(e) Suppose you took the two exposed films and subtracted their optical densities, creating a third image: $D(x, y) = D(x, y; E_h) - D(x, y; E_l)$. Describe in mathematical terms what $D(x, y)$ measures.

5.21 Suppose you have a 20-cm-thick slab of material. You measure the fraction $t(E)$ of x-rays that get through this material (on average) over the energy band 50–150 keV, and discover that it is well modeled by the relation

$$\log_{10} t(E) = -\frac{(E[\text{keV}] - 150)^2}{5{,}000}.$$

(a) Find an expression for the linear attenuation coefficient of this material as a function of x-ray energy E.

Suppose a 5-cm cube of a second material replaced a 5-cm cube of original material from the center of the slab and suppose this second material has linear attenuation coefficient of 0.15 cm^{-2} at 75 keV.

(b) What is the intrinsic contrast of this new material relative to the slab at 75 keV? (Intrinsic contrast is defined by the linear attenuation coefficients alone, as if you could image them directly.)

(c) What is the contrast of this new material if imaged through a standard radiographic system at 75 keV? (Ignore photon noise, detector efficiency, beam divergence, scatter, and magnification effects.)

5.22 A hypothetical x-ray tube has been designed. An x-ray burst from this tube yields exactly 10^4 x-ray photons at energy $E_1 = 60$ keV and 10^5 x-ray photons at energy $E_2 = 65$ keV.

(a) Draw the energy spectrum associated with this x-ray pulse. Explain why the units associated with the numbers 10^4 and 10^5 in the spectrum must be photons-keV.

In the following parts of this question, assume that the above x-ray tube is used as a source in an imaging system to image a phantom, as shown in Figure P5.4. In addition, make the following simplifying assumptions:

- Assume a parallel beam of x-rays that are uniformly distributed over an extended source.
- Assume that the x-rays passing through the phantom undergo the photoelectric effect only; that is, neglect the effect of Compton scatter.

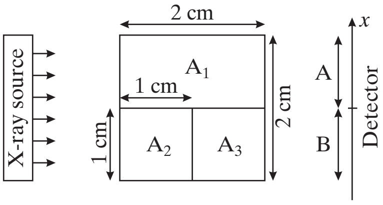


Figure P5.4
A phantom being imaged by an x-ray system. See Problem 5.22.

The phantom consists of three areas A_1, A_2, and A_3, as labeled. The linear attenuation coefficients μ_1, μ_2, μ_3 (units of cm^{-1}) of these three areas at the energy levels present in the x-ray beam are given in the table below.

TABLE P5.1

	E_1	E_2
μ_1	0.2	0.4
μ_2	0.3	0.1
μ_3	0.5	0.4

(**b**) Draw the total number of x-ray photons per cm that hit the detector as a function of position x.

(**c**) Assume the x-ray intensity is proportional to the number of photons. Calculate the contrast of the image observed at the detector as a function of position x assuming that A is the target and B is the background.

(**d**) Make a rough sketch of the profile of optical density as a function of position x assuming x-ray film is used as the detector. Which part of the film is more transparent?

5.23 An x-ray imaging system is set up as shown in Figure P5.5. All length units in the figure and in equations below are in cm. The overall system response can be modeled as

$$I_d(x_d, y_d) = Ks(x_d/m, y_d/m) * t(x_d/M, y_d/M),$$

where x_d, y_d denote coordinates in the detector plane and m and M are the source and object magnification respectively. K is a constant, $s(x, y)$ denotes the x-ray source distribution, and $t(x, y)$ is the spatial transmission function of the object (whose thickness is ignored). Assume the x-ray source can be modeled as a 1-D Gaussian distribution as

$$s(x, y) = S_0 e^{-x^2} \delta(y),$$

where S_0 is a constant.

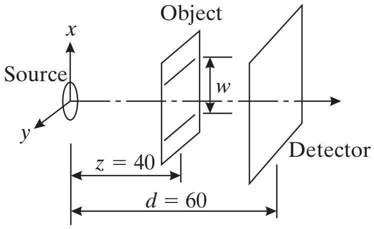


Figure P5.5 Diagram for Problem 5.23.

(**a**) Compute the object magnification M.

(**b**) Compute the source magnification m.

An ideal line phantom is to be imaged by the system, whose transmission function is given by

$$t(x, y) = \delta(x - \frac{w}{2}) + \delta(x + \frac{w}{2}).$$

(c) Determine the image of the phantom on the detector plane.

(d) Determine the minimal value of w such that the images of the two lines on the detector plane can be distinguished from each other.

5.24 A slab of soft tissue with one blood vessel running in the middle is imaged using an x-ray imaging system, as shown in Figure P5.6. For ease of computation, assume the tissue and the vessel both have square-shaped cross sections; the dimensions are shown in the figure. Assume that the x-ray source produces $N_i = 4 \times 10^6$ photons at either 15 keV or 40 keV, and the photons are uniformly shed upon the side of the tissue.

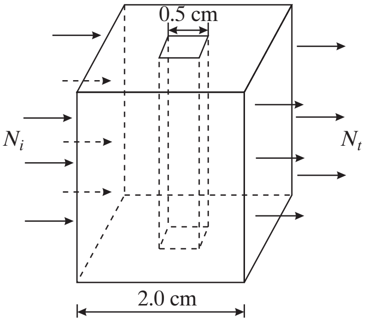


Figure P5.6
Figure for Problem 5.24.

The linear attenuation coefficients μ of the soft tissue, blood, and a radiographic contrast agent at two energy levels are given in the table below.

	Linear Attenuation Coefficient (cm^{-1})		
Energy(keV)	Soft Tissue	Blood	Contrast Agent
15	4.0	3.0	5.0
40	0.4	0.2	20

Ignoring photon noise, detector efficiency, beam divergence, scattering, path length, and magnification effects:

(a) Determine the total number of exiting photons at the two energy levels respectively. At which energy level are more photons absorbed?

(b) Calculate the local contrast of the blood vessel (target) at each energy level. At which energy level is the local contrast of the blood vessel better?

(c) Suppose the contrast agent whose linear attenuation coefficient is given in the table were injected into the blood vessel. Would you expect there to be much difference in the local contrast at 15 keV after injection? How about at 40 keV? (No calculations are needed; just use mathematical reasoning.)

(d) Explain why it is expected that the linear attenuation coefficient of the contrast agent is much larger at 40 keV than at 15 keV?

5.25 An x-ray imaging system is shown in Figure P5.7 to image a bar phantom. The detector is placed on the $z = 0$ plane. The bar phantom with two dark bars is placed on the $z = z_0$ plane, and the x-ray source is placed on the $z = 3z_0$ plane. The two dark bars on the phantom have widths w and are separated by a distance w. The x-ray source fires monochromatic x-ray beams that are uniformly shed upon the phantom.

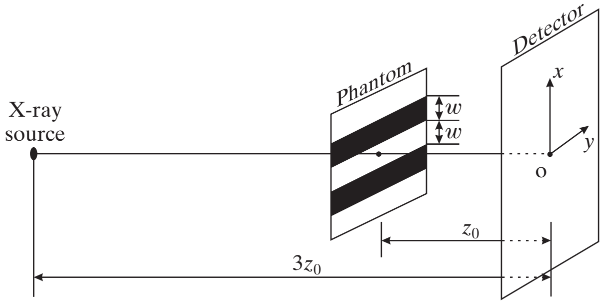


Figure P5.7
A medical imaging system and a bar phantom. See Problem 5.25.

Suppose the dark bars on the phantom absorb 75% of the photons passing through the phantom, and the white bars let all photons go through. Ignore scattering, the inverse square law, obliquity, and image noise.

(**a**) Assume an ideal point source and ignore the thickness of the phantom. Sketch the intensity profile on the detector as a function of position x for $y = 0$. Assume the intensity is 1 at $x = 0$, $y = 0$. Carefully label the axes.

(**b**) What is the contrast of the image on the detector? Assume the dark bars are the targets.

(**c**) The image on the detector will be digitized. Suppose the film/screen combination blurs the image by the following point spread function prior to point sampling

$$h(x, y) = \operatorname{sinc}(\alpha x)\operatorname{sinc}(\beta y),\ \alpha > 0,\ \beta > 0\,.$$

What are the largest sampling periods (in x and y directions) that will give an aliasing-free sampling?

(**d**) Now assume the phantom is a slab of a certain thickness. The dark bars attenuate the x-ray photons by the following linear attenuation coefficient

$$\mu(E) = \ln\left(\frac{640\ \text{keV}}{E}\right)\ \text{cm}^{-1}, \quad 100\ \text{keV} \le E \le 160\ \text{keV}\,,$$

which is a function of energy, E. The x-ray source fires N photons at $E = 160$ keV per unit area during a burst of a unit time. How thick is the phantom such that the dark bars attenuate 75% of the photons?

5.26 Consider an x-ray imaging system using a film-screen detector. Due to the limited dynamic range of film, an essentially constant x-ray intensity is required for the film to be properly exposed. Assume for simplicity that this means that a constant number of photons per unit area is incident on the image detector.

(a) Assuming all other factors are constant, if the x-ray tube kilovolts were increased and the exposure time was appropriately adjusted to produce an optimal film exposure, what would be the effect on SNR? Explain.

(b) Assuming the same conditions as in (a), what would be the effect on the patient dose? Explain.

(c) For each of the following changes, indicate the effect on the subject contrast using this notation: I=increase, D=decrease, N=no effect:

- Increase in patient thickness
- Reduction in kilovoltage
- Increase in detector efficiency
- Reduction in the x-ray field size
- Use of a high atomic number contrast media

(d) Under what conditions is an airgap (space between the patient and the image detector) most effective in reducing the scatter fraction?

(e) What ultimately limits the contrast sensitivity of an x-ray imaging system: geometric unsharpness, the display gamma, the image noise, or the scatter fraction?

(f) The attenuation of a polychromatic x-ray spectrum can be approximately determined using its average photon energy. For a given tube, the average photon energy is mainly determined by: the beam current, the Fourier transform of the x-ray spectrum, the material in the beam path, or the scatter fraction?

5.27 Consider the x-ray projection radiography system shown in Figure P5.8. Relevant dimensions in the figure are $L = 1$ m, $D = 4$ m, $w = 1$ cm, $h = 3$ cm, $R = 0.1$ cm, and $D_{td} = 10$ cm.

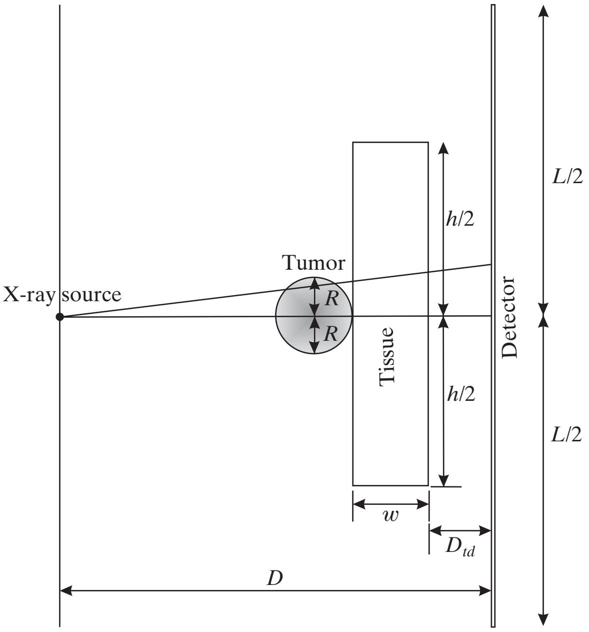


Figure P5.8 An x-ray projection radiography system to image a tumor. (Figure is not to scale.) See Problem 5.27.

A contrast agent is used to enhance the image of the tumor. Assume a 35 keV (monochromatic) x-ray source and linear attenuation coefficients given in the table below.

	Linear Attenuation Coefficient (cm^{-1})		
Energy(keV)	Tissue	Tumor	Tumor with contrast agent
35	1	0.75	10

(a) The K-shell energies of iodine and barium are 33.2 keV and 37.4 keV, respectively. Assuming that either agent could be made into a compound that would go to the tumor, what would be the best agent to use and why?

(b) Let the local contrast of the tumor be defined as $C = (I_t - I_b)/I_b$, where I_t is the intensity at the center of the tumor and I_b is the intensity just outside the edge of the tumor. Neglecting Compton scatter, find both the local contrast C_1 of the tumor before contrast is applied and the local contrast C_2 of the tumor after contrast is applied.

In the next parts, assume that Compton scattering events are taking place throughout the tissue and, for simplicity, assume that $w \approx 0$.

(c) Make a sketch of the photon path that yields a photon hitting the detector with the lowest possible energy. (Assume that only single-scatter Compton events can occur.)

(d) Find the energy of the Compton scattered photon in part (c) (that has the lowest possible energy to be detected given the assumed imaging geometry).

CHAPTER

6

Computed Tomography

6.1 Introduction

A *tomogram* is an image of a plane or slice within the body. A conventional radiograph is 2-D, but it is not a tomogram because its data do not arise from a single plane. X-ray computed tomography (CT) is extremely popular because it generates images of slices of the body, thereby eliminating the artifacts of overlaying structures that dominate conventional radiographs. A typical CT scanner is shown in Figure 6.1. One way to think about the basic mechanism of CT is to imagine taking a series of conventional chest x-rays, where the patient is rotated slightly around an axis running from head to foot between each exposure. Each developed film contains a 2-D projection of the 3-D body, but because they are each exposed at a different angle, there is different information in each one. In fact, each horizontal line within a film gives a 1-D projection of a 2-D axial cross section of the body at that angle. Thus, a collection of horizontal lines, one from each film at the same height, contains information about only one axial cross section. These *projection data* are then used to *reconstruct* cross-sectional images—tomograms.

What can we say about the tissues in an axial cross section from a collection of 1-D projections taken at different angles? Fortunately, we can say a great deal, and this is the basic premise of CT. The transformation that takes 1-D projections of a 2-D object over many angles is called the *2-D Radon transform*, and it has a mathematical inverse. That is, given a collection of projections we can, in theory, reconstruct the axial cross section exactly. Practical implementation of the *inverse 2-D Radon transform* is accomplished in every CT scanner in one fashion or another. Therefore, after studying how a CT scanner acquires these projections (after all, taking a collection of chest x-rays is hardly practical), we will study in detail the 2-D Radon transform and its theoretical and practical inverse.

The advent of x-ray CT in the mid-1960s and its development all the way up to the present day has produced a profound change in the role of diagnostic imaging in medicine. A wide variety of medical conditions is visible in CT images that are not visible in conventional radiographs. Thus, diagnosis and monitoring

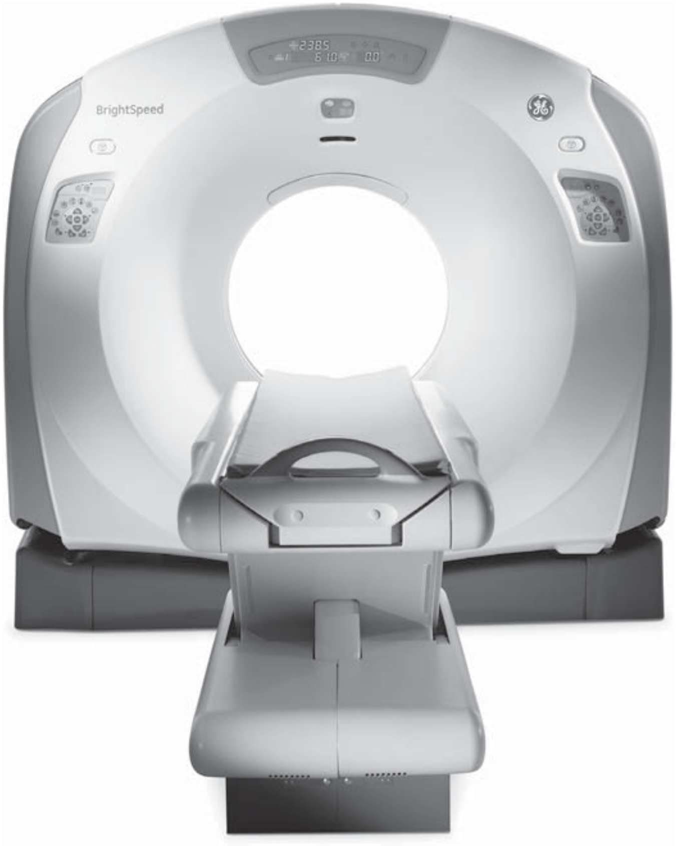


Figure 6.1
A typical CT scanner. (Courtesy of GE Healthcare.)

of many diseases are possible through CT alone; this has reduced the amount of exploratory surgery previously required for these diseases. As a measuring device, the x-ray CT scanner has also come a long way since its origins. Now, through standard calibration procedures, CT scanners measure *CT numbers* in *Hounsfield units*, which are constant from scan-to-scan and across different scanners. Therefore, quantitative analysis is possible and is in common practice today.

Because of the growing importance of CT during the early years, manufacturers put considerable effort toward lowering both the scanning and image reconstruction times. Accordingly, the time required to scan a patient and generate an image has dropped from several hours with the earliest clinical machines to under 100 milliseconds for some current machines. Multiple cross sections can now be collected rapidly using *helical CT* and *multiple-row detector CT* (MDCT); hence, 3-D data sets are readily available. A very important area of current research and development in both academic and industrial arenas involves the display and analysis of 3-D CT data sets.

In this chapter, we present the fundamentals of x-ray CT imaging. Since the properties of x-rays—their generation, detection, and interaction with the human body—have been discussed in Chapter 5, the focus of this chapter is on the various geometries of CT scanners and on a mathematical description of the algorithms necessary to reconstruct images. The mathematical abstraction is necessary since it leads to the actual methods, implemented in software and hardware, used in modern-day clinical scanners. As well, the same mathematics and methods can be used in single photon emission computed tomography (SPECT), positron emission tomography (PET), and magnetic resonance imaging (MRI).

6.2 CT Instrumentation

The fundamental measurement required by a CT scanner is the measurement of x-ray attenuation along a line between an x-ray source and an x-ray detector. In order to reconstruct an image of a 2-D cross section, a collection of such measurements are required along all lines within the cross section. Seven *generations* of basic CT designs have been developed to obtain these fundamental data. Roughly speaking, the seven generations were developed in sequence to improve overall performance, mostly toward faster acquisition of both cross-sectional and volumetric data. Although first- and second-generation scanners are no longer used in medical imaging, it is still important to understand their geometry. In particular, the parallel ray geometry of first-generation scanners is used to explain the reconstruction formulas that underlie all CT scanners. As we describe the CT generations, it may be helpful to refer to Table 6.1, which compares various properties of CT scanners across the generations.

6.2.1 CT Generations

First-generation (1G) scanners are no longer manufactured for medical imaging, but their geometry is still used to explain the theoretical ideas underlying image reconstruction. The geometry of a 1G scanner is shown in Figure 6.2. It consists of a single source, collimated (meaning that its beam is restricted) to a thin line, and a single detector that move in unison along a linear path tangent to a circle that contains the patient. After making a linear scan, the source and detector apparatus are rotated so that a linear scan at a different angle can be made, and so on. The 1G geometry has the advantage that an arbitrary number of *rays*—that is, paths between source and detector—can be measured within a given projection, and an arbitrary number of angular projections may be measured. It has the further advantage that scattered radiation goes mostly undetected (because it misses the detector), and therefore, the measured attenuation of the beam is almost certainly due to tissues along the ray.

A *second-generation (2G) scanner*, as illustrated in Figure 6.3, has additional detectors, forming a detector array, arranged along a line or a circle. As in the 1G scanner, the source and detector array move linearly in unison to cover the field of view (FOV). However, unlike the 1G scanner, the source in a 2G scanner is collimated as a fan beam so that energy is kept within the slice but spread over the detector array. Therefore, while the source and central detector scan the same projection as the 1-D projection, the additional detectors are simultaneously obtaining additional projections from different angles.

With the 2G scanner geometry, we can make a larger rotation after each linear scan and thereby complete a full scan in less time, making the 2G scanner faster. Scattering effects must be considered, however, in the design of

TABLE 6.1

Comparison of CT Generations

Generation	Source	Source Collimation	Detector	Detector Collimation	Source–Detector Movement	Advantages	Disadvantages
1G	Single x-ray tube	Pencil beam	Single	None	Move linearly and rotate in unison	Scattered energy is undetected	Slow
2G	Single x-ray tube	Fan beam, not enough to cover FOV	Multiple	Collimated to source direction	Move linearly and rotate in unison	Faster than 1G	Lower efficiency and larger noise because of the collimation in detectors
3G	Single x-ray tube	Fan beam, enough to cover FOV	Many	Collimated to source direction	Rotate in synchrony	Faster than 2G, continuous rotation using a slip ring	More expensive than 2G, low efficiency
4G	Single x-ray tube	Fan beam covers the FOV	Stationary ring of detectors	Cannot collimate detectors	Detectors are fixed, source rotates	Higher efficiency than 3G	High scattering since detectors are not collimated
5G (EBCT)	Many tungsten anodes in single large tube	Fan beam	Stationary ring of detectors	Cannot collimate detectors	No moving parts	Extremely fast, capable of stop-action imaging of beating heart	High cost, difficult to calibrate
6G (Helical CT)	3G/4G	3G/4G	3G/4G	3G/4G	3G/4G plus linear patient table motion	Fast 3D images	A bit more expensive
7G (Multiple-row detector CT)	Single x-ray tube	Cone beam	Multiple arrays of detectors	Collimated to source direction	3G/4G/6G motion	Fast 3D images	Expensive

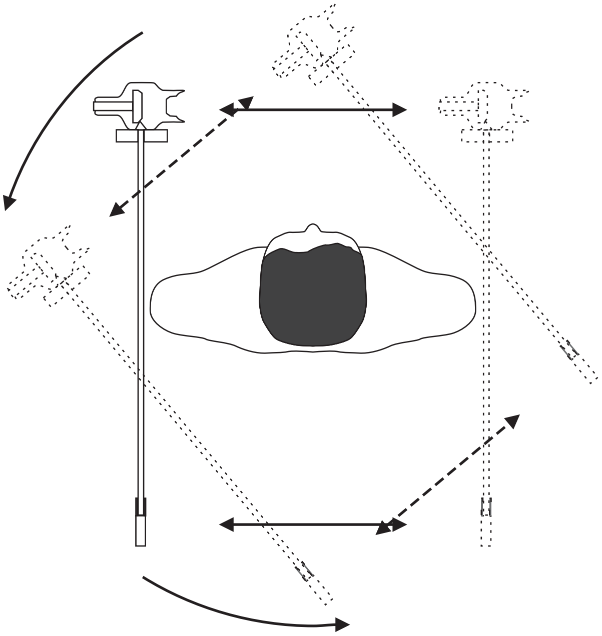


Figure 6.2
The geometry of a first-generation (1G) CT scanner.

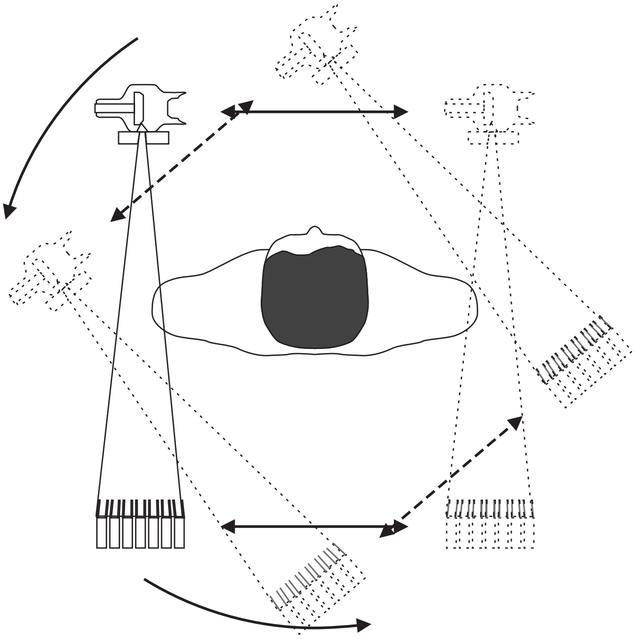


Figure 6.3
The geometry of a second-generation (2G) CT scanner.

a 2G scanner, since radiation scattered from one ray could end up in another detector. Because of this phenomenon, the detectors in a 2G scanner are usually collimated so that they can receive radiation only from the correct direction. Collimation on detection, however, reduces efficiency and increases noise for a given dose. Hence, for a given dose and with an identical number of projections and measurements per projection, the 1G scanner will produce a better image.

Still, the payoff in scan time is so great that slightly increasing the dose to counter the deleterious effects from collimation is worth it.

EXAMPLE 6.1

Consider a 1G or 2G scanner whose source–detector apparatus can move linearly at a speed of 1.0 m/s and that the FOV has a diameter of 0.5 m. Suppose further that 360 projections over 180° are required and that it takes 0.5 s for the source–detector apparatus to rotate one angular increment, regardless of the angle.

Question What is the scan time for a 1G scanner? What is the scan time for a 2G scanner having nine detectors spaced 0.5° apart?

Answer It takes $0.5/1.0 = 0.5$ s to measure one projection. So the time used for making measurements is 360×0.5 s $= 180$ s, and 360 rotations are needed, which take 360×0.5 s $= 180$ s. So the total scan time is $180 + 180 = 360$ s, or 6 minutes.

For the 2G scanner, nine projections are acquired simultaneously. (The required angular increment is $180°/360 = 0.5°$, which agrees with the detector separation.) Therefore, only $360/9 = 40$ rotations are required, which takes 40×0.5 s $= 20$ s. Ignoring the small amount of overscanning required to cover the fan angle itself, the time for each linear scan is still 0.5 s. Therefore, the total scan time is $40 \times 0.5 + 20 = 40$ s. ■

The geometry of a *third-generation (3G) scanner*, as shown in Figure 6.4, has a fan-beam that covers the image region with the source held in a single position. Therefore, the source and detector array need not perform a linear scan; instead they simply rotate in synchrony. To obtain a sufficient number of samples per (fan-beam) projection, there must be a large number of detectors and this pushes the cost of a 3G scanner beyond that of a 2G scanner. The simple

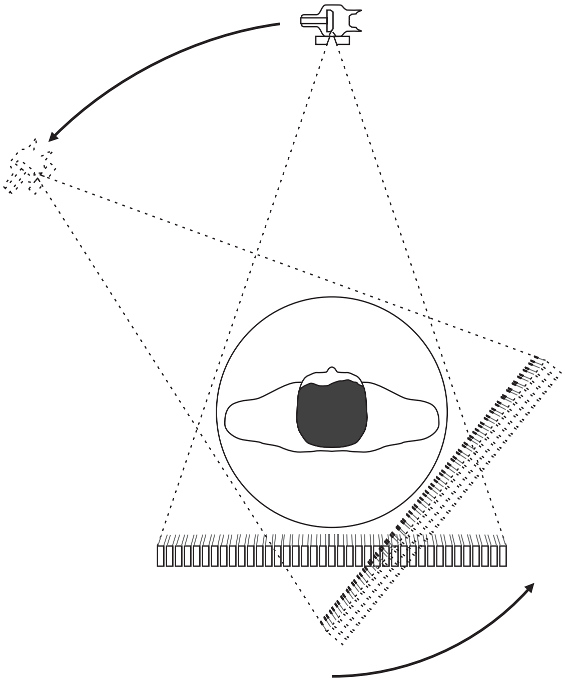


Figure 6.4
The geometry of a third-generation (3G) CT scanner.

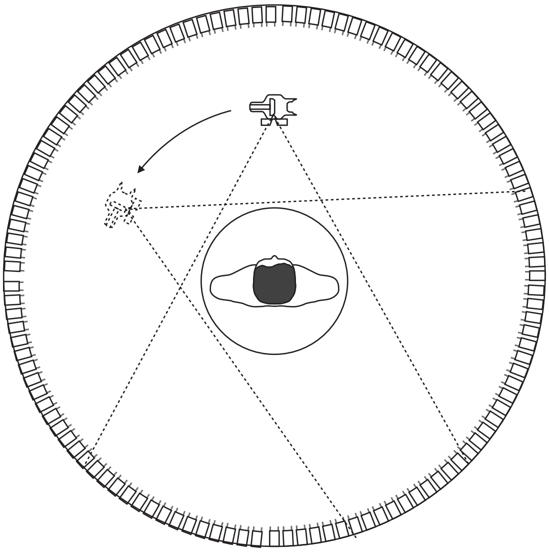


Figure 6.5
The geometry of a fourth-generation (4G) CT scanner.

rotational motion and highly parallel detection capability, however, allows for a dramatic decrease in scan time. A typical scanner acquires 1,000 projections with a fan-beam angle of 30–60 degrees incident upon 500–700 detectors and does this in 1–20 seconds. As in 2G systems, the detectors of a 3G system are collimated to reduce scattering artifacts, so there is some loss of detector collection efficiency. Furthermore, since the detectors must be very small, they cannot be as efficient as those in 1G and 2G systems. Therefore, once again, to obtain the same image quality with a 3G scanner as that obtainable in 1G and 2G systems, the dose must be slightly greater.

A *fourth-generation (4G) scanner*, as shown in Figure 6.5, has a single rotating source with a larger ring of stationary detectors. A variation on this theme has the source outside the detectors with slight gaps between the detectors through which the x-rays can be fired. Collimation cannot be used in this geometry since a detector must receive energy from a source that moves through many positions. Because the detectors are not collimated and because they can be physically large since they lie on a large ring, the detection efficiency of a 4G scanner is higher than that of a 3G scanner. However, because the detectors cannot be collimated, they are also highly susceptible to scattering. This factor is so significant that the image quality in 4G systems is only comparable to that of 3G systems, not better.

One major design feature shared by the first four generations of CT scanners is the presence of only one x-ray tube. This feature is primarily due to the fact that x-ray tubes are expensive and bulky and that they must be constantly calibrated. The need to rotate a bulky object—x-ray tube and/or detectors—limits the overall imaging speed of the scanner. *Fifth-generation (5G) scanners* are designed to solve this problem. Fifth-generation CT scanners, also known as *electron beam computed tomography* (EBCT) scanners, are commercially available at this time but are not in widespread clinical use because of their high cost. EBCT uses a flying electron beam, steered electromagnetically, to hit one of four tungsten anode strips that encircle the patient. X-rays are generated when the electron beam strikes the tungsten anode; the resultant radiation is collimated into a fan-beam, which passes through the patient and is detected on the other side by a stationary ring of detectors, as in 4G CT systems. Since the anodes and detectors are stationary, no moving parts are required, and this allows a full set of fan-beam projection data to be acquired in about 50 milliseconds. EBCT is an expensive design, but because of the extremely small scan-time it is the only commercially available CT method that can capture stop-action images of a beating heart without electrocardiographic (ECG) gating.

Sometimes considered the *sixth-generation (6G) scanner*, the *helical CT* was developed in the late 1980s to address the need for rapid volumetric data acquisition. Today, most of the CT scanners in the United States are capable of helical CT. The reason is simple: With helical CT, a full 60-cm torso scan can be obtained in about 30 seconds, a full 24-cm lung study in 12 seconds, and a detailed 15-cm angiography study in 30 seconds. With these kinds of speeds, motion artifacts can be reduced or eliminated (using breath-holds), and critically ill patients can be scanned quickly. For entry-level scanners, the equipment may cost about US$500,000 for a helical CT unit versus about US$350,000 for a conventional CT unit. A top-of-the-line helical CT unit may reach the US $1 million mark.

A helical CT scanner consists of a conventional arrangement of the x-ray source and detectors (as in 3G and 4G systems) which can continuously rotate. While the tube is rotating and acquiring projection data, the patient table is set into motion, sliding the patient through the source–detector plane. It is easy to see that with this geometry and movement, the position of the source carves out a helix with respect to the patient. Continuous rotation of the large mass comprising the x-ray source and detectors requires what is called *slip ring* technology in order to communicate with the controlling stationary hardware. In particular, power is provided using rings and brushes, while data are passed using optical links. Rotation periods are typically 0.3 to 0.5 seconds per revolution in modern scanners.

A *seventh-generation (7G) scanner* has emerged with the advent of *multiple-row detector* CT (MDCT) scanners. In these scanners, a “thick” fan-beam is used, and multiple (axial) parallel rows of detectors are used to collect the x-rays within this thick fan. (Some scanners have fan beams that are so thick they can be thought of as cone beams.) Essentially, multiple (up to 320) 1-D projections are collected at the same time using this geometry. This leads to more economical use of the x-ray tube since less of the generated beam must be absorbed by lead at the source. Also, provided that the x-ray tube and detectors are far enough apart, the simultaneous data that are collected from the parallel rows of detectors correspond approximately to parallel planes. This means that a “slab” of cross-sectional images can be imaged simultaneously. When a multiple detector array is combined with a helical scanner, the pitch of the helix can be larger, and full 3-D scans can be even faster.

The advent of helical and MDCT has made the requirement for new developments in data processing even more critical. In particular, while a conventional CT might have reconstructed 40 slices over a region of interest, with helical CT and MDCT a clinician might acquire 80–120 slices over the same region in less time. Some extensive studies may comprise 500 slices—for example, a 50 cm scan at 1 mm intervals might be used for imaging the colon. The burden is now to medically interpret or “read” these slices, which takes time. A study was conducted in which helical CT images were used to detect liver lesions. In this study, it was found that reviewers could detect 86–92% of the lesions using all the data, but only 82–85% of the lesions using every other slice. Most clinicians are looking to image processing and computer vision techniques—for example, virtual reality—to reduce the amount of time it takes to read this enormous volume of data. At the same time, it is believed that considerably improved diagnostic accuracy can be achieved.

6.2.2 X-Ray Source and Collimation

Most commercial medical CT scanners use just one x-ray source due to the initial cost and maintenance costs associated with x-ray tubes. In most respects, the x-ray tubes in CT scanners are just like those in projection radiography systems. All use the rotating anode design and are oil-cooled to reduce heat damage to the anode. Some CT scanners operate in pulse mode but most continuously excite the x-ray tube during data acquisition. In most cases, cool-down periods are required between scans to avoid heat damage to the x-ray tube. Even so, with 20 to 30 exposures per examination and 20–30 examinations per day, the x-ray tubes in CT scanners often wear out in less than one year. Maintaining the x-ray tubes through regular calibration and replacing them quite often adds considerably to the cost of operating a CT scanner.

The x-rays generated by a CT x-ray tube require collimation and filtration that is somewhat different than that in projection radiography systems. For one, CT typically requires a fan-beam geometry rather than the cone-beam geometry used in projection radiography. Collimation into a fan, typically between 30 and 60 degrees in fan angle, is accomplished using two pieces of lead that form a slit between them. This collimator is placed as close to the patient as possible, just outside a protective plexiglass tube so that the geometry of the fan has the most constant thickness possible through the patient. A motor controls how wide the slit is so that the fan thickness (or height)—typically 1–10 mm—can be selected by the operator. In MDCT systems, the thickness of the fan is typically larger than in single-slice systems in order to cover the width (perhaps 4–16 cm) of the multiple detector rows, but otherwise the principle is the same. The x-rays pass through a mirror, which can be optically illuminated from below in order to create lines on the patient showing the position and thickness of the x-ray collimation to be used during the scan.

As we will see in Section 6.3, CT image formation requires the x-ray source to be approximated as a monoenergetic source. In order to make the actual CT x-ray source more monoenergetic in practice, more filtering is employed than is typical in projection radiography systems. Generally, copper followed by aluminum is used to narrow the energy spectrum ("harden the beam") of the x-rays entering the patient. Such filtration was discussed in Section 5.2.2.

6.2.3 Dual-Energy CT

Dual-energy imaging is a way to gather more information about tissue characteristics since the linear attenuation of all tissues is a function of energy. This strategy is used in dual-energy x-ray absorptiometry projection scanning to determine bone mineral density in the diagnosis of osteoporosis and is now being introduced in CT for a variety of clinical applications. In the past, dual-energy CT data would be acquired by either scanning the subject twice using different kVp's—for example, 80 kVp and 140 kVp—or by using a pulsed mode that excites the tube at alternating kVp's during a single scan. Despite the clinical utility, these approaches were not very popular due to the increased scan time and other technical challenges.

With the recent introduction of *dual source CT* (i.e., scanners with two x-ray tubes), the situation has changed, and many scanners that are capable of both rapid scanning at a single-energy and simultaneous dual-energy acquisition are being introduced into routine clinical use. Since two separate images are acquired, clinicians are able to study the images separately as in a dual-modality study like PET/CT or combine them into single image or collection of images that reveal different underlying tissue properties based on their linear attenuation coefficients at the two effective energies.

6.2.4 CT Detectors

Most modern scanners use *solid-state detectors*, as shown in Figure 6.6(a). These detectors contain a *scintillation crystal* in the first stage, typically a cadmium tungstate, sodium iodide, bismuth germanate, yttrium-based, or cesium iodide crystal. X-rays interact with the crystal mainly by photoelectric effect, producing photoelectrons, similar to what happens to the phosphor in an intensifying screen. These electrons are excited and emit visible light when they spontaneously de-excite.

This *scintillation* process results in a burst of light. The light is then converted to electric current using a solid-state photodiode that is attached tightly to the scintillator. (Some 4G and 5G scanners use photomultiplier tubes to convert light to electricity; see Section 8.2.3 for a description of such tubes.) For 1G, 2G, and 3G systems, which use line collimators in front of the detectors, the efficiency of each detector may be increased by using thick crystals. However, thinner crystals must be used in 4G systems to allow the detectors to detect off-axis incident photons (as is required by the scanner geometry); unfortunately, thinner crystals have lower efficiency.

In 3G scanners, very small and highly directional detectors are required. Either solid-state detectors, as described above, or xenon gas detectors, as shown in Figure 6.6(b), can be designed to satisfy these requirements. Xenon gas detectors use compressed xenon gas in long, thin tubes, which when ionized generate a current between an anode and cathode (maintained at a high potential difference). These detectors are typically less efficient than solid-state detectors,

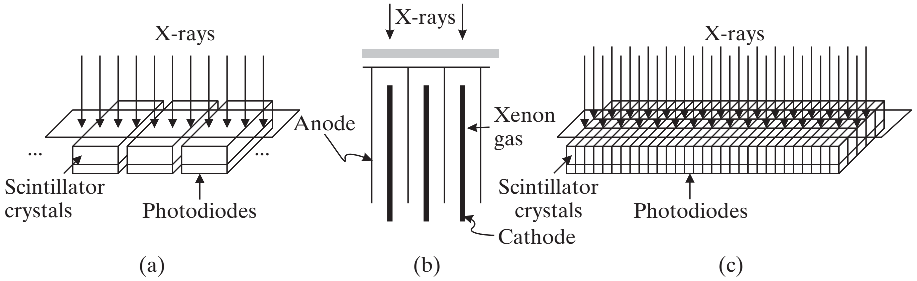


Figure 6.6
(a) Solid-state detectors, (b) xenon gas detectors, and (c) multiple (solid-state) detector array.

but they are highly directional, and therefore appropriate for the geometry of 3G scanners. Solid-state detectors must be accompanied by external collimation in order to provide necessary directivity (i.e., scatter rejection). In order to provide both good resolution and high efficiency, in a single-slice scanner (single-row detector CT or SDCT) each crystal would typically be 1.0 mm × 15 mm in area.

An important concept related to detector dimensions is *slice thickness*. In a typical CT scanner, the axial response (along the z-direction) is a rect function whose width is the slice thickness. The maximum slice thickness is equal to the detector height, 15 mm in the typical SDCT case. Smaller slice thicknesses are created by further restricting the beam using movable blades in the x-ray tube collimator. Slice thickness is controlled at the console and is typically set in millimeter increments, starting from 1 mm.

In a MDCT systems, as shown in Figure 6.6(c), the individual solid-state detectors are typically 1.0 mm × 1.25 mm in size. In these systems, slice thickness is controlled by both the axial beam width and the combined detector height. In particular, MDCT systems have digital switching arrays that can group detectors together in order to synthesize detector widths in multiples of the physical detector heights. Accordingly, slice thicknesses are selected in multiples of the detector height and the beam width is chosen to cover the desired axial acquisition. For example, if there are 16 rows of detectors, each 1.25 mm in height, we could acquire 16 simultaneous projections, each 1.25 mm thick. Alternatively, we could acquire eight simultaneous projections, each 2.5 mm thick, and so on. We will see later that there is a tradeoff in image quality that must be addressed in selecting slice thickness, so it is not always a good idea to simply acquire the thinnest possible slices.

6.2.5 Gantry, Slip Ring, and Patient Table

The *gantry* of a CT system holds the x-ray tube and detectors so that they can be rotated around the patient rapidly and repeatably. The fan angle, size of the detector array, and separation between the x-ray tube and detector array must be capable of imaging a 50 cm (typical) FOV. A full 2-D scan in under 1 second is required. The gantry can typically be tilted in order to facilitate acquisition of nonaxial slices.

Most modern scanners are capable of continuous rotation without requiring a "rewinding" of the source and detectors. Since there are large voltages that must go to the x-ray tube and hundreds of signals that must be transmitted out of the detectors, continuous rotation makes special demands on the mechanical and electrical design, which the *slip ring* solves. A slip ring contains a large cylinder with grooves on the outside so that brushes can make continuous electrical contact with the rotating cylinder. The x-ray tube and detectors are mounted inside the cylinder and are in continuous electrical contact with the (stationary) controlling and data processing electronic hardware.

The patient table is more than just a place to put the patient. In helical scanners, it is an integral part of the data acquisition hardware, since it must be moved smoothly and precisely in synchrony with the source and detector rotation. Even in single-slice scanners, the table's positioning capabilities must be quite flexible. Typically, the table can be extended well out from the gantry and lowered in order to facilitate patient transfer from a hospital bed. It must be capable of “docking” with the scanner and locked into place so that motor-controlled positioning of the patient can be accomplished. Typically, there are positioning lights that can illuminate the patient to show the technician where the slice(s) will be acquired. The table can then be moved to position the patient under the guide lights in order to acquire the desired image(s).

6.3 Image Formation

6.3.1 Line Integrals

In a CT system, the x-ray tube makes a short burst of x-rays that propagate (generally in a fan-beam geometry) through a cross section of the patient. The detectors detect the exit beam intensity integrated along a line between the x-ray source and each detector. As presented in Chapter 5 [see (5.3)], the (integrated) x-ray intensity at any given detector is given by

$$I_d = \int_0^{E_{\max}} S_0(E)E \exp\left[-\int_0^d \mu(s; E)ds\right] dE\,, \tag{6.1}$$

where $S_0(E)$ is the x-ray spectrum and $\mu(s; E)$ is the linear attenuation coefficient along the line between the source and detector.

Unfortunately, the integration over energy in (6.1), while physically correct, is in a mathematically intractable form for CT image reconstruction (to be discussed in Section 6.3.3). To get around this, we use the concept of *effective energy*, $\overline{E}$, which is defined as the energy that, in a given material, will produce the same measured intensity from a monoenergetic source as is measured using the actual polyenergetic source. This concept was first presented in Example 4.4. Given this concept, it is correct to say that

$$I_d = I_0 \exp\left[-\int_0^d \mu(s; \overline{E})\, ds\right]. \tag{6.2}$$

Given a measurement of I_d and knowledge of I_0, (6.2) can be rearranged to yield a basic projection measurement g_d,

$$g_d = -\ln\left(\frac{I_d}{I_0}\right) \tag{6.3}$$

$$= \int_0^d \mu(s; \overline{E})\, ds\,. \tag{6.4}$$

Thus, we can make a very important observation about the function of a CT scanner: The basic measurement of a CT scanner is a line integral of the linear attenuation coefficient at the effective energy of the scanner.

In an actual CT system, the *reference intensity* I_0 must be measured for each detector; this is a calibration step. In fan-beam systems, an auxiliary measurement is typically made using a detector positioned at the end of the array so that there is always air between the source and this detector. Using this measurement, the reference intensities for the other detectors can be determined using prior calibration data of all detectors taken in air.

6.3.2 CT Numbers

A CT scanner reconstructs the value of μ at each pixel within a cross section, using a process that we will describe in detail in the following sections. Different CT scanners, however, have different x-ray tubes, which in turn have different effective energies. Thus, the exact same object will produce different numerical values of μ on different scanners. Worse, since the x-ray tube on a busy CT scanner may need to be replaced about once every year, the same CT scanner will produce a different scan of the same object in successive years. This is clearly not a desirable situation.

In order to compare data from different scanners, which may have different x-ray sources and hence different effective energies, CT numbers are computed from the measured linear attenuation coefficients at each pixel. The *CT number* is defined as

$$h = 1,000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}}, \tag{6.5}$$

and is expressed in Hounsfield units (HU). Clearly, $h = 0$ HU for water, and since $\mu = 0$ in air, we find that $h = -1,000$ HU for air. The largest CT numbers typically found naturally in the body are for bone, where $h \approx 1,000$ HU for average bone, although CT numbers can surpass $h \approx 3,000$ HU for metal and contrast agents. Usually, CT numbers are rounded or truncated to the nearest integer; they are typically reproducible to about ±2 HU between scans and across scanners.

6.3.3 Parallel-Ray Reconstruction

We have learned that the basic CT measurement is a line integral of the effective linear attenuation coefficient within a cross section. But line integrals are not what we desire; we really want a picture of μ, or equivalently its CT number h, over the entire cross section. Therefore, the important question is this: Can we reconstruct a picture of μ given a collection of its line integrals? The answer is "yes," and in this section we explore the theory and practice of reconstruction from projections for parallel-ray geometry.

Geometry Let x and y be rectilinear coordinates in the plane. A line in the plane is given by

$$L(\ell, \theta) = \{(x, y) | x \cos\theta + y \sin\theta = \ell\}, \tag{6.6}$$

where ℓ is the lateral position of the line and θ is the angle of a unit normal to the line, as first introduced in (2.10) and shown in Figure 6.7. The line integral of function $f(x, y)$ is given by

$$g(\ell, \theta) = \int_{-\infty}^{\infty} f(x(s), y(s))\, ds, \tag{6.7}$$

where

$$x(s) = \ell \cos\theta - s \sin\theta, \tag{6.8}$$

$$y(s) = \ell \sin\theta + s \cos\theta. \tag{6.9}$$

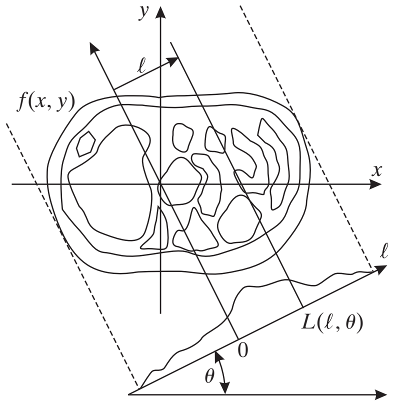


Figure 6.7
The geometry of lines and projections.

This is a parametric form of the line integral; an alternate expression is given by

$$g(\ell,\theta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta+y\sin\theta-\ell)\,dx\,dy\,. \tag{6.10}$$

Here, the sifting property of the 1-D impulse function $\delta(\cdot)$ [see (2.6)] causes the integrand to be zero everywhere except on the line $L(\ell,\theta)$. The integral acts on the impulse function by integrating the values of $f(x,y)$ only along the line—hence, it takes a line integral.

For a fixed θ, $g(\ell,\theta)$ is called a *projection*; for all ℓ and θ, $g(\ell,\theta)$ is called the *2-D Radon transform* of $f(x,y)$. The relationship of a projection to the object $f(x,y)$ is shown in Figure 6.7. If we make the identifications

$$f(x,y)=\mu(x,y;\overline{E})\,, \tag{6.11}$$

$$g(\ell,\theta)=-\ln\left(\frac{I_d}{I_0}\right), \tag{6.12}$$

we see that this mathematical abstraction exactly characterizes the CT measurement situation. In what follows, we assume that $g(\ell,\theta)$ corresponds to the measurements and $f(x,y)$ corresponds to the underlying unknown function or object that we wish to reconstruct. Notice that our definition of a projection corresponds to a collection of line integrals for parallel lines. Hence, these are called *parallel-ray projections*, and they correspond in geometry to 1G CT scanners only. We will find in Section 6.3.4, however, that the formalism and methods developed in this section lead directly to a reconstruction approach for fan-beam projections.

EXAMPLE 6.2

Consider the unit disk given by

$$f(x,y)=\begin{cases}1 & x^2+y^2\le 1\\ 0 & \text{otherwise}\end{cases}\,.$$

Question What is its 2-D Radon transform?

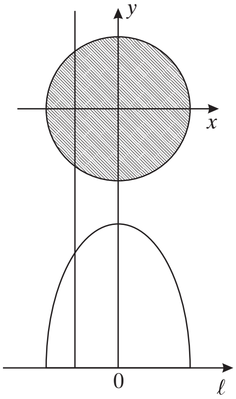


Figure 6.8
Projection of a disk.

Answer This function is circularly symmetric so its projections are independent of the angle. It is therefore sufficient to calculate the projection at $\theta = 0^\circ$ where the lateral displacement ℓ is horizontal and line integrals are vertical. (See Figure 6.8.)

Accordingly, we have

$$g(\ell, \theta) = \int_{-\infty}^{\infty} f(\ell, y)\, dy\,.$$

Here, $g(\ell, \theta) = 0$ when $|\ell| > 1$. When $|\ell| \le 1$, the integral must integrate the function value on the unit disk, just the number 1, from the bottom of the unit circle to the top. Accordingly, in this range we can write

$$g(\ell, \theta) = \int_{-\sqrt{1-\ell^2}}^{\sqrt{1-\ell^2}} dy\,.$$

Performing the integration and supplying the known value outside this range yields

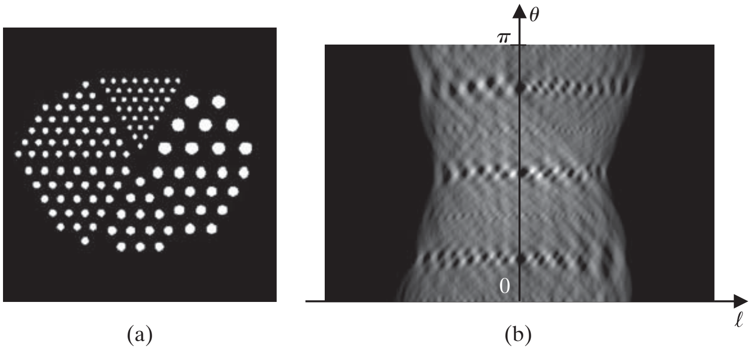
$$g(\ell, \theta) = \begin{cases} 2\sqrt{1-\ell^2} & |\ell| \le 1 \\ 0 & \text{otherwise} \end{cases}\,,$$

which is the desired result. ■

Sinogram An image of $g(\ell, \theta)$ with ℓ and θ as rectilinear coordinates is called a *sinogram*. It is a pictorial representation of the Radon transform of $f(x, y)$ and represents the data that are necessary to reconstruct $f(x, y)$.

An object and its sinogram are shown in Figure 6.9. It is important to observe several features about this sinogram. First, by convention in this book, the bottommost row of the sinogram corresponds to the projection of the object at $\theta = 0$. With reference to Figure 6.7, we see that this is the projection comprising vertical integrals of the object. Since the object in Figure 6.9(a) is widest in the horizontal direction, this gives the widest projection, which is evident in Figure 6.9(b).

Moving up from the bottommost row in Figure 6.9(b), the projections are increasingly narrower and become narrowest at the angle $\theta = \pi/2$. This corresponds to the projection comprising horizontal integrals. The topmost row of the sinogram corresponds to an angle just shy of $\theta = \pi$, returning nearly



(a) (b)

Figure 6.9 (a) An object and (b) its sinogram.

to a projection comprising vertical lines. There is no need to proceed further in angle, since these projections are redundant.

There are other interesting features in Figure 6.9(b). It is possible to see sweeping features that run vertically through the sinogram. These features are actually sinusoidal in character, and if the sinogram were "turned on its side," they would be described by sinusoidal functions. This is the origin of the name *sinogram*. It is straightforward to show that these features correspond to individual objects in the image; the Radon transform of small points in the image corresponds to (vertical) sinusoids in the sinogram. There are also accumulations of bright white spots in certain rows of the sinogram in Figure 6.9(b). Upon careful inspection of the image in Figure 6.9(a), it is apparent that these bright spots correspond to the alignment of bright spots on parallel lines in the projection at these specific angles. In other words, the regular alignment of features in the image along certain directions gives rise to periodic bright spots in the projections, which are revealed as organized features in the sinogram.

Although the sinogram is certainly understandable, it could never be argued that it represents an image of the object. Clearly, it is necessary to take the data represented in the sinogram and process them so that they look like the cross section from which they came. This is the process of *reconstruction*, which we now begin to develop.

Backprojection Let $g(\ell,\theta)$ be the 2-D Radon transform of $f(x,y)$, and consider the projection at $\theta=\theta_0$. In general, there are an infinite number of functions $f(x,y)$ that could give rise to this projection; hence, we cannot determine $f(x,y)$ uniquely from a single projection. Intuition tells us that if $g(\ell,\theta_0)$ takes on a large value at $\ell=\ell_0$, then $f(x,y)$ must be large over the line (or somewhere on the line) $L(\ell_0,\theta_0)$. One way to create an image with this property is to simply assign every point on $L(\ell_0,\theta_0)$ the value $g(\ell_0,\theta_0)$. When we repeat this for all ℓ, the resulting function is called a *backprojection image* and is given formally by

$$b_\theta(x,y)=g(x\cos\theta+y\sin\theta,\theta)\,. \tag{6.13}$$

The backprojection image $b_{30^\circ}(x,y)$ for the example in Figure 6.9 is shown in Figure 6.10(a).

Loosely speaking, the backprojection image at angle θ_0 is consistent with the projection at angle θ_0, but its values are assigned with no prior information

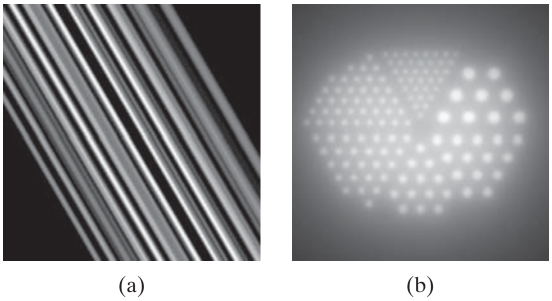


Figure 6.10
(a) A backprojection image and (b) a backprojection summation image.

about the distribution of the image intensities.[1] To incorporate information about the projections at other angles, we can simply add up (integrate) their backprojection images, yielding the *backprojection summation image*

$$f_b(x, y) = \int_0^{\pi} b_\theta(x, y)\, d\theta\,, \tag{6.14}$$

also called a *laminogram*. An example of a backprojection summation image is shown in Figure 6.10(b). Early scanners did this using a discrete approximation to the integral. However, it was soon shown analytically that this is the wrong thing to do; this is also revealed from the blurriness of Figure 6.10(b). The concept is useful, however, because the correct procedure also uses backprojection except that it applies to filtered versions of the projections, as we shall see.

EXAMPLE 6.3

Question Consider the projection $g(\ell, 45^\circ) = \text{sgn}(\ell)$, where $\text{sgn}(\ell)$ is -1 when $\ell \leq 0$ and is $+1$ when $\ell > 0$. What is the backprojection image $b_\theta(x, y)$ at $\theta = 45^\circ$.

Answer We have

$$\begin{aligned} b_{45^\circ}(x, y) &= g\left(x \cos 45^\circ + y \sin 45^\circ, 45^\circ\right) \\ &= g\left(x\frac{\sqrt{2}}{2} + y\frac{\sqrt{2}}{2}, 45^\circ\right) \\ &= g\left(\frac{\sqrt{2}}{2}(x + y), 45^\circ\right)\,. \end{aligned}$$

When $x + y$ is negative, g is -1; when $x + y$ is positive, g is $+1$. Mathematically, this is

$$b_{45^\circ}(x, y) = \begin{cases} -1 & x + y \leq 0 \\ +1 & x + y > 0 \end{cases}\,,$$

which is the desired result and is shown in Figure 6.11. ■

---

[1] Actually, a backprojection image has infinite energy (if its domain is considered to be the infinite plane), and its projection at angle θ_0 is infinite. Therefore, it is not really consistent with the projection.

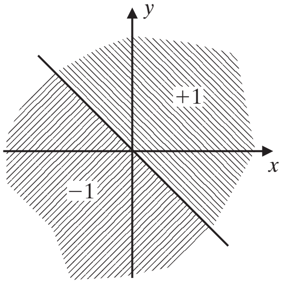


Figure 6.11
A backprojection image.

Projection-Slice Theorem In this section, we develop a very important relationship between the 1-D Fourier transform of a projection and the 2-D Fourier transform of the object.

We begin by taking the 1-D Fourier transform of a projection with respect to ℓ:

$$G(\varrho,\theta) = \mathcal{F}_{1\mathrm{D}}\{g(\ell,\theta)\} = \int_{-\infty}^{\infty} g(\ell,\theta)e^{-j2\pi\varrho\ell}\,d\ell\,, \tag{6.15}$$

where ϱ denotes spatial frequency (like u or v, except in an arbitrary direction). Now we substitute the analytic expression for $g(\ell,\theta)$ given in (6.10) and manipulate this expression:

$$\begin{aligned} G(\varrho,\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - \ell)e^{-j2\pi\varrho\ell}\,dx\,dy\,d\ell \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\int_{-\infty}^{\infty} \delta(x\cos\theta + y\sin\theta - \ell)e^{-j2\pi\varrho\ell}\,d\ell\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi\varrho(x\cos\theta + y\sin\theta)}\,dx\,dy\,, \end{aligned} \tag{6.16}$$

where the last step followed from the sifting property of the delta function [see (2.6)].

The final expression for $G(\varrho,\theta)$ in (6.16) is reminiscent of the 2-D Fourier transform of $f(x,y)$, defined as

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(xu+yv)}\,dx\,dy\,, \tag{6.17}$$

where the variables u and v are the frequency variables in the x and y directions, respectively. In fact, making the identification $u = \varrho\cos\theta$ and $v = \varrho\sin\theta$ leads directly to the equivalence and to the very important relationship

$$G(\varrho,\theta) = F(\varrho\cos\theta, \varrho\sin\theta)\,. \tag{6.18}$$

Equation (6.18) is known as the *projection-slice theorem*, and it is the basis of three important reconstruction methods. The relationship states that the 1-D Fourier transform of a projection is a *slice* of the 2-D Fourier transform of the object. Said another way, the 1-D Fourier transform of the projection equals a line passing through the origin of the 2-D Fourier transform of the object at

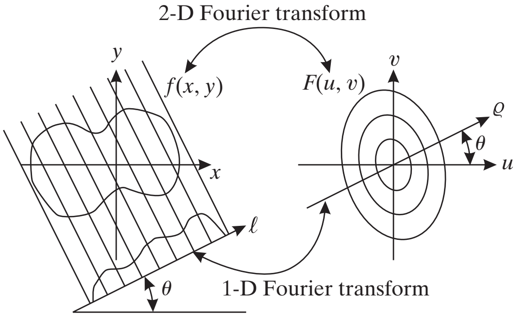


Figure 6.12
The projection-slice theorem in graphical form.

that angle corresponding to the projection. A graphical interpretation is shown in Figure 6.12. We see that ϱ and θ may be interpreted as the polar coordinates of the 2-D Fourier transform.

EXAMPLE 6.4

The projection-slice theorem helps us to understand how angular sampling can influence reconstructions. Suppose that only eight projections are acquired at angles $\theta_i = \pi(i + 0.5)/8$, $i = 0, \ldots, 7$. Figure 6.13 shows the locations of the acquired Fourier data corresponding to these projections.

Question Show that the function $f(x, y) = \cos x$ is invisible at these angles—that is, it will produce projections that are identically zero at these angles.

Answer The 2-D Fourier transform of $\cos x$ is

$$F(u, v) = \pi\,[\delta(2\pi u - 1) + \delta(2\pi u + 1)]\,\delta(v)\,,$$

which has two 2-D impulse functions on the u-axis, as shown in Figure 6.13. On the other hand, the projection-slice theorem tells us that, given the available observations, we will know the Fourier transform of f only on the lines $\{(\varrho\cos\theta_i, \varrho\sin\theta_i) | -\infty < \varrho < \infty\}$ for $i = 0, \ldots, 7$. Notice that although these lines are equally spaced over $[0, \pi]$, they do not include the u-axis. Therefore, they will be zero, and the function $\cos x$ will be invisible. ■

The Fourier Method A conceptually simple reconstruction method, called the *Fourier method*, follows immediately from the projection-slice theorem. One simply takes the 1-D Fourier transform of each projection, inserts it with the

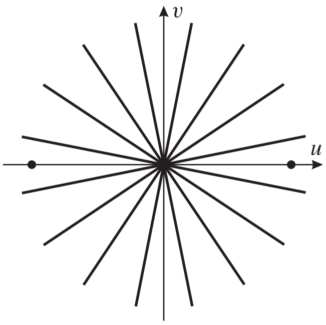


Figure 6.13
See Example 6.4.

corresponding correct angular orientation into the correct slice of the 2-D Fourier plane, and takes the inverse 2-D Fourier transform of the result. Accordingly, we have

$$f(x, y) = \mathcal{F}_{2D}^{-1}\{G(\varrho, \theta)\}\,. \tag{6.19}$$

The Fourier method is not widely used in CT due to the practical problem of interpolating polar data onto a Cartesian grid, and the need to use the relatively time-consuming 2-D inverse Fourier transform. Further analytic manipulation of (6.19), however, leads to improvements, as we shall see below.

Filtered Backprojection The inverse Fourier transform of $F(u, v)$ [see (6.17)] can be written in polar coordinates as

$$f(x, y) = \int_0^{2\pi} \int_0^{\infty} F(\varrho\cos\theta, \varrho\sin\theta) e^{j2\pi\varrho(x\cos\theta + y\sin\theta)} \varrho \, d\varrho d\theta\,. \tag{6.20}$$

Using the projection-slice theorem, we have

$$f(x, y) = \int_0^{2\pi} \int_0^{\infty} G(\varrho, \theta) e^{j2\pi\varrho(x\cos\theta + y\sin\theta)} \varrho \, d\varrho d\theta\,, \tag{6.21}$$

from which it follows that

$$f(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} |\varrho| G(\varrho, \theta) e^{j2\pi\varrho(x\cos\theta + y\sin\theta)} \, d\varrho d\theta\,, \tag{6.22}$$

after some work and use of the fact that $g(\ell, \theta) = g(-\ell, \theta + \pi)$.

From the point of view of integration over ϱ, the term $(x\cos\theta + y\sin\theta)$ in (6.22) is a constant, say ℓ. Hence, (6.22) may be written as

$$f(x, y) = \int_0^{\pi} \left[\int_{-\infty}^{\infty} |\varrho| G(\varrho, \theta) e^{j2\pi\varrho\ell} \, d\varrho \right]_{\ell = x\cos\theta + y\sin\theta} d\theta\,. \tag{6.23}$$

We recognize the inner integral in (6.23) as an inverse 1-D Fourier transform [see (2.63)], where the term $|\varrho|$ makes (6.23) a *filtering* equation. Here, the Fourier transform of the projection $g(\ell, \theta)$ is multiplied by the frequency filter $|\varrho|$ and inverse-transformed. After the inverse transform, the filtered projection is backprojected (this is accomplished by replacing ℓ with $x\cos\theta + y\sin\theta$), which is followed by a "summation" of all filtered projections. This reconstruction approach is appropriately called *filtered backprojection*, and it is a considerable improvement over the Fourier method in speed and flexibility. The term $|\varrho|$ is known as the *ramp filter* because of its appearance in Fourier space.

An intuitive understanding of the role of the ramp filter can be found from inspection of Figure 6.13. Here, the straight application of the Fourier method leads to sampling that is inversely proportional to ϱ. Additional area or "weight" is required to compensate for the sparser sampling at higher frequencies.

Convolution Backprojection From the convolution theorem of Fourier transforms [see (2.91)], we may write (6.23) as

$$f(x,y) = \int_0^\pi \left[\mathcal{F}_{1D}^{-1}\{|\varrho|\} * g(\ell,\theta)\right]_{\ell = x\cos\theta + y\sin\theta} d\theta\,. \tag{6.24}$$

Then, defining $c(\ell) = \mathcal{F}_{1D}^{-1}\{|\varrho|\}$, we have

$$f(x,y) = \int_0^\pi [c(\ell) * g(\ell,\theta)]_{\ell = x\cos\theta + y\sin\theta}\, d\theta \tag{6.25}$$

$$= \int_0^\pi \int_{-\infty}^{\infty} g(\ell,\theta) c(x\cos\theta + y\sin\theta - \ell)\, d\ell d\theta\,, \tag{6.26}$$

where (6.26) results from substitution of the convolution integral and is the equation for *convolution backprojection*. Performing a convolution rather than a filtering operation (a pair of Fourier transforms with a multiplication in-between) is generally more efficient if the impulse response is narrow. It is so in this case, and generally most CT scanners perform (some form of) convolution backprojection rather than filtered backprojection.

Unfortunately, $c(\ell)$ does not exist, since $|\varrho|$ is not integrable (and therefore its inverse Fourier transform is undefined). However, various formal expressions involving generalized functions (e.g., delta functions and their derivatives) or limits of functions have been developed. While these expressions are useful in a theoretical setting, we require a practical approach to design an impulse response that can be used in an actual CT scanner. This is accomplished by *windowing* $|\varrho|$ with a suitable windowing function $W(\varrho)$, such as a square, Hamming, or cosine window. Here, windowing denotes the use of a filter, in addition to the ramp filter, that modifies the observed projection. In practice, therefore, convolution backprojection (CBP) algorithms use the approximate impulse response

$$\tilde{c}(\ell) = \mathcal{F}_{1D}^{-1}\{|\varrho| W(\varrho)\}\,. \tag{6.27}$$

It is also customary to set the filter value at $\varrho = 0$ to be nonzero, in order to produce the correct reconstructed average image value.

Reconstruction in Three Steps Filtered backprojection (6.23) uses three basic steps to reconstruct an image from a sinogram: (1) filtering, (2) backprojection, and (3) summation. Convolution backprojection (6.24) uses the same three steps except that the filtering operation is implemented using a convolution. The reconstructed images resulting from these two approaches will be identical except for round-off errors due to numerical implementation issues.

Figure 6.14 illustrates the filtering (or convolution) step. Each row of the sinogram in Figure 6.14(a) is filtered using a windowed ramp filter to yield the filtered sinogram in Figure 6.14(b). The ramp filter is essentially a high-pass filter and its value at $f = 0$ is 0. As a result, high frequency detail is accentuated, the background value is now zero (gray in the filtered sinogram), and there are negative values (dark values in the filtered sinogram). A projection and its corresponding filtered projection are also shown in this figure for comparison.

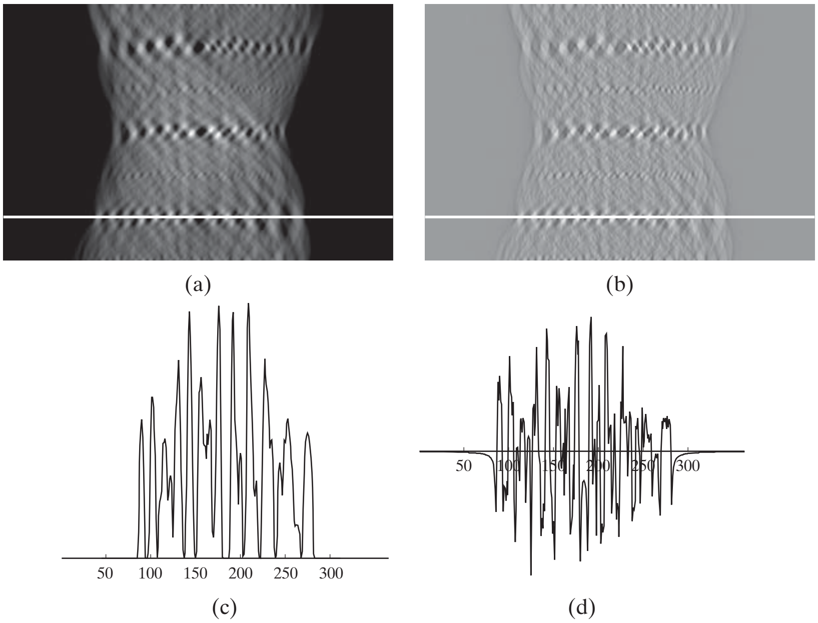

(a) (b)
(c) (d)

Figure 6.14
Convolution step:
(a) Original sinogram;
(b) filtered sinogram;
(c) profile of sinogram row [white line in (a)];
and (d) profile of filtered sinogram row [white line in (b)].

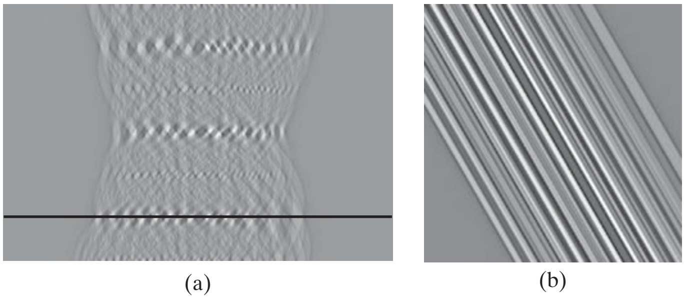
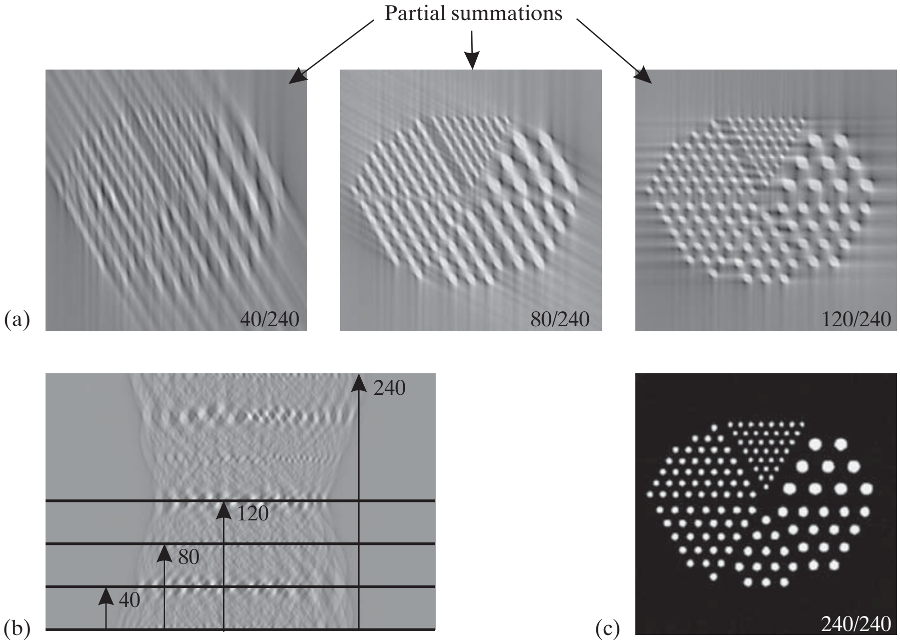

(a) (b)

Figure 6.15
Backprojection step.
(a) A selected filtered projection and (b) the backprojection image created from the filtered projection.

We have already seen the backprojection operation. Here, backprojection is applied to each filtered projection, rather than the raw projection. An example is shown in Figure 6.15. Since the filtered projections have both positive and negative values, backprojection images also have positive and negative values. It can be understood on an intuitive level that the presence of negative values will permit the reconstruction of a zero value outside the region of support of the object; of course, this requires the summation of many backprojection images for this to happen.

The last step in reconstruction is the summation step, a process that is illustrated in Figure 6.16. Notice that the summation process is accomplished using an "accumulator" concept. Starting with a "zero" image, the first (filtered) backprojection image is added. Then, the second backprojection image is added, and so forth. There is no need to store hundreds of backprojection images. The overall process of filtered backprojection is very time and memory efficient, which accounts for its great popularity over the past three decades.



Filtered sinogram

Final reconstruction

Figure 6.16 Summation step. (a) Partial summation images using just 40, 80, and 120 (out of 240 total) projections. (b) The filtered projections that are used in the partial summation images shown in (a). (c) The image that is reconstructed using all 240 filtered projections.

6.3.4 Fan-Beam Reconstruction

As discussed in Section 6.2, all modern commercial scanners have a fan-beam source-to-detector arrangement. Therefore, in practice it is necessary to reconstruct images from fan-beam projections rather than parallel-ray projections. In this section, we develop a CBP reconstruction algorithm that can be applied to a fan-beam geometry.

Geometry There are three basic fan-beam geometries to consider: (1) those that have equal angles between the measured ray-paths, (2) those that have equal detector spacing, and (3) those that have equal angles and equal detector spacing. The third geometry can only be satisfied if the detectors are placed along a circular arc whose center is at the source. In this section, we consider the equal-angle case with detectors positioned on a circular arc. A reconstruction formula for the case of equal detectors on a straight line can be developed using analogous means.

We consider the fan-beam geometry shown in Figure 6.17. This is a 3G geometry in which the source and detector rotate together around the laboratory origin, also called the rotational *isocenter*. A fan-beam projection, measured at each rotational position of the source, is denoted $p(\gamma, \beta)$ [in analogy to $g(\ell, \theta)$], where γ is the angular position of a given detector and β is the angular position of the source, as shown in Figure 6.17. Notice that both γ and β are angles measured in the counterclockwise direction; however, their origins are different and so are their reference directions (physical orientations when taking on zero values). In particular, β measures the source position around the isocenter, and it is zero when the source is resting on the $+y$-axis. On the other hand, γ is

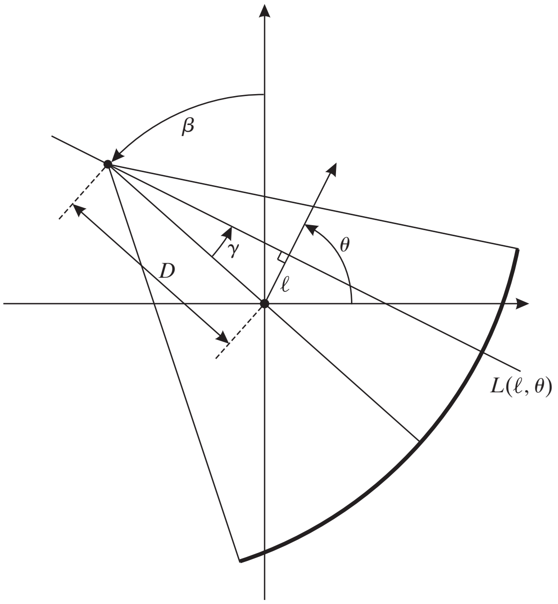


Figure 6.17
Fan-beam geometry.

measured around the source position, and it is zero for rays that pass through the center of the detector array.

It is easiest to derive the fan-beam CBP formula by starting with what we already know about parallel-ray reconstruction. Within the fan-beam projection we have identified a line $L(\ell, \theta)$ in Figure 6.7, where ℓ is the lateral displacement and θ is the angular orientation, exactly as before. We can see from the geometry that the parameters of this line, ℓ and θ, are related to the fan-beam parameters by

$$\theta = \beta + \gamma\,, \tag{6.28}$$

$$\ell = D \sin \gamma\,, \tag{6.29}$$

where D is the distance from the source to the origin (isocenter).

Fan-Beam Reconstruction Formula Our development of a fan-beam reconstruction formula begins with the equation for parallel-ray CBP [(6.26)], which we repeat here for convenience:

$$f(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} g(\ell, \theta) c(x \cos \theta + y \sin \theta - \ell)\, d\ell\, d\theta\,. \tag{6.30}$$

From the geometry of parallel-ray tomography,

$$f(x, y) = \frac{1}{2} \int_0^{2\pi} \int_{-T}^{T} g(\ell, \theta) c(x \cos \theta + y \sin \theta - \ell)\, d\ell\, d\theta\,, \tag{6.31}$$

where we have assumed that $g(\ell, \theta) = 0$ for $|\ell| > T$. Thus, in (6.31) we consider rotation over the full circle and objects that are zero outside the disk with radius T, centered at the origin.

Let (r, ϕ) be polar coordinates in the plane. Then $x = r\cos\phi$, $y = r\sin\phi$, and $x\cos\theta + y\sin\theta = r\cos\phi\,\cos\theta + r\sin\phi\,\sin\theta = r\cos(\theta - \phi)$. Substituting this last expression into (6.31) gives

$$f(r, \phi) = \frac{1}{2}\int_0^{2\pi}\int_{-T}^{T} g(\ell, \theta)c(r\cos(\theta - \phi) - \ell)\,d\ell\,d\theta\,, \tag{6.32}$$

which is a parallel-ray reconstruction formula for f written in polar coordinates.

We now want to integrate over γ and β instead of ℓ and θ. This is a transformation of coordinates, where the transformation is given in (6.28) and (6.29). The Jacobian of this transformation is $D\cos\gamma$, and direct substitution yields

$$\begin{aligned} f(r, \phi) = \frac{1}{2}\int_{-\gamma}^{2\pi-\gamma}\int_{\sin^{-1}\frac{-T}{D}}^{\sin^{-1}\frac{T}{D}} & g(D\sin\gamma, \beta + \gamma) \\ & \times c(r\cos(\beta + \gamma - \phi) - D\sin\gamma)D\cos\gamma\,d\gamma\,d\beta\,. \end{aligned} \tag{6.33}$$

The following simplifications to this integral can be made:

- The limits of the outer integral can be replaced by 0 (lower limit) and 2π (upper limit), since the functions are periodic in β with period 2π.
- The expression $\sin^{-1}\frac{T}{D}$ represents the largest angle γ_m that needs to be considered given that the object is contained in a disk of radius T. Thus, we can replace the upper and lower limits on the inner integral by γ_m and $-\gamma_m$, respectively.
- We recognize the fan-beam projection as

$$p(\gamma, \beta) = g(D\sin\gamma, \beta + \gamma)\,. \tag{6.34}$$

This gives the basic fan-beam reconstruction formula:

$$f(r, \phi) = \frac{1}{2}\int_0^{2\pi}\int_{-\gamma_m}^{\gamma_m} p(\gamma, \beta)c\left(r\cos(\beta + \gamma - \phi) - D\sin\gamma\right)D\cos\gamma\,d\gamma\,d\beta\,. \tag{6.35}$$

Fan-Beam Convolution Backprojection In this section, we manipulate the fan-beam reconstruction formula of (6.35) into a form that more closely resembles ordinary CBP.

Consider an arbitrary point in the cross section, given by the polar coordinates (r, ϕ), as shown in Figure 6.18. Its position can also be defined relative to the source and detector positions using the angle γ' and radius D', as shown in the figure. The argument of $c(\cdot)$ can be written in simpler form using these coordinates (after some trigonometric manipulation) as

$$r\cos(\beta + \gamma - \phi) - D\sin\gamma = D'\sin(\gamma' - \gamma)\,, \tag{6.36}$$

where γ is an arbitrary angle as depicted in Figure 6.17. This yields

$$f(r, \phi) = \frac{1}{2}\int_0^{2\pi}\int_{-\gamma_m}^{\gamma_m} p(\gamma, \beta)c(D'\sin[\gamma' - \gamma])D\cos\gamma\,d\gamma\,d\beta\,, \tag{6.37}$$

where D' and γ' are determined by r and ϕ.

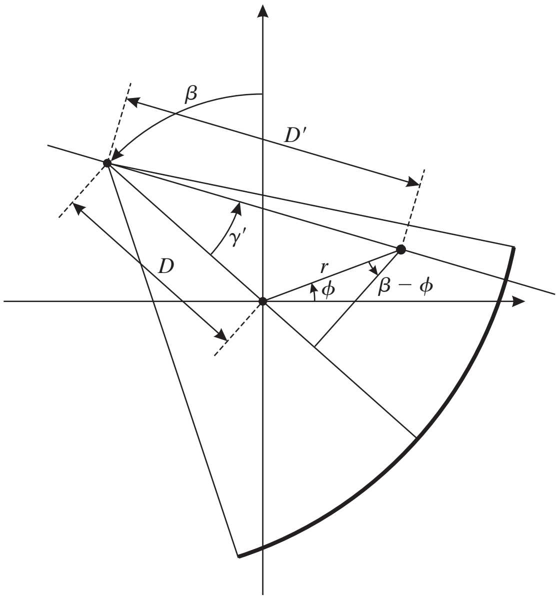


Figure 6.18
Rotating polar coordinates for fan-beam geometry.

It can be shown (see Problem 6.18) that

$$c(D' \sin \gamma) = \left(\frac{\gamma}{D' \sin \gamma} \right)^2 c(\gamma) \,. \tag{6.38}$$

Accordingly, looking back at (6.37), we see that it is useful to define

$$c_f(\gamma) = \frac{1}{2} D \left(\frac{\gamma}{\sin \gamma} \right)^2 c(\gamma) \,. \tag{6.39}$$

Combining (6.37) and (6.38) with this definition, we find that

$$f(r, \phi) = \int_0^{2\pi} \frac{1}{(D')^2} \int_{-\gamma_m}^{\gamma_m} \tilde{p}(\gamma, \beta) c_f(\gamma' - \gamma) \, d\gamma \, d\beta \,, \tag{6.40}$$

where $\tilde{p}(\gamma, \beta) = \cos \gamma \, p(\gamma, \beta)$, which is our final fan-beam reconstruction formula.

To more clearly identify the "convolution" part and the "backprojection" part, we split the reconstruction formula of (6.40) into two pieces:

$$q(\gamma, \beta) = \tilde{p}(\gamma, \beta) * c_f(\gamma) \,, \tag{6.41}$$

$$f(r, \phi) = \int_0^{2\pi} \frac{1}{(D')^2} q(\gamma', \beta) \, d\beta \,. \tag{6.42}$$

Here, $q(\cdot)$ is simply a filtered projection where the impulse response $c_f(\cdot)$ is a weighted (or windowed) version of $c(\cdot)$, the usual (approximate) ramp filter [as in (6.27)]. However, the backprojection operator is a bit different since D' depends on the image coordinate (r, ϕ) or, equivalently, (x, y). Therefore, each filtered projection backprojects along the ray-paths of integration but does so with a weighting that depends on the distance from the source. Thus, one should think of the phrase *convolution weighted-backprojection* when the projections are fan-beams.

6.3.5 Helical CT Reconstruction

Consider the helical CT scenario in which there is a single detector array and the patient table is continuously moved through the gantry. In this case, the nature of the acquired data is different than in single-slice CT. Since x-rays move at the speed of light, any acquired projection still corresponds to a single plane within the patient; however, because of the continuous movement of the patient, no two projections will correspond to the same plane. A given (fan-beam) projection is therefore identified uniquely by its longitudinal position z_j, measured relative to the patient. The index j, in fact, not only identifies the position z_j of the projection but also determines its angle β_j.

Suppose there are M angles acquired over 360°. Then, the acquired angles repeat such that $\beta_j - \beta_{j+M} = 0$. How far has the patient moved when the angle repeats? This is determined by the *pitch* of the helix, which is given by

$$\zeta = z_{j+M} - z_j \,, \tag{6.43}$$

which is in turn determined by the speed of gantry rotation and the table speed.

Suppose that we want to reconstruct an image at the longitudinal position z. In conventional (single-slice) tomography, we would need many projections taken at a collection of angles within the axial plane passing through z. In the helical geometry, however, we are not guaranteed even to have one such projection (because z may not correspond to z_n for any n). It is standard practice to "create" the needed projections using linear interpolation of the projections that are measured in nearby slices. For example, suppose we would like the projection corresponding to the angle β_j. We know that the projection angle β_j is acquired at all linear positions z_{j+kM} for $k = 0, 1, 2, \ldots$. It is straightforward to find k such that $z_{j+kM} \le z < z_{j+[k+1]M}$. We know that the scanner acquires projections at angle β_j at both positions z_{j+kM} and $z_{j+[k+1]M}$. Linear interpolation between these two projections yields an estimate of the desired projection

$$\begin{aligned}\hat{p}_z(\gamma, \beta_j) = {} & \frac{z - z_{j+kM}}{\zeta} p_{z_{j+[k+1]M}}(\gamma, \beta_j) \\ & + \frac{z_{j+[k+1]M} - z}{\zeta} p_{z_{j+kM}}(\gamma, \beta_j) \,,\end{aligned} \tag{6.44}$$

which is valid for any position γ within the projection. This procedure can be repeated for all $j = 1, \ldots, M$, noting that the choice of k depends on both z and j. In this way, a complete set of projections for fan-beam reconstruction of the 2-D slice at z is created. A fan-beam reconstruction formula can then be used to reconstruct a picture of the slice.

Many factors will affect the quality of a helical CT image. In particular, the quality of interpolated data will be affected by the pitch of the helix. The pitch is determined by the rate of rotation of the gantry and the speed of the table feed. If the pitch is too big, while the fan is too thin, then aliasing artifacts will appear. To compensate, the fan could be made thicker, which provides a natural analog low-pass filter. Then, the images will be free from aliasing but will be blurry. In clinical scanners today, protocols are preset into the scanner for certain types of studies. These protocols balance the conflicting requirements of image quality with FOV requirements and speed of acquisition.

6.3.6 Cone Beam CT

With the goal of rapid coverage and reconstruction of 3-D volumes, it is natural to consider the use of a cone beam radiation pattern and an area detector (rather than a fan-beam and a linear detector). A standard fluoroscope, for example, can rotate the cone-beam source and 2-D detector in a circle around the body, electronically recording 2-D projection images at a collection of different angles. Also, the geometry of MDCT systems with large-area detectors is correctly understood to be in the form of a cone beam. In either case, the data within the plane of the circular source path are exactly what one achieves with a conventional fan-beam CT scanner; thus, an image can be reconstructed by conventional means for this central image plane. It turns out, however, that there is not sufficient coverage to yield mathematically correct reconstructions in any other plane through the object. The situation is further complicated by the fact that projections are often *truncated*; that is, the body extends beyond the x-ray cone.

Despite these complications, so-called *cone beam reconstruction algorithms* have been developed to produce excellent 3-D reconstructions. The Feldkamp algorithm is the most well known and has been shown to produce excellent reconstructions for the geometry described above.

6.3.7 Iterative Reconstruction

For the past 40 years, commercial CT systems have all used variants of CBP for image reconstruction. However, current systems are rapidly changing over to *iterative reconstruction* methods because they offer equivalent (or better) reconstruction quality at substantially reduced dose. We describe iterative reconstruction methods in Chapter 9, when describing emission tomography systems.

Iterative reconstruction methods are more computationally demanding than CBP. Since emission tomography's data requirements are lower than that of most other imaging modalities (because larger and thus fewer voxels are used to represent a 3D volume), iterative reconstruction methods were introduced roughly two decades ago for emission tomographic reconstruction and are thus more fully developed in that application. Also, emission tomography is significantly limited by noise, especially compared with most other imaging modalities, and iterative reconstruction methods are particularly adept at handling noise.

The application of iterative reconstruction methods to CT is conceptually straightforward and follows the same development as in Chapter 9 for emission tomography.

6.4 Image Quality in CT

While the theory presented in the previous sections captures the ideal behavior of CT systems, in practice there are limitations to what one can achieve. It is not possible to achieve arbitrarily high resolution, since detectors have finite widths and there are sampling limitations, both in the number of projections and in the number of ray-paths per projection. As well, since the patient will absorb x-rays, we cannot use an arbitrarily large dose, and this limits the attainable signal-to-noise ratio (SNR). A third problem is beam hardening, which is a phenomenon caused by energy selective absorption of x-rays by human tissues (as discussed in Section 5.2.2). In this section, we describe these phenomena and indicate how they affect the appearance of reconstructed CT images.

6.4.1 Resolution

Filtered backprojection [see (6.23)] is an exact formula for the inverse Radon transform. As noted in (6.27), however, the ideal (ramp) filter $|\varrho|$ is not realizable, so an approximate filter $W(\varrho)|\varrho|$ using a window function $W(\varrho)$ [see (6.27)] must be used in practice. In addition, the line integrals themselves cannot be imaged exactly due to the finite size of the detectors. In fact, CT detectors are *area detectors* that locally integrate the underlying true signal, an effect that can be modeled using an additional filter $S(\varrho)$ [see (3.81)], which is the Fourier transform of the indicator function $s(\ell)$ defining a detector. Combining these effects, the filtered backprojection formula becomes

$$\hat{f}(x,y) = \int_0^{\pi} \left[\int_{-\infty}^{\infty} G(\varrho,\theta)S(\varrho)W(\varrho)|\varrho|e^{j2\pi\varrho\ell}\, d\varrho \right]_{\ell = x\cos\theta + y\sin\theta} d\theta\,. \qquad (6.45)$$

The reconstruction in (6.45) is approximate, and the resulting function is a blurred version of $f(x,y)$. We can discover a relation between $\hat{f}(x,y)$ and $f(x,y)$ by manipulating (6.45).

First, we recognize that (6.45) is precisely the inverse Radon transform of the function

$$\hat{g}(\ell,\theta) = \mathcal{F}^{-1}\{G(\varrho,\theta)S(\varrho)W(\varrho)\}\,. \qquad (6.46)$$

By the 1-D convolution theorem,

$$\hat{g}(\ell,\theta) = g(\ell,\theta) * \tilde{h}(\ell)\,, \qquad (6.47)$$

where

$$\tilde{h}(\ell) = s(\ell) * w(\ell)\,. \qquad (6.48)$$

The result of filtered backprojection can therefore be interpreted as the inverse Radon transform of the blurry projections $\hat{g}(\ell,\theta)$.

In order to determine how the blurry projections cause the image of the object to be blurred, it is convenient to draw upon the *convolution property* of the Radon transform. In particular, it can be shown that the Radon transform of the convolution of two functions is the convolution of their Radon transforms (see Problem 6.9)

$$\mathcal{R}\left\{f * h\right\} = \mathcal{R}\{f\} * \mathcal{R}\{h\}\,. \qquad (6.49)$$

Notice that the convolution on the left in (6.49) is two-dimensional, while that on the right is one-dimensional. Comparing (6.47) with (6.49) allows us to identify

$$\mathcal{R}\{h\} = \tilde{h}(\ell)\,, \qquad (6.50)$$

and therefore, upon taking the inverse Radon transform of both sides,

$$h(x, y) = \mathcal{R}^{-1}\{\tilde{h}(\ell)\}\,. \tag{6.51}$$

From the convolution property of the Radon transform (6.49), we can conclude that the reconstructed object is given by

$$\hat{f}(x, y) = f(x, y) * \mathcal{R}^{-1}\{\tilde{h}(\ell)\}\,. \tag{6.52}$$

Therefore, the reconstructed estimate $\hat{f}(x, y)$ is a blurry version of $f(x, y)$, and the resolution of the system can be characterized by the point spread function (PSF) $\mathcal{R}^{-1}\{\tilde{h}(\ell)\}$.

To find the PSF $h(x, y)$, we use the projection-slice theorem (6.18). The Fourier transform of $\tilde{h}(\ell)$ is given by

$$\tilde{H}(\varrho) = S(\varrho)W(\varrho)\,, \tag{6.53}$$

which is independent of θ. Accordingly, the 2-D Fourier transform of $h(x, y)$ must be circularly symmetric and given by

$$\mathrm{H}(q) = S(q)W(q)\,. \tag{6.54}$$

It follows that the PSF is also circularly symmetric and given by the inverse Hankel transform [see (2.113)]

$$\mathrm{h}(r) = \mathcal{H}^{-1}\{S(\varrho)W(\varrho)\}\,. \tag{6.55}$$

Finally, the reconstructed image is given by

$$\hat{f}(x, y) = f(x, y) * \mathrm{h}(r)\,, \tag{6.56}$$

where $r = \sqrt{x^2 + y^2}$. The PSF is circularly symmetric and can be characterized by a full width at half maximum (FWHM) given the detector width, which then determines $S(\varrho)$, and the ramp filter window function $W(\varrho)$.

EXAMPLE 6.5

Suppose a CT system uses rectangular detectors having width d and a rectangular window function with highest frequency $\varrho_0 \gg 1/d$.

Question What is the approximate PSF of this CT system?

Answer Rectangular detectors with width d blur the underlying projection with impulse response function

$$s(\ell) = \mathrm{rect}\left(\frac{\ell}{d}\right)\,.$$

In the frequency domain, this is represented by the filter

$$S(\varrho) = d\,\mathrm{sinc}(d\varrho)\,.$$

The ramp filter window function can be written as

$$W(\varrho) = \mathrm{rect}\left(\frac{\varrho}{2\varrho_0}\right) .$$

The first zero of $S(\varrho)$ is at frequency $\varrho = 1/d$. Since $\varrho_0 \gg 1/d$, the frequency cutoff defined by the detectors is much less than that of the ramp filter, and the ramp filter window function can be ignored. Therefore, the (circularly symmetric) 2-D transfer function is (approximately) given by

$$\mathrm{H}(q) \approx d\,\mathrm{sinc}(dq) .$$

From Table 2.3,

$$\mathcal{H}\{\mathrm{sinc}(r)\} = \frac{2\,\mathrm{rect}(q)}{\pi\sqrt{1-4q^2}} .$$

Using the fact that the forward and inverse Hankel transforms are the same, and [see (2.120)]

$$\mathcal{H}\{\mathrm{f}(ar)\} = \frac{1}{a^2}\mathrm{F}(q/a) ,$$

we can work out the following (approximate) impulse response:

$$\begin{aligned} \mathrm{h}(r) &= \mathcal{H}^{-1}\{d\,\mathrm{sinc}(d\varrho)\} \\ &= d\frac{1}{d^2}\frac{2\,\mathrm{rect}(r/d)}{\pi\sqrt{1-4(r/d)^2}} \\ &= \frac{2\,\mathrm{rect}(r/d)}{\pi\sqrt{d^2-4r^2}} . \end{aligned}$$

■

6.4.2 Noise

Measurement Statistics A CT detector measures the intensity of the x-ray pulse after losses in the patient. There is intrinsic noise in such a measurement, coming from the Poisson nature of x-rays [see (3.52)]. In addition, body attenuation and limited detector efficiency reduce the measured number of photons, further increasing noise.[2] We will make many approximations in this section in arriving at a useful approximate expression for the noise in a CT image. The first approximation is that we will assume that the CT system is monoenergetic, operating at the effective energy $\overline{E}$ of the scanner. In this case, the statistical behavior of the scanner can be understood by converting the observed intensity at a given detector I_d to an observed photon count as $N_d = I_d A \Delta t / \overline{E}$.

We denote the basic CT measurement for the ith detector and jth angle by

$$g_{ij} = -\ln\left(\frac{N_{ij}}{N_0}\right) , \qquad (6.57)$$

---

[2] The x-ray beam itself arises from a random process, and the number of x-rays in the beam is a Poisson random variable. Body attenuation and detector efficiency produce *random deletions* (in statistical jargon) of these x-rays, and the number of x-rays in the detected beam is also a Poisson random variable.

where N_0 is the incident photon count and N_{ij} is the number of detected photons. Unfortunately, because the absorption of photons is a statistical phenomenon, the same experiment repeated again would probably yield a different measurement. In fact, N_{ij} is a Poisson random variable with mean

$$\overline{N}_{ij} = N_0 \exp\left(- \int_{L_{ij}} \mu(s)\, ds \right) , \tag{6.58}$$

where L_{ij} is the ray-path from the source to detector i for the jth projection angle. Equation (6.58) is just another way of writing the line integral introduced in (6.1). Notice that the mean depends on the number of incident photons and on the line integral of the linear attenuation coefficient along the ray-path between the source and detector. In fact, the mean is the desired measurement.

The basic measurement g_{ij} is a transformation of the random variable N_{ij}, however, not a transformation of the mean of N_{ij}. Therefore, g_{ij} is a random variable and, assuming N_0 is large, its mean $\overline{g}_{ij}$ and variance $\mathrm{var}(g_{ij})$ are given by

$$\overline{g}_{ij} \approx -\ln\left(\frac{\overline{N}_{ij}}{N_0} \right) , \tag{6.59}$$

$$\mathrm{var}(g_{ij}) \approx \frac{1}{\overline{N}_{ij}} . \tag{6.60}$$

We assume that the random variables N_{ij} are independent, and it follows that the random variables g_{ij} are also independent. Given this second-order characterization of the basic CT measurements, we are now in a position to determine the second-order statistics of the images that we create from these measurements. These second-order statistics will permit us to define a signal-to-noise ratio of reconstructed images, which is a measure of the quality of reconstructions.

Image Statistics In this section, we develop an expression for the mean and variance of the reconstructed linear attenuation coefficients. For simplicity, we assume parallel-ray geometry. It is mathematically convenient to use a discrete approximation to the convolution backprojection (CBP) integral in (6.26), which is given by

$$\hat{\mu}(x, y) = \frac{\pi T}{M} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} g_{\theta_j}(iT)\, \tilde{c}\left(x \cos\theta_j + y \sin\theta_j - iT\right) , \tag{6.61}$$

where M is the number of projections taken over the range $[0, \pi)$, $N + 1$ (odd) is the number of ray-paths per projection, T is the physical spacing between detectors, and $g_{ij} = g_{\theta_j}(iT)$. Notice that $\tilde{c}(\cdot)$ is a realizable approximation to the ramp filter. The mean of the reconstructed image is therefore

$$\mathrm{mean}[\hat{\mu}(x, y)] = \frac{\pi T}{M} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} \overline{g}_{ij}\, \tilde{c}\left(x \cos\theta_j + y \sin\theta_j - iT\right) , \tag{6.62}$$

which for large N_0, M, and N is exactly what we want. Therefore, CBP has the desirable property that the mean of the reconstructed image approaches exact reconstruction as the quality of measurement increases.

Since g_{ij} are assumed to be independent random variables, the variance of the sum is the sum of the variances (see (3.56)), given by

$$\sigma^2(x,y) = \mathrm{var}[\hat{\mu}(x,y)] \tag{6.63}$$

$$= \frac{\pi^2 T^2}{M^2} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} \mathrm{var}[g_{ij}] \left[c\left(x\cos\theta_j + y\sin\theta_j - iT\right)\right]^2 \tag{6.64}$$

$$= \frac{\pi^2 T^2}{M^2} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} \frac{1}{\overline{N}_{ij}} \left[c\left(x\cos\theta_j + y\sin\theta_j - iT\right)\right]^2 . \tag{6.65}$$

To proceed, we make a rather drastic approximation: $\overline{N}_{ij} \approx \overline{N}$. That the mean number of detected photons is a constant is clearly false for nearly all objects; however, by using this approximation, we can develop some important relationships that are otherwise obscured in the summations.

Using the approximation $\overline{N}_{ij} \approx \overline{N}$,

$$\sigma^2(x,y) = \sigma_\mu^2 = \frac{\pi^2 T^2}{M^2 \overline{N}} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} \left[c\left(x\cos\theta_j + y\sin\theta_j - iT\right)\right]^2 . \tag{6.66}$$

For large N and M, we may make the approximation

$$\frac{\pi}{M} \sum_{j=1}^{M} T \sum_{i=1}^{N} \left[c\left(x\cos\theta_j + y\sin\theta_j - iT\right)\right]^2$$

$$\approx \int_0^{\pi} \int_{-\infty}^{\infty} \left[c\left(x\cos\theta + y\sin\theta - \ell\right)\right]^2 \, d\ell d\theta \tag{6.67}$$

$$= \pi \int_{-\infty}^{\infty} \left[c(\ell)\right]^2 \, d\ell \tag{6.68}$$

$$= \pi \int_{-\infty}^{\infty} |C(\varrho)|^2 \, d\varrho , \tag{6.69}$$

where the last equality follows from Parseval's theorem [see (2.96)]. For a rectangular window with bandwidth ϱ_0 applied to the ramp filter,

$$\pi \int_{-\infty}^{\infty} |C(\varrho)|^2 \, d\varrho = \pi \int_{-\varrho_0}^{\varrho_0} \varrho^2 \, d\varrho = \frac{2\pi\varrho_0^3}{3} . \tag{6.70}$$

Thus, for a rectangular windowed ramp filter and the approximation $\overline{N}_{ij} = \overline{N}$, the reconstructed image variance is independent of (x,y), and is given by

$$\sigma_\mu^2 \approx \frac{\pi T}{M} \frac{1}{\overline{N}} \frac{2\pi\varrho_0^3}{3} \tag{6.71}$$

$$\approx \frac{2\pi^2}{3} \varrho_0^3 \frac{1}{M} \frac{1}{\overline{N}/T} . \tag{6.72}$$

We now interpret this expression for image noise variance with the following observations:

- If we increase the bandwidth of the rectangular window, the image variance also increases. This follows one's natural intuition that noise has high frequency components, while the image bandwidth tends to diminish at high frequencies. We will see in emission tomography (Chapter 9) that the design of windowing functions that optimally balance the increase in noise contributions with the loss of image sharpness is very important.
- If we decrease T, the spacing between detectors, the variance decreases. Notice, however, that if the size of each detector is forced to decrease in order to do this, or if the detector efficiency decreases, then the decreased spacing will be offset by a lower $\overline{N}$.
- If we increase $\overline{N}$, the variance decreases. In general, one increases N_0 to do this. One could also increase $\overline{N}$ by increasing the incident x-ray energy, which lowers the amount of x-ray absorption. However, one would then have a loss of contrast, a subject that is considered in the following section on SNR.
- If we increase M, the variance decreases. Thus, the more angles the better. This assumes constant acquisition time per angle—it would not be true with constant total scan time. (However, increased M for constant total scan time could still provide benefits from improved angular sampling.)
- The fraction $\overline{N}/T$ indicates the average number of photons per unit distance along the detector array. In fan-beam geometries, we can increase this ratio, and hence decrease the image variance, by increasing N_0.

Image SNR Image variance is an important concept when we are trying to measure the linear attenuation coefficient. However, in radiology we must also consider the contrast between tissues, since this is what creates the visual effect of objects or structures with boundaries in a medical image. Therefore, any useful definition of SNR must contain both contrast and noise. We define SNR here as

$$\mathrm{SNR} = \frac{C\overline{\mu}}{\sigma_\mu}, \tag{6.73}$$

where C is the fractional change in μ from $\overline{\mu}$, $\overline{\mu}$ is the mean linear attenuation coefficient, and σ_μ is the standard deviation of the measurement (the square root of the variance). Equation (6.73) is thus the CT version of the differential SNR introduced in (3.69). Combining (6.72) with (6.73) gives

$$\mathrm{SNR} = \frac{C\overline{\mu}}{\pi}\varrho_0^{-3/2}\sqrt{\frac{3}{2}(\overline{N}/T)M}\,. \tag{6.74}$$

In a good CT scanner design, we set $\varrho_0 \approx k/d$, where $k \approx 1$ and d is the width of a detector. This gives

$$\mathrm{SNR} \approx 0.4kC\overline{\mu}d^{3/2}\sqrt{(\overline{N}/T)M}\,. \tag{6.75}$$

If, in fact, $d = T$ (as in a 3G scanner), then

$$\mathrm{SNR} \approx 0.4kC\overline{\mu}d\sqrt{\overline{N}M}\,. \tag{6.76}$$

In the fan-beam case,

$\overline{N}_f$: mean photon count per fan-beam,

D : number of detectors,

L : length of detector array,

and, therefore, $\overline{N} = \overline{N}_f/D$ and $d = L/D$, which yields

$$\mathrm{SNR} \approx 0.4kC\overline{\mu}LD^{-3/2}\sqrt{\overline{N}_f M}\,. \tag{6.77}$$

Therefore, in the fan-beam case we have the odd situation that increasing the number of detectors actually *lowers* the SNR. The reason, it turns out, is that convolution of the projections with the ramp filter couples the noise between detectors and does so to a greater extent as the number of detectors increases. Why shouldn't we reduce the number of detectors to, say, three or even one? Although such a system would give excellent SNR, we know from Section 6.4.1 that the resolution would be abysmal.

EXAMPLE 6.6

Consider a fan-beam CT system with one source, D detectors, M angles, and J by J reconstructed images, where $D = M = J = 256$. Assume that the width of each detector is $d = 0.25$ cm and the ramp filter uses a rectangular window with cutoff $\varrho_0 = 1/d$. The scanner is used to image a lesion with contrast $C = 0.005$ embedded in water ($\overline{\mu} = 0.15$ cm^{-1}).

Question We require the image to have a SNR of at least 20 dB. What is the minimum number of photons per projection at the detectors that is required in order to meet this SNR constraint?

Answer Since $\mathrm{SNR(dB)} = 20\log_{10}\mathrm{SNR}$, we have

$$\mathrm{SNR} = 10^{20\mathrm{dB}/20} = 10.$$

Since $\mathrm{SNR} = C\overline{\mu}/\sigma_\mu = 10$, we have

$$\sigma_\mu = \frac{0.005}{10} \times 0.15\ \mathrm{cm}^{-1} = 7.5 \times 10^{-5}\ \mathrm{cm}^{-1}.$$

Thus,

$$\sigma_\mu^2 = 5.625 \times 10^{-9}\ \mathrm{cm}^{-2} = \frac{2\pi^2}{3}\frac{\varrho_0^3 T}{M\overline{N}}.$$

The number of photons per projection is

$$P_p = \overline{N}D = \frac{2\pi^2}{3}\frac{\varrho_0^3 T}{M\sigma_\mu^2}D.$$

Since $\varrho_0 = 1/d$, $T = d = 0.25$ cm, and $D = M$, we have

$$P_p = \frac{2\pi^2}{3} \frac{1}{0.25^2} \frac{1}{\sigma_\mu^2} = 1.87 \times 10^{10} \text{ minimum}.$$

■

6.4.3 Artifacts

This section identifies and briefly explains some of the artifacts that can appear in CT images. Several of these are depicted in Figure 3.15.

Aliasing The projections will be aliased if they are undersampled. The effect on a projection is that higher frequency information will appear as lower frequency artifacts, as discussed in Section 3.6. These artifacts will continue through the process of reconstruction (CBP, Fourier method, etc.), and appear as artifacts in the image. The most readily apparent of these artifacts are *streak artifacts* which appear to emanate, in particular, from small bright objects within the image. An insufficient number of projections also causes aliasing. These artifacts also appear in the reconstructed images, often as streaks emanating from object boundaries at points which have small radii of curvature. These streaks may be either dark or bright, depending on the precise location of the objects and edges within the field-of-view.

Aliasing may be eliminated by sampling at a high enough (spatial sampling) rate, assuming the object is band limited, or by low-pass filtering prior to sampling. Both of these goals are achieved to some degree in current clinical scanners. High sampling rates are achieved by providing numerous detectors within the fan-beam and by sampling at numerous angles. Ultimately, the number of detectors is limited by physical size, efficiency, and the fact that smaller detectors receive fewer photons and, hence, have higher measurement variance. The number of angles is limited by storage capacity time (both to scan and reconstruct), and dose. Low-pass filtering (along each projection) is achieved by the physical aperture of each detector, which acts as a boxcar filter. There is no effective low-pass filtering in angle.

Another potential source for artifacts is in the backprojection-summation process, which must supply filtered projection values for a finite set of points (x_i, y_j) (pixels) in the plane. The critical issues are how many pixels will be used over the field-of-view and what will be the method of interpolation. Too few, and we produce an aliased version of the correct image, showing moiré patterns that appear as a structured textures that do not really exist. For such a coarse image, the projections should be low-pass filtered before backprojection.

A general rule of thumb for designing CT systems is that the number of detectors should be approximately equal to the number of projections that should be approximately equal to the number of points on the side of a reconstructed image. A typical 3G system has around 700 detectors, acquires 1,000 projections, and reconstructs a 512 × 512 image.

Beam Hardening *Beam hardening* is caused by energy-selective attenuation of the x-rays, as discussed in Section 5.2.2, and refers to the phenomenon in which the mean energy of the x-ray spectrum increases while propagating through the body. In the human body, the linear attenuation coefficient decreases with increasing E, as shown in Figure 4.8. Therefore, low-energy photons are preferentially absorbed. Hence, since all CT x-ray sources have a distribution (spectrum) of energies, the propagating beam becomes richer in high-energy photons.

We had concluded in Section 6.2 that the concept of effective energy could be used for polyenergetic sources. However, this concept assumes that the exit spectrum is the same as the entering spectrum. When this is not true, and standard reconstruction methods are used, the so-called *interpetrous lucency artifact* may appear. (It is called ‘‘interpetrous’’ because it often occurs in bone.) This artifact, which is particularly apparent in head scans (because of the outer ring of bone), creates a kind of halo effect around the brain parenchyma. In other parts of the body, streak artifacts, particularly at the tips of bones and at metal pieces within the body (shrapnel, surgical clips, screws, etc.) may also appear. Methods exist for compensating for or eliminating these artifacts, including preprocessing projection data, postprocessing images, and dual-energy imaging. In general, radiologists simply know to expect them and interpret the images accordingly.

Other Artifacts Three other sources of artifacts in CT images are worth mentioning: (1) electronic or system drift, (2) x-ray scatter, and (3) motion. The electronic or system artifact of particular interest is that which is caused by miscalibration or gain drift. The most dramatic effect is caused by the complete failure of a detector. In a 3G scanner, this situation (usually) causes the detector to report zero photons, and thus 100% attenuation at all angular positions. The resulting reconstruction contains a ring artifact whose radius is equal to the position of the defective detector from the detector array’s center.

X-ray scatter generally causes a convolutional blurring of each projection.[3] A blurred projection creates a blurry image. It is possible to prevent scatter by collimation and/or selective energy detection, but this comes at the expense of lower efficiency. It is also possible to compensate for scatter by deconvolution, but this has the tendency to further increase the noise in the projections. One way that has been used to measure the amount of scattering present is off-plane detectors.

Finally, motion is a difficult problem. Since a complete scan takes (typically) between 0.3 seconds (for a single slice) and 10 seconds (for a full thoracic scan), the heart is almost certainly going to have significant motion during the scan. Breathing is going to be a problem on the longer scans or when patients cannot hold their breath. It is possible to gate the data acquisition so that data will be taken only at a certain stage in the cardiac cycle and/or breathing cycle. The data acquired over successive cycles can then be pieced together and treated as if they were taken simultaneously. These methods are used on some machines in certain clinical situations where deemed necessary.

---

[3] This may have the desirable property of low-pass filtering the projection if aliasing is a problem.

6.5 Summary and Key Points

Computed tomography solves the problem of loss of contrast present in projection radiography that is caused by superposition of overlying structures, by focusing directly (and exclusively) on the cross-sectional image slice of interest. This strategy significantly improves contrast, at the expense of spatial resolution, but the gain in contrast is so important that CT is ubiquitous in medical imaging practice. In this chapter, we presented the following key concepts that you should now understand:

1. A *tomogram* is an image of a cross-sectional plane or slice within or through the body.
2. *X-ray computed tomography* (CT) produces tomograms of the distribution of linear attenuation coefficients, expressed in *Hounsfield units*.
3. There are currently seven *generations* of CT scanner design, which depend on the relation between the x-ray source and detectors, and the extent and motion of the detectors (and patient table).
4. The *basic imaging equation* is identical to that for projection radiography; the difference is that the ensemble of projections is used to reconstruct cross-sectional images.
5. The most common reconstruction algorithm is *filtered backprojection*, which arises from the *projection-slice theorem*.
6. In practice, the reconstruction algorithm must consider the *geometry* of the scanner—parallel-beam, fan-beam, helical-scan, or cone-beam.
7. As in projection radiography, *noise* limits an image's SNR.
8. Other *artifacts* include *aliasing*, *beam hardening*, and—as in projection radiography—*inclusion of Compton scattered photons*.

Further Reading

Carlton, R.R. and Adler, A.M. *Principles of Radiographic Imaging: An Art and a Science*, 5th ed. Clifton Park, NY: Delmar Cengage Learning, 2012.

Kak, A.C. and Slaney, M. *Principles of Computerized Tomographic Imaging*. New York, NY: IEEE Press, 1988.

Kalender, W.A. *Computed Tomography*, 3rd ed. Munich: Publicis MCD Verlag, 2011.

Hsieh, J. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. 2nd ed. Bellingham, WA: SPIE Press, 2009.

Macovski, A. "Physical Problems of Computerized Tomography." *Proceedings of the IEEE* 71, 1983: 373–78.

Macovski, A. *Medical Imaging Systems*. Englewood Cliffs, NJ: Prentice Hall, 1983.

Mahesh, M. *MDCT Physics: The Basics*, Philadelphia, PA: Lippincott Williams and Wilkins, 2009.

Wolbarst, A.B. *Physics of Radiology*, 2nd ed. Norwalk, CT: Appleton and Lange, 2005.

Zeng, G. *Medical Image Reconstruction: A Conceptual Tutorial*, Heidelberg: Springer, 2010.

Problems

Instrumentation

6.1 A CT calibration experiment measures two Hounsfield numbers: $h_m^W = 10$ for water and $h_m^A = -1,100$ for air.

(**a**) Find expressions for a and b so that

$$\begin{aligned} ah_m^W + b &= h^W , \\ ah_m^A + b &= h^A . \end{aligned}$$

(**b**) What are the correct CT numbers h^W and h^A for water and air, respectively?

(**c**) What are the values of a and b for this calibration experiment?

6.2 Suppose a 6G CT scanner has a patient table which moves at a speed of 2 cm/s. The x-ray source detector apparatus rotates at a speed of 4π radians per second. Also assume that it takes 1 ms to measure a projection.

(**a**) What is the pitch of the helix?

(**b**) How many projections does the system measure over a 2π angle?

(**c**) How long does it take to do a 60 cm torso scan?

Radon Transform

6.3 Show that the Radon transform is a linear operator.

6.4 Show that the Radon transform of $f(x - x_0, y - y_0)$ is $g(\ell - x_0 \cos\theta - y_0 \sin\theta, \theta)$.

6.5 Find the 2-D Radon transform $g(\ell, \theta)$ of $f(x, y) = \exp(-x^2 - y^2)$. (Hint: Use the rotational symmetry of $f(x, y)$ to simply your integration.)

6.6 Show that the function $h_\ell(\ell)h_\theta(\theta)$ cannot represent a 2-D Radon transform $g(\ell, \theta)$ unless $h_\theta(\theta)$ is a constant.

6.7 (**a**) Find the 2-D Radon transform $g(\ell, \theta)$ of the 2-D function $f(x, y) = \cos 2\pi f_0 x$. Show that *filtered* backprojection produces the correct reconstruction.

(**b**) Repeat part (a) for $f(x, y) = \cos 2\pi ax + \cos 2\pi by$ and $f(x, y) = \cos 2\pi(ax + by)$.

6.8 A first-generation CT scanner is used to image a unit-square shaped object (i.e, length of each side = 1). The object is surrounded by air and has a constant linear attenuation coefficient of μ_0. The coordinate system is set up such that the origin is at the object center, and the x- and y-axes are parallel to the sides of the object.

(**a**) Write a mathematical expression for the linear attenuation function $\mu(x, y)$. (Hint: Use the rect function.)

(**b**) What is the Fourier transform of $\mu(x, y)$?

(**c**) Write a mathematical relationship between the projection $g(\ell, \theta)$ (computed using the observed x-ray intensities) and $\mu(x, y)$.

(**d**) Using the projection-slice theorem, find $G(\varrho, \theta)$.

(e) Take the inverse Fourier transform of $G(\varrho, \theta)$ to find an expression for $g(\ell, \theta)$.

(f) Sketch $g(\ell, 30^\circ)$ (include axis labels) and sketch its backprojection image $b_{30^\circ}(x, y)$.

6.9 Prove the convolution property of the radon transform:

$$\mathcal{R}\left\{f * h)\right\} = \mathcal{R}\{f\} * \mathcal{R}\{h\}.$$

CT Reconstruction

6.10 The "unit square function" is given by

$$s(x, y) = \begin{cases} 1 & -1 \leq x \leq 1, -1 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Let $g_s(\ell, \theta)$ be the 2-D Radon transform of $s(x, y)$.

(a) Show that $g_s(\ell, \theta + \pi/2) = g_s(\ell, \theta)$.

(b) Show that $g_s(\ell, -\theta) = g_s(\ell, \theta)$.

(c) Write an expression for $g_s(\ell, \theta)$, $-\infty < \ell < +\infty$, $0 \leq \theta < \pi$ given knowledge of $g_s(\ell, \theta)$ only in the range $-\infty < \ell < +\infty$, $0 \leq \theta < \pi/4$.

(d) Sketch $g_s(\ell, \theta)$ at $\theta = 0$, $\theta = \pi/8$, and $\theta = \pi/4$.

(e) Determine an expression for $g_s(\ell, \theta)$, $0 \leq \theta < \pi/4$.

6.11 The *mass* of an object is defined as

$$m = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy$$

and the *center-of-mass* is defined as $\mathbf{c} = (c_x, c_y)$ where

$$c_x = \frac{1}{m} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y)\, dxdy\,,$$
$$c_y = \frac{1}{m} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y)\, dxdy\,.$$

The mass of a projection is defined as

$$m_p(\theta) = \int_{-\infty}^{\infty} g(\ell, \theta)\, d\ell$$

and its center-of-mass is given by

$$c_p(\theta) = \frac{1}{m_p(\theta)} \int_{-\infty}^{\infty} \ell g(\ell, \theta)\, d\ell\,.$$

(a) Show that $m_p(\theta) = m$.

(b) Show that $c_p(\theta) = c_x \cos\theta + c_y \sin\theta$.

(c) Define the *triangle function* as

$$f(x,y) = \begin{cases} 1 & 0 \le y \le 1 - |x| \\ 0 & \text{otherwise} \end{cases} .$$

Find $m_p\left(\frac{\pi}{4}\right)$ and $c_p\left(\frac{\pi}{4}\right)$ for the triangle function.

6.12 Consider an object comprising two small metal pellets located at $(x, y) = (2, 0)$ and $(2, 2)$ and a piece of wire stretched straight between $(0, -2)$ and $(0, 0)$.

(a) Sketch this object.

Assume that N photons are fired at each lateral position ℓ in a parallel-ray configuration. For simplicity, assume that each metal object stops 1/2 the photons that are incident upon it no matter what angle it is hit.

(b) Sketch the number of photons you would expect to see as a function of ℓ for $\theta = 0°$ and $\theta = 90°$.

(c) Draw the projections you would see at $\theta = 0°$ and $\theta = 90°$.

(d) Sketch the backprojection image you would get at $\theta = 0°$ (without filtering).

6.13 Consider the object in Figure P6.1 that is an equilateral triangle centered at the origin, each side of which is of length a and one side of which is parallel to the x axis. Suppose the object has constant linear attenuation coefficient μ and is being imaged in a 1G CT scanner. Assume $\mu = 1$ and $a = 6$.

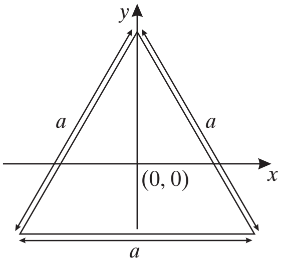


Figure P6.1
Object geometry for Problem 6.13.

(a) Find a formula for the projection $g(\ell, 60°)$ and sketch it.

(b) What is the value of $b_{60°}(0, a/4)$ for $g(\ell, 60°)$?

(c) Find $F(\varrho \cos 60°, \varrho \sin 60°)$.

6.14 In finding an approximation $\tilde{c}(\ell)$ for the ramp filter in filtered backprojection, you decide to multiply $|\varrho|$ by the triangular window of width ϱ_0 given by

$$W(\varrho) = \begin{cases} 1 - \frac{|\varrho|}{\varrho_0} & |\varrho| \le \varrho_0 \\ \\ 0 & \text{otherwise} \end{cases} .$$

(a) Derive an expression for the approximate filter $\tilde{c}(\ell)$.

(b) Describe what happens as $\varrho_0 \to \infty$.

6.15 Show that

$$\int_0^{2\pi}\int_0^{\infty} \varrho G_\theta(\varrho)e^{j2\pi\varrho\omega\cdot x}\, d\varrho\, d\theta = \int_0^{\pi}\int_{-\infty}^{\infty} |\varrho| G_\theta(\varrho)e^{j2\pi\varrho\omega\cdot x}\, d\varrho\, d\theta.$$

6.16 In this problem, we want to figure out what is "wrong" with the backprojection summation image given by

$$f_b(x,y) = \int_0^{\pi} g(x\cos\theta + y\sin\theta, \theta)d\theta\,,$$

and see how to fix it.

(**a**) Show that the 2-D Radon transform of the 2-D delta function $\delta(x,y)$ is $\delta(\ell)$. (Hint: Use the projection slice theorem.)

(**b**) Show that the backprojection summation image of $\delta(x,y)$ is $\dfrac{1}{\sqrt{x^2+y^2}}$. (Hint: Write (x,y) in polar coordinates (r,ϕ), then use the fact that $\sin\theta \approx \theta$ for small θ and the fact that $\delta(a\ell) = \delta(\ell)/|a|$.)

(**c**) Show that the backprojection summation image of $\delta(x-x_0, y-y_0)$ is $1/\sqrt{(x-x_0)^2+(y-y_0)^2}$—that is, that the backprojection summation operator is shift invariant.

(**d**) Show that $f_b(x,y) = f(x,y) * (1/\sqrt{x^2+y^2})$.

(**e**) Describe how one could, in principle, recover $f(x,y)$ from $f_b(x,y)$. Describe any problems you may foresee with your approach.

6.17 With each measured projection $g(\ell,\theta)$ (viewed as a function of ℓ) a backprojection image $b_\theta(x,y)$ can be generated.

(**a**) Write down a formula for $b_\theta(x,y)$.

Now assume that we know only one projection: $g(\ell, 30^\circ) = e^{-|\ell|}$.

(**b**) What is the value of $b_{30^\circ}(1,2)$?

(**c**) Can you determine $b_{45^\circ}(1,2)$? If yes, what is it? If no, why not?

(**d**) Can you determine $b_{210^\circ}(1,2)$? If yes, what is it? If no, why not?

(**e**) Sketch $b_{30^\circ}(x,y)$.

Suppose $g(\ell,\theta)$ is sampled so that only the following values are known: $g(nT,\theta)$, $-\infty < n < \infty$.

(**f**) Can you determine $b_{30^\circ}(1,2)$ from the sampled projection $g(n, 30^\circ)$? If yes, what is it? If no, say why not and suggest an approximate value.

(**g**) Can you determine $b_{30^\circ}(2,1)$ from the sampled projection $g(n, 30^\circ)$? If yes, what is it? If no, say why not and suggest an approximate value.

6.18 For fan beam geometry, prove the following relationship is satisfied by the ramp filter $c(\cdot)$:

$$c(D'\sin\gamma) = \left(\frac{\gamma}{D'\sin\gamma}\right)c(\gamma)\,.$$

Image Quality

6.19 In a CT system, each projection is actually obtained using a uniform scanning beam of width W instead of an infinitesimal pencil beam. Find the resultant estimate $\hat{f}(x,y)$ of the function $f(x,y)$ when CBP is used.

6.20 Find the SNR of the CT reconstruction of a lesion immersed in a 20-cm (diameter) cylinder of water whose attenuation coefficient is 5% different than that of the water. A scanned source is used providing 100 projections at 0.1 R (roentgen) per projection. The detector and beam dimensions are 2.0×2.0 mm. Make appropriate assumptions about the reconstruction filter. Assume $\overline{\mu} = 0.15\ \text{cm}^{-1}$ and that the detectors are touching each other. Also assume that there are 2.5×10^{10} photons/cm^2 per roentgen.

6.21 Joe is designing a new CT scanner. It will be a fan-beam design, with one source and D detectors. He will use M angles, reconstruct images with N by N pixels, and strictly enforce the design rule of thumb $D = M = N$. The width of each detector d will be selected to be as large as possible in order to completely fill a 1 meter detector array with D detectors. The reconstruction ramp filter will be designed using a rectangular window with bandwidth $\varrho_0 = 1/d$.

(**a**) Joe requires that a lesion with contrast $C = 0.005$ embedded in water ($\overline{\mu} = 0.15\ \text{cm}^{-1}$) will have a SNR of at least 20 dB. Suppose $D = M = N = 300$. What is the minimum number of photons per projection at the detectors that is required in order to meet this SNR constraint.

(**b**) Assume the patient has 0.125 m^2 cross-sectional area (lying completely in each fan beam). Assume that the patient is exposed to 1 roentgen when 2.5×10^{10} photons pass through 1 cm^2 of tissue, and assume that the dose to the patient in rads is equal to the exposure in roentgens. What is the largest $D(= M = N)$ that can be selected which will satisfy the SNR requirement of part (a) and will keep the total dose for the entire cross-sectional scan less than or equal to 2 rads. (Neglect attenuation due to the patient and assume the detectors have 100% efficiency.)

6.22 Your problem is to determine the number of detectors D to fit into an array of fixed length L in order to optimize the SNR of a third-generation CT scanner.

(**a**) Assume the approximate ramp filter is $\tilde{c}(\ell) = \mathcal{F}^{-1}\{|\varrho| W(\varrho)\}$, where $W(\varrho)$ is a rectangular window with (single-sided) bandwidth ϱ_0. Assume that $\varrho_0 = \min\{d^{-1}, \varrho_{\max}\}$, where d is the width of a detector and $\varrho_{\max}$ is a fixed constant. Assume that the number of projections M is given by $1.5D$. Find an expression for the SNR as a function of D.

(**b**) To avoid aliasing a $J \times J$ pixel reconstructed image it is decided to choose $\varrho_{\max} = J/(2L)$. We also restrict the number of detectors to lie in the range $1 \le D \le J$. What D gives the best SNR?

6.23 Suppose we have a first-generation CT scanner with a mechanical problem. For each projection angle, it measures zero at $\ell = 0$. Now we use the scanner to image a uniform disk.

(**a**) Sketch the sinogram.

(**b**) What would the reconstructed image look like?

(**c**) If instead the scanner always zeroes the measurement at $\ell = \ell_0$, what would the reconstructed image look like?

Applications and Advanced Topics

6.24 Suppose we know that $f(x,y) = \sum_{j=1}^{n} f_j \phi_j(x,y)$. Define the ith line integral as $g_i = \int_{L_i} f(x,y)ds, \quad i = 1,\ldots,m$, and the ith measured line integral as $y_i = g_i + v_i, \quad i = 1,\ldots,m$, where v_i is noise. Now define the following vectors $f = [f_1 \cdots f_n]^{\mathrm{T}}$, $y = [y_1 \cdots y_m]^{\mathrm{T}}$, and $v = [v_1 \cdots v_m]^{\mathrm{T}}$. Then in matrix notation, our measurements are given by $y = Hf + v$, where H is an $m \times n$ matrix.

(**a**) Determine H_{ij}, the (i,j)th entry in H.

(**b**) In an ideal world we would find that $v = 0$, $m = n$, and H^{-1} exits In this case, $f = H^{-1}y$ gives an exact reconstruction of f (and we never even spoke of the Radon transform!). Discuss the meaning of the following situations:

(**1**) H^{-1} exists but $v \neq 0$.

(**2**) $v = 0$ but $m < n$.

(**3**) $v = 0$ but $m > n$.

(**c**) Suppose $m > n$ and assume that the rank of H is n. Find an expression for $\hat{f}$, the vector that minimizes $E = (y - Hf)^{\mathrm{T}}(y - Hf)$. (Hint: Multiply out E and take the derivative with respect to each component of f and set each to zero. Organize your answer using matrix notation and solve for f.)

(**d**) In a typical system we may have 256 × 256 pixels in a reconstructed image and 360 × 512 line integral measurements. Given these sizes, explain the main difficulty with the above approach to image reconstruction.

6.25 (**a**) Find the 2-D Radon transform $g_\theta(\ell)$, for $0 \leq \theta \leq \pi/4$ only, of the unit square indicator function:

$$f(x,y) = \begin{cases} 1 & -1/2 \leq x,y \leq 1/2 \\ 0 & \text{elsewhere} \end{cases} .$$

(**b**) Sketch $g_\theta(\ell)$ for some θ, where $0 < \theta < \pi/4$.

(**c**) Find $\int_{-\infty}^{\infty} g_\theta(\ell)\, d\ell$.

(**d**) Suppose only two projections, for $\theta = 0$ and $\theta = \pi/2$, are available. Find an expression for the approximate summation backprojection image $\hat{f}_b(x,y)$ (given only these two images) and make a sketch of the result.

(**e**) In general, is it possible to reconstruct $f(x,y)$ perfectly given only a finite number of projections? Explain your answer.

6.26 A first-generation CT scanner is used to image the object shown in Figure P6.2 (surrounded by air). Suppose for $E > 100$ keV

$$\mu_1(E) = 1.0 \exp -E[\text{keV}]/100[\text{ keV}] \ \text{cm}^{-1},$$

$$\mu_2(E) = 2.0 \exp -E[\text{ keV}]/100[\text{ keV}] \ \text{cm}^{-1}.$$

Suppose that each incident photon burst is polychromatic, but has only two photon energies, and that there are 10^6 photons at 100 keV, and 0.5×10^6 at 140 keV.

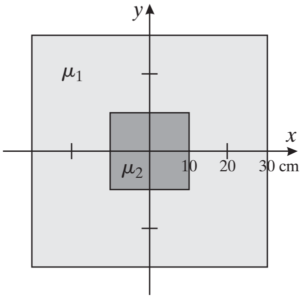


Figure P6.2
Object geometry for Problem 6.26.

(**a**) Calculate the incident intensity of the x-ray burst.

(**b**) Ignore the dimension of the detector in this part. Find the intensity I_d as a function of x for the measured projection at $\theta = 0°$. Sketch it.

(**c**) Calculate the local contrast of the middle square given the measured projection $I_d(x)$. What is the local contrast of the projection $g(\ell, 0)$?

(**d**) Suppose the detector has a width of 1 cm. Find the measured projection $I_d(x)$ (x being a continuous variable) in this case, and sketch it. Will this change affect the local contrast? Explain.

6.27 A CT system is designed to acquire M parallel-ray projections of $f(x, y)$ at angles $\theta_i = \pi i/M, i = 0, \ldots, M-1$. The collection of projections measured by this system (modeled to be continuous in ℓ) is defined to be the *M-projection Radon transform*:

$$\mathcal{D}_M f = \{g(\ell, \theta_i) | \; i = 0, \ldots, M-1\},$$

where $g(\ell, \theta_i) = \int\int f(x, y)\delta(x \cos\theta_i + y \sin\theta_i - \ell)dxdy$ is the projection at angle θ_i.

(**a**) Draw a picture showing the four positions of the detector array required to determine $\mathcal{D}_4 f$. Label which projection is acquired at each position.

(**b**) Knowledge of $\mathcal{D}_M f$ gives partial knowledge of the Fourier transform $F(u, v)$ of $f(x, y)$. Draw a picture showing the set of points in the u-v plane where $F(u, v)$ is known given $\mathcal{D}_4 f$. Indicate which projection gave rise to each subset.

(**c**) Now let $G_1 = \mathcal{D}_M f_1$. Define $G_2 = \mathcal{D}_M(f_1 + f_2)$, where $f_2 = \cos 2\pi f_x x \cos 2\pi f_y y$, $f_x = \cos(3\pi/16)$ and $f_y = \sin(3\pi/16)$. Show that $G_2 = G_1$ for $M = 4$.

(**d**) For f_2 given above, find the minimum value of M in order to guarantee that $G_2 \neq G_1$.

(**e**) The result of part (c) means that you can add a function to the original without changing its M-projection Radon transform. In this context, such functions are called a *ghost functions*. In theory, is it possible to make M large enough so that there are no ghost functions of $\mathcal{D}_M f$? Explain your answer. What happens in practice?

(**f**) Suppose one used large detectors, effectively low-pass filtering each projection upon observation. Would f_2 still be a ghost function? Explain.

6.28 Consider an object comprising three squares of the same size with width 20 cm, as shown in Figure P6.3. The origin of the coordinate system is located at the center of the middle square. The linear attenuation coefficient in these three regions are $\mu_1 = 0.1\ \text{cm}^{-1}$, $\mu_2 = 0.2\ \text{cm}^{-1}$, and $\mu_3 = 0.3\ \text{cm}^{-1}$, respectively.

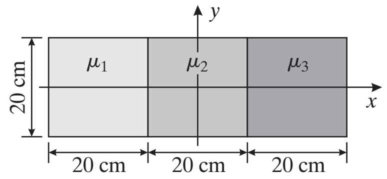


Figure P6.3
Object geometry for Problem 6.28.

Assume that parallel ray geometry is used.

(a) Find $g(\ell, 0^\circ)$.

(b) Find $g(\ell, 90^\circ)$.

(c) Find $g(\ell, 45^\circ)$. Sketch it and carefully label the axes.

(d) Sketch $b_{45^\circ}(x, y)$. Find the value $b_{45^\circ}(1, 1)$.
Now, assume we are using a 3G CT scanner. The source–detector distance is 1.5 m.

(e) What is the smallest possible circular FOV to image the entire object shown in the figure? What is the shortest length of the detector array that will "cover" the FOV?

(f) Suppose the detector array has 256 elements. Using the CT "rule of thumb," how many angles should be acquired? What is the pixel size of the reconstructed image (assuming the image covers the entire FOV.)

6.29 A square object with side-width of 40 cm is imaged by a first generation CT scanner.

(a) What type of collimators should be used for the source and the detector?

(b) What is the smallest circular FOV to cover the entire object?

(c) Suppose the smallest angular increment of the scanner is 0.25° and we strictly obey the CT "rule of thumb." How many line integrals should we measure for each angle? If the reconstructed image is just large enough to cover the FOV, what is the size of the pixels?

(d) We found that the reconstructed image was noisy. Is it possible to double the SNR given that we cannot change the 0.25-degree angular increment and will not violate the rule of thumb? Explain.

(e) Suppose the object has a constant linear attenuation coefficient of $0.1\ \text{cm}^{-1}$, sketch $g(\ell, 45^\circ)$ and carefully label the axes.

(f) Sketch $b_{45^\circ}(x, y)$. Find the value $b_{45^\circ}(10, 10)$.

6.30 You are designing a CT scanner. You always use the rules $M = D = J$ and $\varrho_0 = 1/d$. You want a circular field-of-view (FOV) having a 60 cm diameter, so your image will be a square with 60 cm sides.

(a) You want your system to be able to resolve two point sources separated by 1 mm, to that at least one square pixel can always be put between them no matter what their orientation relative to the pixel area. What is the minimum number of detectors you must use?

You decide to use 925 detectors, each 0.8 mm wide and fit side-by-side on a line. You decide to place the x-ray source 180 cm away from the center of the detector array (orthogonal to the array, of course).

(**b**) Will your FOV fit within the fan? Explain your answer.

Suppose $\overline{\mu} = 0.2\ \text{cm}^{-1}$ and suppose there are an average of 1.5×10^{11} photons hitting the detector array for each projection.

(**c**) What is the approximate SNR in dB of a tumor mass whose linear attenuation coefficient is 0.25cm^{-1}.

6.31 You have the good fortune to acquire a used 3G CT scanner that has slip ring technology (continuously rotating 360-degree x-ray tube and detector array). Everything is in good working order except that only one detector channel is operational. Fortunately, you are able to switch this channel between detectors so that you can acquire data from any detector, but from only one detector for each pulse of the x-ray tube. The tube can be pulsed continuously at a rate of one pulse per millisecond. The gantry can rotate at one revolution per second. The geometry of the fan beam is shown in Figure P6.4.

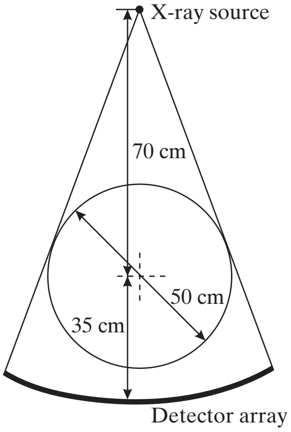


Figure P6.4
Scanner geometry for Problem 6.31.

(**a**) Determine the fan angle from the depicted geometry.

(**b**) Suppose 703 detectors are packed into the array depicted. Determine the spacing between the detectors.

(**c**) Suppose data from only the central detector is acquired (as fast as possible) over a single counterclockwise rotation of the gantry (from the x-ray source on the top back to the x-ray source on the top). What line integrals are acquired? Draw the trajectory of data acquired over this one acquisition cycle on a sinogram diagram. Are any line integrals repeated? Explain.

(**d**) Now suppose that data from only the bottom-left detector (with reference to Figure P6.4) is acquired, again over a single counterclockwise rotation of the gantry with the fastest possible rotation and x-ray tube pulsing. What line integrals are acquired this time?

Draw the trajectory on a sinogram diagram and explain whether there are repeated line integrals or not.

(e) Devise a strategy based on the ideas in (c) and (d) that will scan a sinogram without redundancy. In order to approximately obey the "CT rule of thumb," how many rotations will be required to scan the sinogram? Explain.

(f) What is so bad about this CT scanner that it should never be used to scan patients?

6.32 Object motion can cause artifacts in CT images. This happens in part because object motion produces an observation that is not a legitimate Radon transform of any object. Assume parallel ray geometry.

(a) Prove that the Radon transform of $\delta(x, y)$ is $\delta(\ell)$? Sketch its sinogram.

(b) Prove the following theorem:

$$\mathcal{R}f(x - x_0, y - y_0) = g(\ell - x_0 \cos\theta - y_0 \sin\theta, \theta).$$

(c) Find the Radon transform of $\delta(x - 1, y)$. Sketch its sinogram.

We now assume that while the scanner acquires data over $\theta \in [0, \pi/2]$ the object is $\delta(x, y)$ and while the scanner acquired data over $\theta \in [\pi/2, \pi)$ the object is $\delta(x - 1, y)$. In other words, the object *moved* during the scan from $(0, 0)$ to $(1, 0)$.

(d) Sketch the sinogram that is acquired by this scan.

(e) Prove the following:

$$\int_{-\infty}^{\infty} \ell g(\ell, \theta) d\ell = q_x \cos\theta + q_y \sin\theta\,,$$

where

$$q_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x, y) dx dy\,,$$

$$q_y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x, y) dx dy.$$

(f) Show that the sinogram acquired by the scanner when the object moved cannot be the Radon transform of *any* object.

6.33 The plot in Figure P6.5(a) shows the input energy spectrum of a polychromatic x-ray source and the linear attenuation coefficient of a material as a function of photon energy.

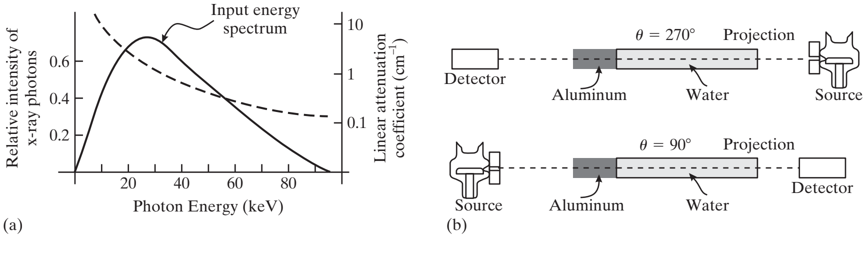


Figure P6.5
Problem 6.33.

(a) Sketch the energy spectrum of the x-ray beam after it passes through the material. Explain your sketch.

(b) Suppose that an object is imaged using a 1G CT scanner. In Figure P6.5(b) there are two different positions of the source and detector while imaging this same object. Explain why the two measurements should ideally be the same but are actually different in practice.

(c) The fact that the two measurements are different causes artifacts in CT reconstruction. One way to reduce these artifacts is to acquire two measurements with two different input x-ray spectra, one having higher mean energy than the other. Describe how you might obtain these two measurements using the same x-ray tube.

(d) This so-called *dual-energy CT* setup requires calibration. Four measurements are made on two test objects (Figure P6.6) using the two input spectra, yielding measurements given in the table below. The values in the table are line integrals (not x-ray intensities or numbers of photons). Explain why g_{c1}^{L} is greater than g_{c1}^{H}.

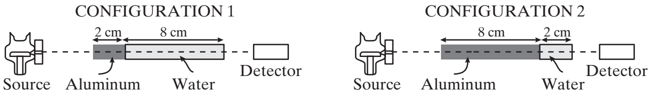


Figure P6.6
Problem 6.33.

	Configuration 1	Configuration 2
Higher photon energy (keV)	$g_{c1}^{H} = 2.2$	$g_{c2}^{H} = 4.3$
Lower Photon energy (keV)	$g_{c1}^{L} = 3.16$	$g_{c2}^{L} = 6.79$

(e) Suppose you had access to a 75 keV (monoenergetic) x-ray source. Given $\mu(\text{aluminum, 75 keV}) = 0.7\ \text{cm}^{-1}$ and $\mu(\text{water, 75 keV}) = 0.1866\ \text{cm}^{-1}$, find the line integrals, g_{c1}^{75} and g_{c2}^{75}, for each of the two configurations.

(f) For calibration, coefficients are to be determined so that measurements using the high and low spectra can be combined in a linear equation to yield what is expected from a monoenergetic 75 keV x-ray source. Given your result from part (e), find the coefficients a^{L} and a^{H} such that

$$g_{c1}^{75} = a^{L} g_{c1}^{L} + a^{H} g_{c1}^{H}\,,$$
$$g_{c2}^{75} = a^{L} g_{c2}^{L} + a^{H} g_{c2}^{H}\,.$$

(g) For an arbitrary object, your dual-energy CT scanner will obtain two sinograms $g^{L}(\ell,\theta)$ and $g^{H}(\ell,\theta)$. Write a mathematical expression for the reconstructed object, $\mu(\text{object, 75 keV})$, given these sinograms.

PART

III

Nuclear Medicine Imaging

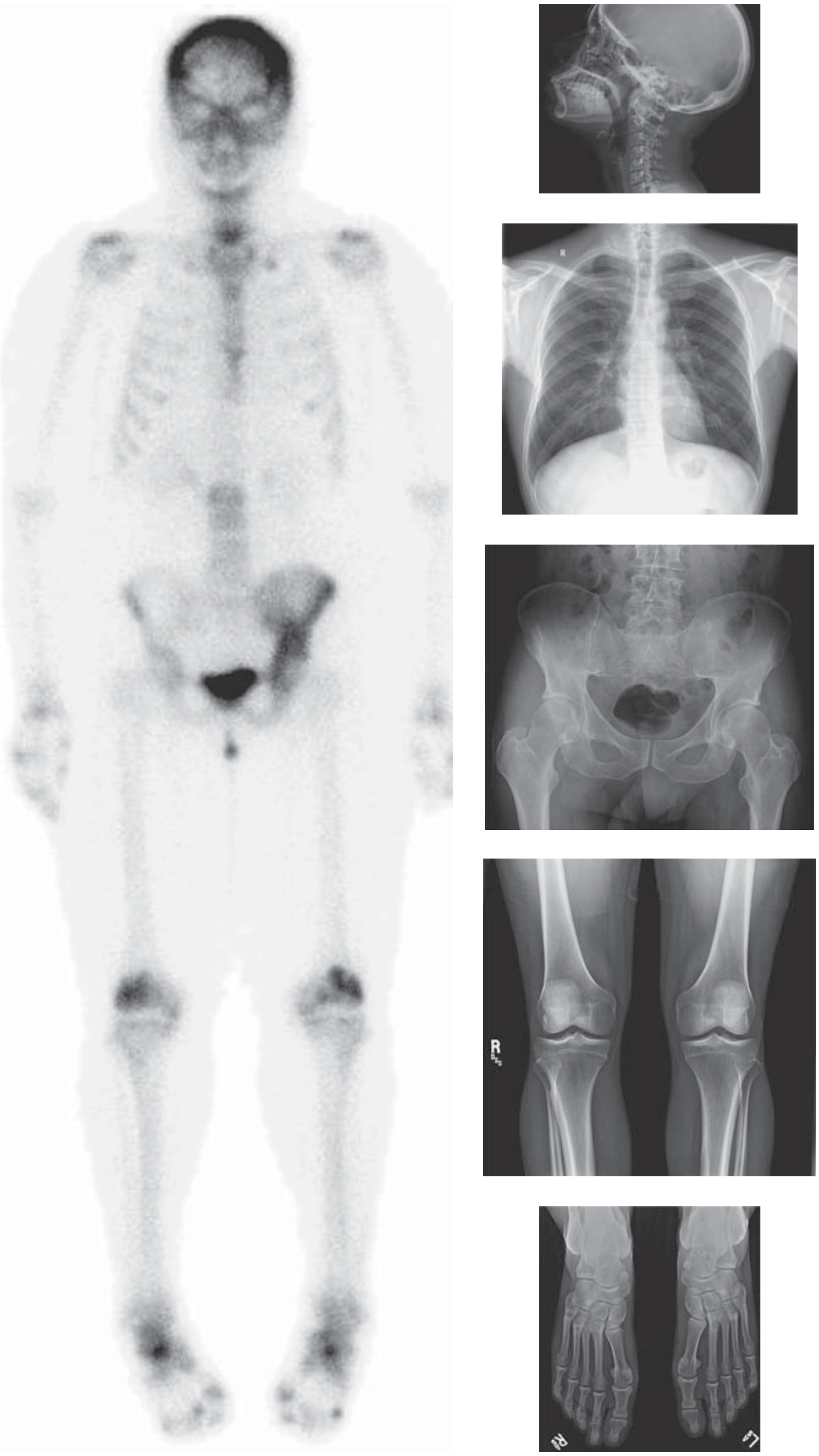
Overview

In the preceding part of the book, we considered projection radiography and computed tomography, two imaging modalities that rely on the *transmission* of photons through the body to form images. We now turn our attention to nuclear medicine, an imaging modality that relies on the *emission* of photons from within the body.

In contrast to projection radiography and computed tomography, in nuclear medicine, which we cover in this part of the book, the biological behavior of a substance's biodistribution in the body is of interest; each molecule of the substance is labeled with a radioactive atom. Here, the ionizing radiation emitted when the radioactive atom undergoes radioactive decay is used to determine the location of the molecule within the body; the ionizing radiation is of no medical interest per se. Since the biodistribution of the radiolabeled substance—the *radiopharmaceutical* or *radiotracer*—is determined by the body's physiological and biochemical functioning, nuclear medicine is considered a *functional* imaging modality.

Figure III.1 makes this point. Figure III.1(a) shows a whole-body bone scan. In this anterior view, image intensity is proportional to the distribution of a radiotracer that reflects the metabolic activity of bone. (The radiotracer is ultimately metabolized via the kidneys, which is why there is a *hot spot* of activity in the urinary bladder in the middle of the image.) Figure III.1(b) shows a stacked series of separate projection radiographs of the corresponding areas (from different patients). In these images, as discussed in the previous part of the book, image intensity reflects the varying absorption of transmitted x-rays through bone and other tissues.

In general, a radiotracer is injected into a peripheral arm vein of the patient, or the patient inhales or ingests the radiotracer. Specialized instrumentation produces images of the internal distribution of radioactivity, which is assumed to mirror the distribution of the compound of interest. These images are compared with known distributions in different disease states. Because there are hundreds



(a) (b)

Figure III.1 Representative (a) emission image and (b) transmission images. Image (a) courtesy of Harvey Ziessman, MD, Johns Hopkins University and Hospital. Images in (b) courtesy of GE Healthcare.

of different radiotracers routinely available in nuclear medicine, there are literally hundreds of different nuclear medicine studies, each of which assesses the function of a different physiologic process or organ system within the body. The technology of nuclear medicine permits direct measurements to be made of body processes in humans that in the past could be examined only in experimental animals. In order to most accurately depict the biodistribution of radiotracers, instrumentation that emphasizes high image quality and quantitative accuracy has been developed. Digital image processing plays an important role in nuclear medicine, not only in image enhancement but also in extraction of quantitative information about physiological function.

Nuclear medicine is used whenever a physician needs information on physiologic or biochemical function. For example, the two most common nuclear medicine procedures are bone scanning and myocardial perfusion imaging. Bone scanning, as shown in Figure III.1(a), looks at the metabolic activity of bones; this is complementary to the structural (anatomical) information from a projection radiograph (e.g., Figure III.1(b)). A projection radiograph can show a fracture; a bone scan can show active metabolism during the healing process. Similarly, coronary angiography (a projection radiography technique), which depicts the anatomy of the coronary arteries, shows the vessels that supply blood to the heart muscle, whereas myocardial perfusion imaging shows the distribution of blood flow in the muscle. Figure III.2 illustrates the use of nuclear medicine to depict myocardial perfusion. The images show the distribution of a specific radiotracer that distributes in heart muscle according to blood flow. The left ventricle of the heart has the greatest blood flow and is by far the most prominent

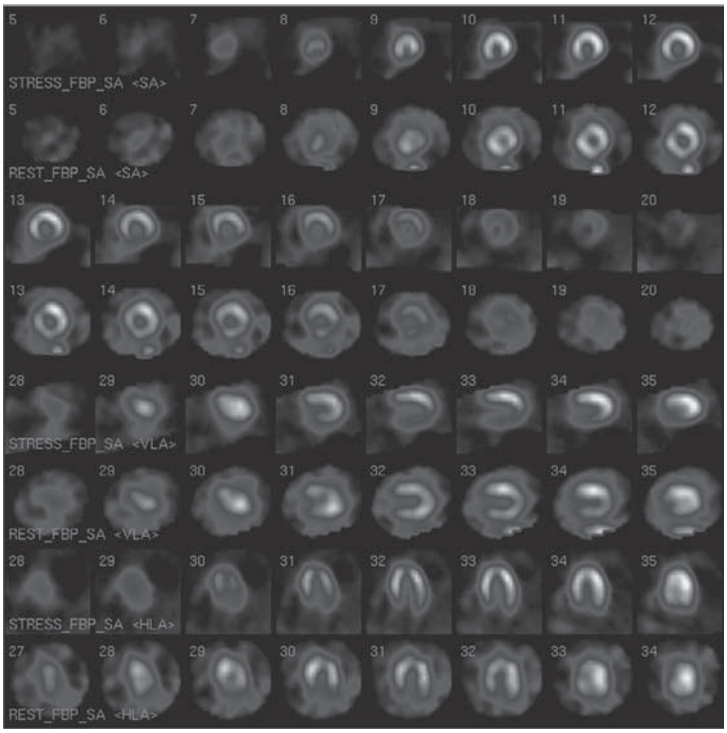


Figure III.2
Myocardial perfusion imaging via SPECT. Top four rows: short axis slices. Next two rows: horizontal long axis slices. Bottom two rows: vertical long axis slices. Image courtesy of Harvey Ziessman, MD, Johns Hopkins University and Hospital.

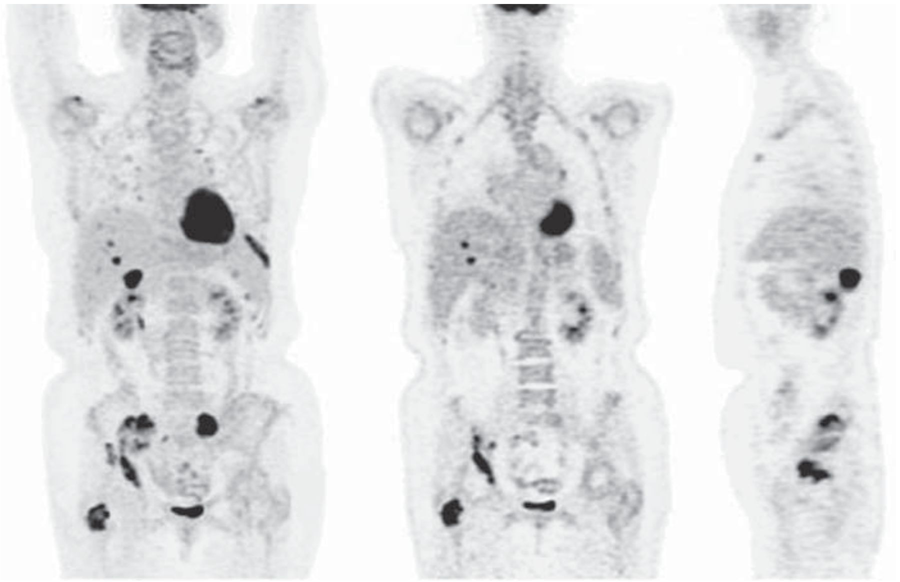


Figure III.3
Whole-body FDG-PET. Left: anterior reprojected view. Middle: posterior reprojected view. Right: left lateral reprojected view. Images courtesy of Harvey Ziessman, MD, Johns Hopkins University and Hospital.

structure in the images. The images represent tomographic slices from a 3-D *single-photon emission computed tomography* (SPECT) study. In contrast, the image in Figure III.1(a) is a projection image, which in nuclear medicine is often called a *planar* image or *planar scintigraphy*.

In the same way that computed tomography is the 3-D tomographic modality based on projection radiography, SPECT is the 3-D nuclear medicine modality based on planar scintigraphy. In Figure III.2, the first four rows are *short axis* slices (as though the heart were sliced like a bread loaf), the next two rows are *horizontal long axis* slices (orthogonal to the short axis), and the final two rows are *vertical long axis* slices (orthogonal to the short axis and also to the horizontal long axis). These slices thus represent orientations relative to the heart, not the long axis of the (entire) body.

Arguably the most advanced form of nuclear medicine imaging is *positron emission tomography* (PET), which utilizes specialized radiotracers and instrumentation. Figure III.3 depicts a whole-body PET scan (in anterior, posterior, and left lateral reprojection views) of the distribution of fluorine-18-labeled deoxyglucose (FDG), whose distribution reflects glucose metabolism. FDG-PET is the most commonly performed PET study with a focus on cancer detection and staging. These reprojection views are common in FDG-PET and are formed from the 3-D reconstructed dataset.

The Physics of Nuclear Medicine

CHAPTER 7

7.1 Introduction

Nuclear medicine relies on *radiopharmaceuticals* introduced into the body to trace the spatial and temporal distribution of the underlying physiological and biochemical processes that govern the radiopharmaceutical's distribution. Radiopharmaceuticals are usually introduced in very small quantities that do not themselves affect the physiologic or biochemical processes. Radiopharmaceuticals are thus often called *radiotracers*, reflecting both the fact that they are introduced into the body in trace quantities and that they trace body processes. Radiopharmaceuticals consist of both a chemical compound (the "pharmaceutical" portion of the word) and a radioactive atom (the "radio" portion of the word). When the radioactive atom undergoes *radioactive decay*, a process we describe below, radiation is emitted that leaves the body. External imaging devices, such as *scintillation cameras*, record these radiation emissions coming from the patient and produce either a planar, two-dimensional image or cross-sectional, tomographic images.

Depending on the specific radiotracer, different physiological or biochemical functions are imaged. This is strikingly different than projection radiography or computed tomography. In these two *transmission* imaging modalities, while the specific characteristics of the signal depend on specific imaging and instrument parameters, the basic type of information does not change from image to image; it depends on tissue attenuation characteristics, as we have discussed. In nuclear medicine, however, each different radiotracer produces a depiction of a completely different physiological or biochemical function, so the basic information itself is different. Further, since the radiation that forms the basis of the image is not transmitted through the body, but arises from the radioactive atoms within the body, nuclear medicine imaging is referred to as *emission* imaging.

In this chapter, we consider the physical processes that give rise to radionuclides, radioactive decay, and the emissions that form the basis of nuclear medicine imaging.

7.2 Nomenclature

Recall from Chapter 4 that an atom consists of a nucleus of protons and neutrons, which together are called *nucleons*, surrounded by orbiting electrons. The *atomic number* Z is equal to the number of protons in the nucleus and defines the element. The *mass number* A is equal to the number of nucleons in the nucleus. The term *nuclide* refers to any unique combination of protons and neutrons that forms a nucleus. If a particular nuclide is radioactive (i.e., it can undergo radioactive decay), it is termed a *radionuclide*. Nuclides are typically denoted by either ${}^{A}_{Z}X$ or X-A, where X is the element symbol.

Atoms with the same atomic number but different mass number (i.e., different numbers of neutrons) are called *isotopes*. For example, carbon-11 is an isotope of carbon that decays by positron decay and is used in positron emission tomography, as discussed in Chapter 9. Because isotopes have the same number of protons (and hence electrons), they are chemically identical. Atoms with the same mass number but different atomic numbers are called *isobars*. For example, carbon-11 decays to boron-11; the two are isobars. Atoms with the same number of neutrons are called *isotones*. Atoms with the same atomic and mass number (i.e., the same nuclide) but with different energy levels are called *isomers*. For example, technetium-99m decays to technetium-99; the two are isomers.

This nomenclature becomes important when we consider the radionuclides used in nuclear medicine. Radioactive isotopes of certain elements serve as the source of ionizing radiation that we image in nuclear medicine. As discussed below, we are particularly interested in obtaining certain radiations that drive our choice of nuclides.

7.3 Radioactive Decay

7.3.1 Mass Defect and Binding Energy

The sum of the masses of the constituents of an atom (i.e., the protons, neutrons, and electrons) is greater than the atom's actual mass. The difference between the sum of the masses of the atom's constituents and the actual mass is called the *mass defect*.

As an example, consider stable carbon-12. An atom of carbon-12 has 6 protons, 6 neutrons, and 6 electrons. On the atomic scale, it is common to express mass in *unified atomic mass units* (u), where 1 u is exactly 1/12 the mass of a carbon-12 atom. The mass of a proton is 1.007276 u, the mass of a neutron is 1.008665 u, and the mass of an electron is 0.000548 u. The mass defect of carbon-12 is therefore $6 \times 1.008665 + 6 \times 1.007276 + 6 \times 0.000548 - 12 = 0.098934$ u.

Einstein's famous relationship, $E = mc^2$, states that mass and energy are related and that matter and energy cannot be created or destroyed, they can only be transformed from one form into the other. Since mass and energy are related by Einstein's equation, there is an amount of energy "missing" from the atom (equivalent to the mass defect) that would otherwise be predicted to be present based on the constituents of that atom. This missing energy is termed the *binding energy* and can be computed from $E = \Delta mc^2$, where Δm is the mass defect and c is the speed of light (3×10^8 m/s). The unit of energy that we have been using throughout the book is the *electron volt* (eV), which is defined as the amount of energy an electron gains when it is accelerated across a voltage potential of 1 volt. Based on $E = mc^2$, one u is equivalent to 931 MeV.

In general, a more massive nuclide is more likely to have a larger mass defect, and therefore a larger binding energy, without necessarily being more "tightly bound." Thus, it is often more appropriate to consider the binding energy per nucleon rather than the total binding energy per se. As an example, we can again consider carbon-12. Since the mass defect of a carbon-12 atom is 0.098934 u, the binding energy is $0.098934 \times 931 = 92.1$ MeV. Since carbon-12 has 12 nucleons, the binding energy/nucleon $= 92.1/12 = 7.67$ MeV/nucleon. A graph of the binding energy per nucleon as a function of mass number is shown in Figure 7.1. This graph is often referred to as the *curve of binding energy*.

Binding energy applies to both the protons and neutrons in the nucleus and to the orbiting electrons. An electrostatic attractive force exists between particles with opposite charge (e.g., positively charged protons in the nucleus and negatively charged orbiting electrons). An electron in an inner orbit is attracted to the positively charged nucleus with a greater force than an electron farther away. The energy required to remove an electron completely from an atom is the *electron binding energy*, which is greater for electrons in the orbits closer to the nucleus because of the greater attractive force of the nucleus for inner electrons.

Why don't the protons in the nucleus repel each other? Normally, an electrostatic repulsive force exists between particles with the same charge;

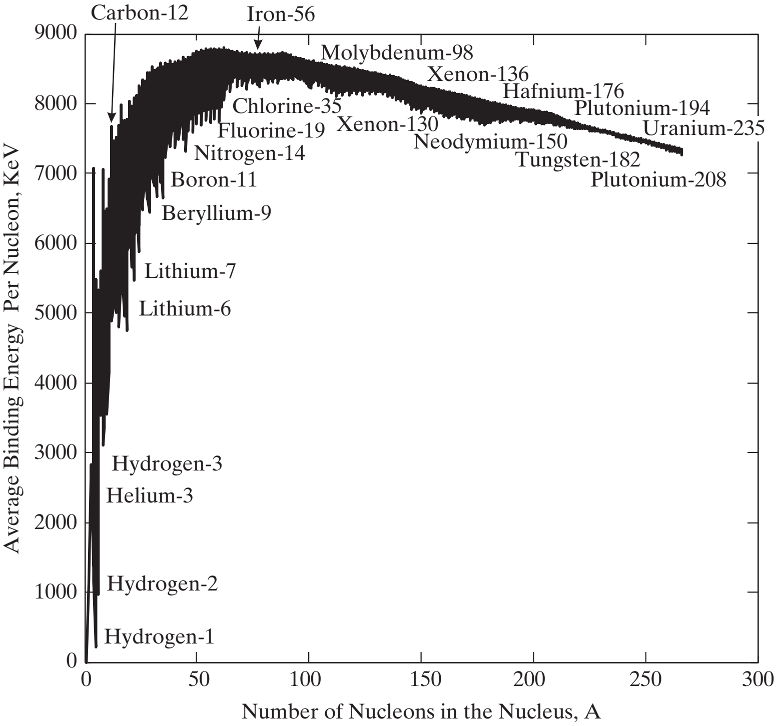


Figure 7.1
Average binding energy per nucleon as a function of mass number. All known nuclides are plotted; approximate positions of example nucleons are indicated.

protons repel each other when separated by a distance greater than the diameter of a nucleus. Within the nucleus, however, an attractive force, the *nuclear* or *strong force*, is responsible for holding neutrons and protons together. The energy required to separate the constituent protons and neutrons in a nucleus is the *nuclear binding energy*.

Radioactive decay is the process by which an atom rearranges its constituent protons and neutrons to end up with lower inherent energy. Radioactive decay occurs spontaneously, and energy is released in the process. The result of radioactive decay is an atom (the *daughter* atom) with less inherent energy than that which preceded it (the radioactive *parent* atom). This change in energy is reflected in the nuclear binding energy discussed above. Since binding energy is the amount of energy missing from an atom, the daughter atom has a higher binding energy per nucleon than the parent atom.

7.3.2 Line of Stability

If the different unique combinations of protons and neutrons as a nucleus that are found in nature are cataloged, they can be separated into two groups: nonradioactive nuclides, which are stable, and radioactive nuclides, which are called *radionuclides*. In general, the total number of nucleons, and the ratio of neutrons to protons, determine whether a nuclide is stable or radioactive.

A graph of the number of neutrons as a function of the number of protons for each unique stable nuclide produces a curve that follows the line of identity at low to intermediate atomic numbers (i.e., the nucleus has the same number of protons as neutrons), and then diverges up, so that for stable nuclides at higher atomic numbers there are more neutrons than protons in the nucleus. This curve, shown in Figure 7.2, is called the *line of stability*, because it depicts the data from all the stable nuclides. (It is called a "line" in practice even though it is not a straight line.) When plotted on the same graph, radionuclides fall off this line of stability. One way of conceptualizing radioactive decay is as the attempt of any radioactive atom that is "off" of this line of stability to reach the line of stability. In order to move toward the line of stability, an atom changes its proton to neutron ratio in the process of radioactive decay.

The greater the binding energy per nucleon, the more stable the atom. Thus, a second way to conceptualize radioactive decay is as a process an atom undergoes to increase its binding energy per nucleon. Since the daughter atom will always have a higher binding energy per nucleon than the parent atom,

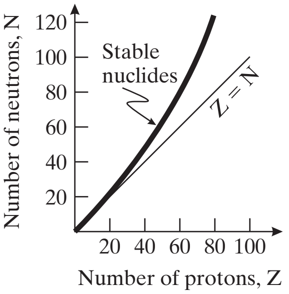


Figure 7.2
Stable elements having proton/neutron pairings that lie on the *line of stability*.

Figure 7.1 can also be helpful in visualizing the relationships between parents and daughters in radioactive decay.

For example, radon-222 is the daughter of radium-226. Radium-226 undergoes radioactive decay by alpha decay (see below) to produce radon-222. The alpha particle is emitted from the nucleus of a radium-226 atom during decay and carries off the "excess" energy of the atom. The binding energy/nucleon of radium-226 and radon-222 is 7.96 MeV/nucleon and 7.99 MeV/nucleon, respectively.

7.3.3 Radioactivity

The term *radioactivity* describes how many radioactive atoms are undergoing radioactive decay every second. It does not reflect what type of radiation is being emitted or the energy of that radiation. The common unit for radioactivity is the *curie*, abbreviated as Ci, where 1 Ci $= 3.7 \times 10^{10}$ disintegrations per second (dps). A *disintegration* refers to an atom undergoing radioactive decay. The SI unit of radioactivity is the *becquerel*, abbreviated as Bq, where 1 Bq = 1 dps. Clearly,

$$1 \text{ Ci} = 3.7 \times 10^{10} \text{ Bq}\,. \tag{7.1}$$

Radioactivity in the range of mCi or MBq is common in nuclear medicine procedures.

EXAMPLE 7.1

The intensity of radiation incident on a detector at range r from a radioactive source is given by

$$I = \frac{AE}{4\pi r^2}\,,$$

where A is the radioactivity of the material and E is the energy of each photon.

Question For technetium-99m with radioactivity of 1 mCi, what is the intensity at a distance of 20 cm?

Answer From Table 7.1 (see page 250), we know that the photon energy for technetium-99m is

$$E = 140 \text{ keV}\,.$$

The radioactivity is $A = 1 \text{ mCi} = 3.7 \times 10^7$ Bq. So the intensity at a distance 20 cm from the source is

$$I = \frac{3.7 \times 10^7 \text{ Bq} \times 140 \text{ keV}}{4\pi (0.2 \text{ m})^2} = 1.03 \times 10^{10} \frac{\text{keV}}{\text{sec} \cdot \text{m}^2}\,.$$

■

7.3.4 Radioactive Decay Law

The radioactive decay law states that the loss of atoms in a radioactive source per unit time is proportional to the number of radioactive atoms. Thus, letting

N represent the number of radioactive atoms in the source (and treating it as a continuous quantity), we have

$$-\frac{dN}{dt} = \lambda N\,, \tag{7.2}$$

where λ, the constant of proportionality, is called the *decay constant*. The decay constant has units of inverse time, and it is a constant for a given radionuclide. Assuming there are N_0 atoms at time $t = 0$, this expression can be integrated to determine the number of atoms N_t at time t, yielding

$$N_t = N_0 e^{-\lambda t}\,. \tag{7.3}$$

The *radioactivity* A of a source is defined as the number of atoms disintegrating per unit time. (Do not confuse this A with the mass number A of a nuclide or the area element in computing fluence.) Combining this fact with (7.2) and (7.3),

$$A = -\frac{dN}{dt} = \lambda N\,, \tag{7.4}$$

and

$$A_t = A_0 e^{-\lambda t}\,. \tag{7.5}$$

Either (7.3) or (7.5) is commonly referred to as the *radioactive decay law*. The factor that multiplies the initial value (N_0 or A_0) is called the *decay factor* (DF), and is given by

$$\mathrm{DF} = e^{-\lambda t}\,. \tag{7.6}$$

Recall that an exponential curve never goes to zero, so some radioactivity always remains. The *half-life* $t_{1/2}$ is the time it takes for the radioactivity (or the number of radioactive atoms) to decrease by a factor of 2. By definition,

$$\frac{A_{t_{1/2}}}{A_0} = \frac{1}{2} = e^{-\lambda t_{1/2}}\,. \tag{7.7}$$

Taking the natural logarithm of both sides and rearranging yields

$$t_{1/2} = \frac{0.693}{\lambda}\,. \tag{7.8}$$

Thus, the half-life and decay constant have a fixed relationship across radionuclides. As with the decay constant, the half-life is a constant for a given radionuclide (but differs across radionuclides).

EXAMPLE 7.2

Consider two radionuclides P and Q. Suppose the half-life of P is twice that of Q, $t^P_{1/2} = 2t^Q_{1/2}$. At $t = 0$, we have N_0 atoms of both radionuclides.

Question When will the radioactivities of two radionuclides be equal?

Answer Since $t_{1/2}^P = 2t_{1/2}^Q$, the decay constants of two radionuclides have the relation $\lambda_P = \lambda_Q/2$. So at $t = 0$, the radioactivities for P and Q are

$$A_0^P = \lambda_P N_0 \qquad \text{and} \qquad A_0^Q = \lambda_Q N_0 = 2A_0^P\,.$$

Based on the radioactive decay law, we have

$$A_t^P = A_0^P e^{-\lambda_P t} \qquad \text{and} \qquad A_t^Q = A_0^Q e^{-\lambda_Q t} = 2A_0^P e^{-2\lambda_P t}.$$

By equating A_t^P and A_t^Q, we have the following equation for t:

$$e^{-\lambda_P t} = 2e^{-2\lambda_P t}.$$

The solution is $t = \dfrac{\ln 2}{\lambda_P} = t_{1/2}^P$. So at $t = t_{1/2}^P$, the radioactivities of two radionuclides are equal. ■

7.4 Modes of Decay

It is important to understand the different modes of radioactive decay as these govern the different types of ionizing radiation produced. There are four main modes of decay: (1) *alpha decay*, which results in emission of an alpha particle; (2) *beta decay*, which results in emission of a beta particle; (3) *positron decay*, which results in emission of a positron; and (4) *isomeric transition*, which results in emission of a gamma ray. These ionizing radiations fall into two classes: *particulate radiation* and *electromagnetic radiation*. The main particulate ionizing radiations resulting from radioactive decay are alphas, betas, and positrons. Alpha particles consist of two protons and two neutrons, beta particles are identical to electrons except that they originate from the nucleus of an atom, and positrons are antimatter electrons. The electromagnetic ionizing radiations resulting from radioactive decay are gamma rays. Of these, in medical imaging we are only concerned with *positrons* (used for positron emission tomography in Chapter 9) and *gamma rays* (used for planar scintigraphy in Chapter 8 and single photon emission computed tomography in Chapter 9).

7.4.1 Positron Decay and Electron Capture

A nuclide with a ratio of neutrons to protons too low for stability can undergo electron capture or positron decay. These processes can be thought of as the transformation of a proton into a neutron, although that is not what literally occurs. A *positron* β^+ is emitted in the process of positron decay:

$$p \to n + \beta^+ + \nu\,,$$

where p is a proton, n is a neutron, and ν is a neutrino. For example, carbon-11 decays by positron emission to boron-11:

$${}^{11}_{6}\mathrm{C} \to {}^{11}_{5}\mathrm{B} + \beta^+ + \nu\,.$$

A positron is actually an antimatter electron having a +1 charge and the same rest mass as that of an electron. The neutrino, emitted with the positron, is a massless, chargeless subatomic particle. For a given radionuclide undergoing positron decay, the total kinetic energy of the positron plus the neutrino is a constant; however, the division of this total energy among the positron and neutrino varies from emission to emission.

Since a positron is an antimatter electron, an "unusual" atomic process occurs when a positron and electron meet. After a positron is emitted, it travels a short distance (several millimeters) in the material in which it is emitted, depositing its kinetic energy. (Its kinetic energy is transferred to the material by both collisional and radiative transfer, just as with a beta particle or energetic electron.) It then meets a free electron in the tissue, and—because the positron is an antimatter electron—mutual annihilation occurs. From conservation of energy, two 511 keV annihilation photons appear (511 keV is the energy equivalent to the rest mass of an electron or positron from $E = mc^2$); from conservation of momentum, they are emitted 180° back-to-back.

Atoms that undergo positron decay have an alternative decay mode—the nucleus can actually capture an electron:

$$p + e^- \rightarrow n + \nu \, .$$

Typically, an inner shell (K or L shell) electron is captured as these are in the closest physical proximity to the nucleus. A given radionuclide source will undergo electron capture at a certain fraction of time, and positron decays the other, but a given atom will (of course) only undergo decay once, by one of the two possible modes.

7.4.2 Isomeric Transition

A radionuclide may decay to a more stable nuclide that has the same atomic and mass numbers; both the parent and daughter are not only the same element but the same isotope of that element. In such a case, the parent usually represents a transient, *metastable* state with higher energy. This excess energy is released in the form of gamma rays. Recall from Chapter 4 that gamma rays are in practice indistinguishable from the x-rays that are now familiar to you. (Some textbooks indicate that the distinction between x-rays and gamma rays is based on wavelength or energy, but it is actually based on site of origin: gamma rays from the nucleus and x-rays from the electron cloud. In practice, most gamma rays have higher energies than many x-rays; hence, the common but incorrect distinction is based on wavelength or energy.)

An example of an isomeric transition is

$$\text{Cs-137} \rightarrow \text{Ba-137m} \rightarrow \text{Ba-137} + \gamma \, , \tag{7.9}$$

where γ represents the 662 keV gamma ray photon released when Ba-137m decays to Ba-137. Here, the "m" in Ba-137m represents the metastable state.

7.5 Statistics of Decay

Radioactive decay is a random process. If you conduct two experiments in which you start with exactly N_0 atoms of the same radionuclide in each experiment, and then count the number of decays over a period of exactly one thousandth of a half-life, you are very likely to get two different counts. The difference you observe is not due to experimental error but due to the inherently random nature of radioactive decay. This tells us immediately that the radioactive decay law described in the previous section [(7.3) and (7.5)] must be considered as a description of average behavior; the radioactive decay law predicts the *mean* number of radioactive atoms and activity not the exact instantaneous number and amount.

For large numbers of radioactive atoms and for periods of time much smaller than their half-lives, the random behavior of radioactive decay is governed by the Poisson distribution, which was first introduced in Section 3.4.3 and is given by

$$\Pr[N = k] = \frac{a^k e^{-a}}{k!}, \tag{7.10}$$

where a is a parameter of the distribution. This is a *probability mass function* describing the probability that a random variable N will equal k in a given experiment. It can be shown that both the mean and variance of a Poisson random variable are equal to a, the only parameter of the distribution, as noted in (3.53) and (3.54).

As an example, suppose that we have N_0 identical radioactive atoms having a long half-life. Over a short period of time Δt, only a tiny fraction of these atoms will disintegrate; therefore, we can assume that N_0 is constant over this time. From the radioactive decay law [see in particular (7.2)], we would expect to see exactly $\Delta N = \lambda N_0 \Delta t$ disintegrations over this short time period. In fact, over repeated measurements we would actually see a variable number of disintegrations. We would also notice that the *average* number of disintegrations would tend toward this predicted value. This value turns out to be the mean of a Poisson distribution, which describes the true random nature of radioactive decay. That is, for this experiment we have

$$a = \lambda N_0 \Delta t,$$

and the Poisson distribution governing the decay of these radioactive atoms is

$$\Pr[\Delta N = k] = \frac{(\lambda N_0 \Delta t)^k e^{-\lambda N_0 \Delta t}}{k!}. \tag{7.11}$$

Let us consider the probability that there are no disintegrations in an extremely small time interval. Evaluating (7.11) for $k = 0$ yields

$$\Pr[\Delta N = 0] = e^{-\lambda N_0 \Delta t}, \tag{7.12}$$

which for very small Δt is approximated by

$$\Pr[\Delta N = 0] \approx 1 - \lambda N_0 \Delta t\,. \tag{7.13}$$

Since for an extremely small interval, there is going to be only one disintegration, or none at all, and since probabilities must sum to 1, we can interpret $\lambda N_0 \Delta t$ as the probability of having one disintegration from all N_0 radioactive atoms in the time interval Δt. This gives us another way to interpret the radioactive decay constant λ: It can be thought of as the probability of radioactive decay per radioactive atom per unit (small) time.

If Δt is treated as a parameter and allowed to vary, then (7.11) characterizes a time-varying process $\Delta N(\Delta t)$, which is known as a *Poisson counting process*. Such a process counts the number of events (disintegrations, in this case) that have occurred over a certain period of time. In this case, the quantity λN_0 is known as the *Poisson rate* (see Example 3.7). It has units of events (disintegrations) per second and can be thought of as a type of *intensity* of the Poisson process. It is also a measure of the activity A of the decay process. However, it must be reemphasized that the Poisson process behavior takes place over time periods much smaller than the half-life of a radionuclide. On the other hand, the radioactive decay describes the change in A over much longer periods of time.

EXAMPLE 7.3

A patient study needs to be completed in no more than 10 minutes with at least 3.5 million counts of photons to achieve the desired image quality.

Question Suppose we detect 6 K photons in the first second. What is the minimal half-life of the radionuclide for the study to be successful?

Answer During the first second, the number of photons detected is

$$\Delta N = \int_0^1 \lambda N_0 e^{-\lambda t} dt = N_0(1 - e^{-\lambda}) = 6\ \text{K}\,.$$

We need at least 3.5 million counts in 10 minutes, so

$$\Delta N = \int_0^{600} \lambda N_0 e^{-\lambda t} dt = N_0(1 - e^{-600\lambda}) \geq 3{,}500\ \text{K}\,.$$

The above two equations give us

$$\frac{1 - e^{-600\lambda}}{1 - e^{-\lambda}} \geq \frac{3{,}500}{6}\,.$$

Solving the inequality equation, we have

$$\lambda \leq 9.45 \times 10^{-5}\ \text{s}^{-1}.$$

The minimal half-life is

$$t_{1/2} = \frac{0.693}{9.45 \times 10^{-5} \text{Sec}^{-1}} = 7{,}333\ \text{s} \approx 2\ \text{hr}\,.$$

■

7.6 Radiotracers

There are about 1,500 known radionuclides, of which 200 or so can be purchased. Out of these, however, only a dozen or so are suitable for nuclear medicine for several reasons. First, we desire ‘‘clean’’ gamma ray emitters—that is, ones that do not also emit alpha and beta particles (because these particles would only contribute to patient radiation dose without usefully contributing to image formation, as they would never make it out of the patient’s body). Positron emitters are also suitable because the positrons rapidly annihilate with electrons to produce gamma rays. Second, unlike projection radiography and CT, where we require attenuation of the radiation to produce image contrast, in nuclear medicine we would prefer that there would be no attenuation of the radiation. This is because in nuclear medicine we are trying to determine the location of the emitters; attenuation simultaneously adds to the patient dose and reduces the signal we can detect. This requirement suggests that the energy of the gamma rays should be high so that they leave the body with little attenuation. On the other hand, we must be able to actually detect the radiation once it leaves the body, and the higher the energy the less likely the gamma rays will interact in the detector. For these reasons (and others, as we will see later), the gamma emitters used in nuclear medicine have energies in the range 70–511 keV.

Another important property of a radiotracer is its half-life. In general, the half-life of a radiotracer depends on both the physical half-life of the radionuclide and the biological half-life of the pharmaceutical; this biological half-life arises from body processes that metabolize or clear the pharmaceutical. The *effective half-life* T_e combines the physical T_p and biological T_b half-lives:

$$\frac{1}{T_e} = \frac{1}{T_p} + \frac{1}{T_b} \, . \tag{7.14}$$

In general, we need to be able to form images in a matter of minutes, not seconds or hours. If the radionuclide decays in seconds, there is hardly enough time to administer and metabolize the compound before the source is lost. On the other hand, if it takes hours to create an image, patient motion will be a serious problem and natural metabolic processes will change the distribution of the radiotracer in the body over the period of the exam. Furthermore, a long half-life increases the radiation dose to the patient. Therefore, the ideal effective half-life of the radionuclide is short but not too short. Typical times are on the order of minutes to several hours. Because of this relatively short time requirement, some radiotracers are made on-site in *generators* or *cyclotrons*, while others are ordered from a nearby radiopharmacy. A list of the more commonly used radionuclides and some of their properties is given in Table 7.1.

An important characteristic of a radionuclide is that it must be useful and safe to ‘‘trace’’ within the body, either by itself or attached to a compound. Some of the radionuclides in Table 7.1 are more useful than others in this respect. Iodine, for example is a naturally occurring substance in the body, accumulating in the thyroid gland. Iodine-123 or I-131 in a sodium salt can be administered orally and measured in the thyroid to assess thyroid function. Technetium-99m can be used to label diethylene triamine pentaacetic acid (DTPA), which is filtered by the kidneys, and serial images of the kidneys can be used to assess

TABLE 7.1

Common Radionuclides in Nuclear Medicine

Gamma Emitters

Z	Nuclide	Half-life	Photon Energy (keV)
24	Chromium-51	28 d	320
31	Gallium-67	79.2 h	92, 184, 296
34	Selenium-75	120 d	265
38	Strontium-87m	2.8 h	388
43	Technetium-99m	6 h	140
49	Indium-111	2.8 d	173, 247
	Indium-113m	1.73 h	393
53	Iodine-123	13.3 h	159
	Iodine-125	60 d	35, 27
	Iodine-131	8.04 d	364
54	Xenon-133	5.3 d	81
80	Mercury-197	2.7 d	77
81	Thallium-201	73 h	135, 167

Positron Emitters

Z	Nuclide	Half-life	Positron Energy (keV)
6	Carbon-11	20.3 min	326
7	Nitrogen-13	10.0 min	432
8	Oxygen-15	2.1 min	696
9	Fluorine-18	110 min	202
29	Copper-64	12.7 h	656
31	Gallium-68	68 min	1,900
33	Arsenic-72	26 h	3,340
35	Bromine-76	16.1 h	3,600
37	Rubidium-82	1.3 min	3,150
53	Iodine-122	3.5 min	3,100

Source: Wolbarst, 1993.

renal function. Gaseous O_2 in which one oxygen atom has been replaced by oxygen-15 is used to measure blood flow and to assess oxygen metabolism with positron emission tomography. Fluorodeoxyglucose (FDG) is used by the body like glucose, except that the (labeled) fluorine-18 atoms remain where the molecule is first used. Imaging the uptake of FDG in the brain, for example, is thought to reveal aspects of the mental processes involved in motor, perceptual, and cognitive tasks.

A final and somewhat subtle desirable characteristic for gamma emitters is that they should have monoenergetic (single energy) decay. The reason for this is that with monoenergetic emissions, energy-sensitive detection can be used to discriminate the primary photons from those that have been Compton scattered. It can be seen from Table 7.1 that most of these common radionuclides are monoenergetic. Even those that are not monoenergetic, such as Tl-201 and Ga-67, emit only two or three different energy photons, in contrast to a typical polyenergetic x-ray beam, which has a continuous spectrum of energies.

Technetium-99m (Tc-99m) is by far the most commonly used gamma ray producing radionuclide in nuclear medicine. It is easily generated on-site using a generator containing molybdenum-99, whose radioactive decay (half-life is 67 hours) produces Tc-99m. Tc-99m can in turn be tagged to a variety of molecules, including sulfur colloids, glucoheptonate, albumin macroaggregates, and phosphates, which can be used to track a variety of physiological processes in the body. Tc-99m has a half-life of 6 hours and a monoenergetic gamma ray energy of 140 keV, which are both ideal in terms of balancing the various radionuclide requirements. As we describe in the following chapters the instrumentation related to planar scintigraphy and SPECT, it will be useful to think of Tc-99m as the radioactive source and to work out examples with this source as well.

EXAMPLE 7.4

Consider an experiment that uses dual radionuclides, technetium-99m and thallium-201. At $t = 0$, we have equal number of atoms of both radionuclides.

Question When will the count rate of technetium-99m be smaller than 20 percent of the total count rate?

Answer The average number of photons from technetium-99m is given by

$$\Delta N_{Tc}(t) = \lambda_{Tc} N_0 e^{-\lambda_{Tc} t}.$$

Similarly, for thallium-201, we have

$$\Delta N_{Tl}(t) = \lambda_{Tl} N_0 e^{-\lambda_{Tl} t}.$$

When the counts of technetium-99m are smaller than 20 percent of the total counts, we must have

$$4\Delta N_{Tc}(t) \leq \Delta N_{Tl}(t).$$

From Table 7.1, we know $\lambda_{Tc} = 0.693/6 \text{ hr}^{-1}$ and $\lambda_{Tl} = 0.693/73 \text{ hr}^{-1}$. Solving the above equation, we have

$$t = 36.65 \text{ hr}.$$

■

7.7 Summary and Key Concepts

Nuclear medicine makes use of radiotracers (compounds that are chemically labeled with radioactive atoms), which depend on the radioactive decay of radionuclides. The choice and behavior of specific radionuclides significantly influences the utility of the radiotracer. In this chapter, we presented the following key concepts that you should now understand:

1. *Nuclear medicine* produces images that depict the distribution of a *radiotracer*; this distribution is governed by body function, not structure.
2. Nuclear medicine makes use of *radionuclides*; these are unique nuclear species representing radioactive atoms, which emit ionizing radiation upon (spontaneous) decay.

3. A given radionuclide is characterized by its *decay mode* (which indicates the type of ionizing radiation emitted) and *half-life* (the time for half the radioactive atoms to decay, on average).
4. *Radioactive decay* is a random process that is governed by the *Poisson distribution*. Its statistical properties are determined by the number of radioactive atoms, time, and a *decay constant*, which is a characteristic of a given radionuclide.
5. Radiotracers make use of radionuclides that emit radiation of appropriate type and energy, have half-lives that are not too short or too long, and can be used safely in the body to reveal physiology or function.

Further Reading

Cherry, S.R., Sorenson, J.A., and Phelps, M.E. *Physics in Nuclear Medicine*, 4th ed. Philadelphia, PA: W.B. Saunders, 2012.

Christian, P.E. and Waterstram-Rich, K.M. *Nuclear Medicine and PET/CT: Technology and Techniques*, 7th ed. New York, NY: Elsevier/Mosby, 2012.

Rollo, F.D. *Nuclear Medicine Physics: Instrumentation and Agents*. St. Louis, MO: C. V. Mosby, 1977.

Wolbarst, A.B. *Physics of Radiology*, 2nd ed. Norwalk, CT: Appleton and Lange, 2005.

Problems

Fundamentals of Atoms

7.1 Show that 1 u is equivalent to 931 MeV.

7.2 A deuteron has 1 proton and 1 neutron and has a mass of 2.01355 u. Compute its mass defect and binding energy.

Radioactive Decay and Its Statistics

7.3 Compute the mean and variance of a Poisson distribution with parameter a.

7.4 Suppose there are 1×10^9 radioactive atoms in a given sample of iodine-123 (^{123}I), which has a half-life of 13 hours.

- **(a)** What is the radioactivity of the original sample?
- **(b)** How many radioactive atoms can be expected to be present after 24 hours?
- **(c)** What is the probability that there will be 1×10^8 radioactive atoms left after 24 hours?

7.5 Consider the problem in Example 7.1 and suppose we have 1×10^{12} atoms of technetium-99m to start with. What is the intensity measured at $t = 0$ and at $t = 1$ hour?

7.6 **(a)** How long will it take for a radioactive sample with activity 1 Ci to decay to activity 1 Bq if the half-life is τ.

- **(b)** What approximate order of magnitude would you want the half-life of a radioactive tracer used in nuclear medicine to be? Milliseconds? Seconds? Minutes? Hours? Days? Weeks? Years? Explain your answer.

7.7 Consider the imaging system shown in Figure P7.1. The $D \times D$ square detector is on the x-y plane. The center of the detector coincides with the origin of the x-y plane. A point source, S, is on the z-axis a distance R away from the detector plane. The intensity on the detector is defined as the energy of photons detected on a unit area of detector in unit time. If the radioactivity of the source at time $t = 0$ is A_0,

(a) What is the intensity at each point on the detector plane as a function of time?

(b) What is the average intensity on the plane as a function of time?

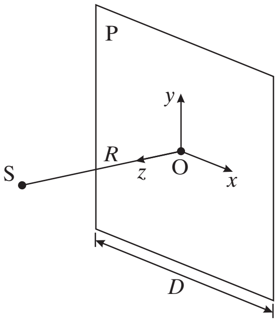


Figure P7.1
A simple imaging system with a radioactive source and a planar detector. See Problem 7.7.

7.8 (a) Show that the decay factor DF is related to the half-life by

$$\mathrm{DF} = e^{0.693t/T_{1/2}} \,.$$

(b) The *average lifetime* τ of a radioactive sample is given by $\tau = 1/\lambda$. Find an expression for τ in terms of the half-life.

7.9 A vial containing ^{99m}Tc is labeled "2 mCi/ml @ 8 a.m."

(a) What is the activity of the sample at 4 p.m. on the same day?

(b) What volume should be withdrawn at 4 p.m. on the same day to prepare an injection of 1.5 mCi for a patient?

7.10 (a) Assume that 10^8 atoms (N_0) are undergoing radioactive decay. If after 864,000 seconds (10 days) there are 9.9212×10^6 parent atoms left, what is the half-life of this radionuclide?

(b) If for this experiment the Poisson rate is 267.42 disintegrations per second, what is the average number of disintegrations over 0.01 seconds?

(c) What is the probability that there are more than two disintegrations over this time period?

7.11 At 1:00 p.m., we have 8 g of $^{21}_{11}$Ms, which has a half-life of 2 hours. How much will have decayed by 5:00 p.m.?

7.12 A radioactive element is labeled 3 mCi/ml @ 8 a.m. One hour later its activity is 1 mCi/ml.

(a) What is the half-life of this element?

(b) What is the activity of the same sample at 12 noon on the same day?

(c) What volume should be withdrawn at 12 noon on the same day to prepare an injection of 1.5 mCi for a patient?

Radiotracers

7.13 (**a**) Explain whether each radiotracer listed below is good or bad for medical imaging.

(i) $E_\gamma = 30$ KeV, $t_{1/2} = 7$ hours;

(ii) $E_\gamma = 150$ KeV, $t_{1/2} = 5$ hours; and

(iii) $E_\gamma = 200$ KeV, $t_{1/2} = 10$ days.

(**b**) The activity of an object at $t = 0$ is 4×10^{10} dps. Its activity reduces to 1×10^{10} dps after 5 hours. Compute the expected number of radioatoms present at $t = 0$.

7.14 A doctor referred a patient to have a SPECT scan using a specific radiotracer. That radiotracer was not available in the lab, but another radiotracer with the same emitted photon energy and half-life was available. Should the technician make the substitution? Explain.

Planar Scintigraphy

CHAPTER

8

8.1 Introduction

As in diagnostic x-ray imaging, nuclear medicine imaging evolved from projection imaging to tomographic imaging. Projection studies in nuclear medicine, called *planar scintigraphy*, have always used the *Anger scintillation camera*, a type of electronic detection instrumentation. The corresponding tomographic imaging method, called *single photon emission computed tomography* (SPECT), uses one or more rotating Anger cameras to obtain projection data from multiple angles. These imaging methods depend on radiotracers labeled with radioactive atoms whose decay produces a single gamma photon directly. Another imaging method, called *positron emission tomography* (PET), is based on radiotracers labeled with radioactive atoms whose decay produces a positron that is subsequently annihilated, producing two gamma photons. PET is implemented only as a tomographic imaging method and does not have a corresponding projection imaging mode. Together, SPECT and PET are referred to as *emission computed tomography*; they are described in the next chapter. Here, we focus on planar (2-D projection) scintigraphy.

8.2 Instrumentation

The three basic imaging modalities in nuclear medicine—planar imaging, SPECT, and PET—can be logically grouped in two ways. Planar imaging and SPECT use radiotracers that are gamma emitters, while PET uses radiotracers that emit positrons. On the other hand, SPECT and PET require tomographic reconstruction techniques (like CT), while planar imaging forms images by projection (as in projection radiography). In this section, we describe the instrumentation in planar scintigraphy, much of which is also used in SPECT, which is described in Chapter 9. Because of the peculiar nature of positron decay, PET scanners have significantly different instrumentation, which we also describe in Chapter 9.

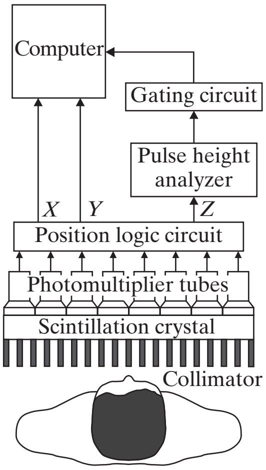


Figure 8.1
Components of an Anger scintillation camera.

As described in the Overview to Part III, nuclear medicine images are based on the distribution of radioactivity. Unlike projection radiography and CT, we are thus interested not in total intensity but in the detected decay rate of the source, typically expressed as *counts* per time. Nuclear medicine images are thus formed event-by-event.

The *Anger scintillation camera*, or *gamma camera*, was invented in the late 1950s by Hal Anger of the Donner Laboratory at the University of California, Berkeley. It is the most commonly used imaging instrument in nuclear medicine today. The complete camera system consists of a multihole lead collimator, a 10- to 25-inch circular, square, or rectangular sodium iodide scintillation crystal, an array of photomultiplier tubes on the crystal, a positioning logic network, a pulse height analyzer, a gating circuit, and a computer, as shown in Figure 8.1. We now describe each of these components in detail.

8.2.1 Collimators

The *collimator* is a 1- to 2-inch thick slab of lead of the same dimensions as that of the scintillation crystal, with a geometric array of holes in it. The lead that separates each hole is called a *septum*; collectively the lead represent *septa*. The collimator provides an interface between the patient and the scintillation crystal by allowing only those photons traveling in an appropriate direction (i.e., those that can pass through the holes without being absorbed in the lead) to interact with the crystal. Thus, the collimator discriminates against photons based on their direction of travel, and restricts the field of view of the crystal. In this regard, the collimator is not like a lens (although some authors draw that comparison) because it does not steer or focus the photons the way a lens does.

There are several types of collimators used with Anger cameras: parallel-hole, converging, diverging, and pinhole (Figure 8.2). The most commonly used collimator is the *parallel-hole collimator*, which consists of an array of parallel holes perpendicular to the crystal face, as shown in Figure 8.2(a). With parallel-hole collimators, the image on the crystal face is the same size as the object. Such collimators are either cast or fabricated from corrugated lead sheets. Originally,

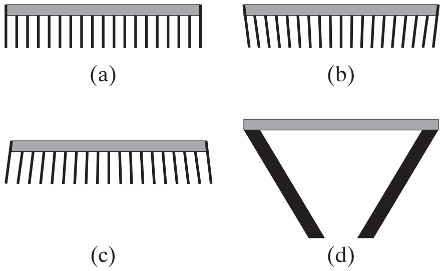


Figure 8.2 Different types of collimators: (a) parallel-hole, (b) converging, (c) diverging, and (d) pinhole.

collimator holes were circular in cross section. This meant that the lead septa were thicker in some areas than in others. Today, collimators have square, hexagonal, or triangular holes, and the septa are of uniform thickness around each hole.

Converging and diverging collimators are available as well but less frequently used in practice. A *converging collimator*, shown in Figure 8.2(b), has an array of tapered holes that aim at a point—the *focal point*—some distance in front of the collimator. The image that is presented to the crystal is a magnified version of the real object. A *diverging collimator*, shown in Figure 8.2(c), is essentially an upside-down converging collimator. Diverging collimators have an array of tapered holes that diverge from a hypothetical focal point behind the crystal. In this case, the image presented to the crystal face is smaller than the real object. Since converging and diverging collimators are simply flipped versions of each other, some collimator carriages actually have an insert that can be flipped either way, in effect producing two collimators in one (a *div/con collimator*).

EXAMPLE 8.1

When a converging or diverging collimator is used in an Anger camera, the image on the crystal face is not of the same size as that of the object.

Question Using a diverging collimator with a focal length of z_f, what is the ratio between the size of the image and the size of the actual object when a planar object is placed parallel to the crystal face at a distance z away from the collimator face?

Answer For a diverging collimator, only the photons traveling towards the focal point can be detected. (Others will be absorbed by the collimator walls.) Using geometry and ignoring the height of the collimator, we have the ratio between the size of the image and the size of the actual object as

$$\frac{S_i}{S_o} = \frac{z_f}{z + z_f}\,.$$

■

Pinhole collimators comprise a different class of collimators. As shown in Figure 8.2(d), these are thick conical collimators with a single 2–5 mm hole in the bottom center. As the source object is moved away from the pinhole, its image on the scintillation crystal gets smaller. However, the image is not always magnified (or minified) compared with the object. In fact, the camera image is magnified (i.e., larger than real size) from the collimator face to a distance equal to the length of the collimator and is then progressively smaller (minified) at larger distances.

The magnification produced by converging and pinhole collimators is useful in those cases when the object being imaged is significantly smaller than the field of view of the camera and when the object's size or the amount of detail within the object challenges the intrinsic spatial resolution capabilities of the system. With magnification, the image presented to the camera face is larger than life-size. This means that detail that is smaller in the "real world" than the resolving capability of the camera may possibly still be resolved, as long as the actual physical size presented to the camera face is larger than the camera's intrinsic resolution. This same concept is what allows us to see small details with a magnifying glass that are invisible to the unaided eye.

8.2.2 Scintillation Crystal

The *scintillation detector* is the most commonly used detector in nuclear medicine because it is more sensitive to electromagnetic radiation than a gas-filled detector. This type of detector is based on the property of certain crystals to emit light photons (*scintillate*) after deposition of energy in the crystal by ionizing radiation. The most commonly used scintillation crystal in nuclear instrumentation is sodium iodide with "thallium doping," which is written as NaI[Tl]. Since NaI[Tl] crystals absorb moisture from air, the crystals are hermetically sealed in aluminum "cans." Because the aluminum absorbs alphas and betas, NaI[Tl] detectors are generally used only for the detection of x-rays and gamma rays.

The scintillation crystals in a gamma camera are typically 10–25 inches in diameter and 1/4–1 inch thick. The thicker crystals are used for high-energy gamma rays, while the thinner crystals are used for low-energy gamma rays. Of course, any camera has only one crystal, so the user purchases a system with a crystal thickness matched to his or her needs. Like the screens that are used in projection radiography, thicker crystals stop more radiation than thinner crystals, but they also have poorer spatial resolution—so there is a tradeoff between efficiency and resolution just as there is in projection radiography. We will have more to say about this later in the chapter.

8.2.3 Photomultiplier Tubes

Each gamma photon that interacts in the scintillation crystal (by a photoelectric or Compton scattering process) produces a burst of light in the crystal, comprising thousands of light (scintillation) photons. This light is reflected and channeled out of the back of the crystal, through a glass plate and is incident upon an array of *photomultiplier tubes*. Each photomultiplier tube (PMT), shown in Figure 8.3, is a vacuum tube with two important components: a sensitive front surface, called the *photocathode*, and a series of electrodes, called *dynodes*. The PMT serves two important functions: it converts a light signal into an electrical signal and it amplifies this electrical signal.

The front surface of a PMT has a transparent window through which light travels in order to hit a photosensitive surface—the *photocathode*, an extremely thin layer of an alloy such as cesium and antimony. For every four to five light photons incident upon the photocathode, one electron is released or ejected from the photocathode by the photoelectric effect.

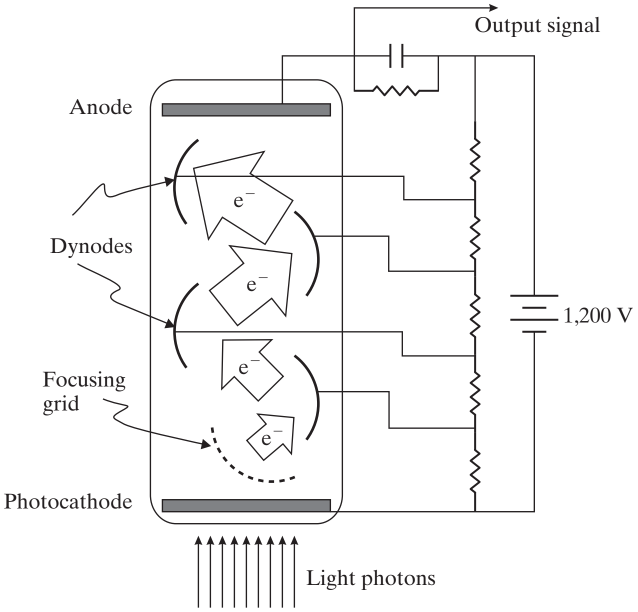


Figure 8.3
A schematic of a photomultiplier tube.

These photoelectrons are accelerated to the first *dynode*, an electrode that is positively charged and positioned at a short distance from the photocathode (see Figure 8.3). For each electron reaching the first dynode, 3–4 electrons are released. The second dynode has a higher voltage than the first; thus, the electrons liberated from the first dynode are accelerated to the second dynode. Each of these electrons in turn liberates three to four electrons from this second dynode. This process is repeated at 10–14 successive dynodes in the PMT, and 10^6 to 10^8 electrons reach the tube's anode (i.e., the final, most positively charged electrode) for each electron liberated from the photocathode.

These electrons comprise the output of the PMT, and they form a current pulse. The PMT outputs such a pulse each time a gamma photon interacts and deposits energy in the NaI[Tl] crystal, triggering the sequence of events in the tube as just described. This current pulse is directed into a preamplifier circuit, which amplifies the signal to provide a voltage pulse with a peak voltage—that is, its *pulse height*—that ranges from a few millivolts to a few volts.

As stated above, an array of PMTs covers the back of the crystal. This array means that light produced in the scintillation crystal is detected no matter where it is produced. An array is also necessary to produce the signals used in event positioning. The original Anger camera had seven PMTs, but modern gamma cameras have anywhere from 37 to 91 tubes or more, often arranged in a hexagonal pattern, as shown in Figure 8.4. Generally, more PMTs mean better spatial resolution and image uniformity, but this comes with a higher initial cost, more difficult calibration procedures, and more costly maintenance. The number of tubes is determined by the size and shape of both the crystal and each individual PMT. The tubes must be smaller, of course, when there are more of them arrayed over the scintillation crystal. Early PMTs had round cross sections, whereas current PMTs often have hexagonal cross sections to better "tile" the scintillation crystal without gaps between tubes.

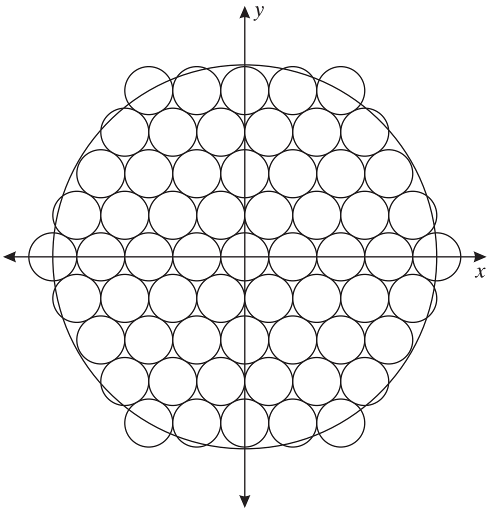


Figure 8.4
Example arrangement of 61 PMTs on the face of a circular Anger camera. Rectangular scintillators are most commonly used today and more complex shapes and arrangements of the PMTs are used to fit those crystals.

8.2.4 Positioning Logic

When a gamma photon interacts with the crystal, thousands of scintillation photons are produced, and every PMT produces an output pulse. The goal of the Anger camera's *positioning logic* circuitry is to determine both where the event occurred on the face of the crystal (see Figure 8.4) and the combined output of all the tubes, which represents the light output of the crystal (which in turn represents the energy deposited by the gamma photon). These output signals are denoted as X and Y for the estimated two-dimensional position of the event and Z for the total light output (which is proportional to the total energy deposited in the crystal by that gamma photon). The amplitude of a given tube's output is directly proportional to the amount of light (number of scintillation photons) its photocathode receives. The tubes closest to the scintillation event will have the largest output pulses, while those farther away will have smaller output pulses. By analyzing the spatial distribution of pulse heights, the location of a single scintillation event (X, Y) can be determined quite accurately—to within a fraction of the diameter of a PMT, in fact. We will discuss this process later, but for now it is important to remember that the *rate* of (X, Y, Z) signals or pulses coming from the camera is proportional to total radioactivity, and the *size* or *height* of each Z-pulse is proportional to the energy deposited in the crystal by the gamma photon for the event that (X, Y, Z) represents.

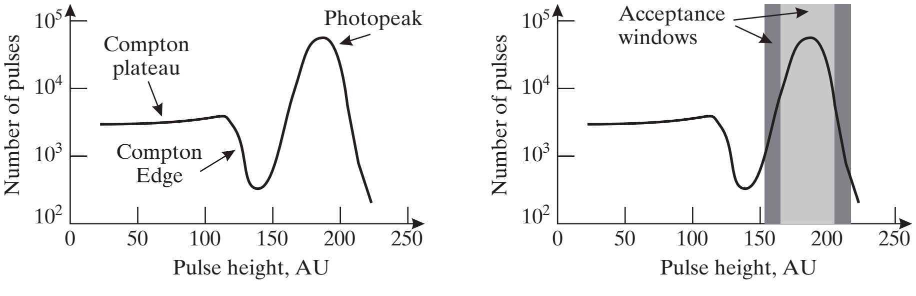
8.2.5 Pulse Height Analyzer

The desired goal of the Anger camera is to create an image that portrays the distribution of radioactivity (i.e., sites and numbers of radioactive atoms) within the patient. Because the collimator only allows those photons traveling in predetermined directions to interact in the crystal, a line drawn from the scintillation event in the crystal through the nearest collimator hole is presumed to intersect the site of origin of the photon (i.e., the radioactive atom it came from) in the patient. If the photon has been Compton scattered in the patient, a line drawn through its direction of flight will not intersect its site of origin but instead the site of the Compton interaction. Thus, photons scattered into the field of view could be falsely attributed to activity at the sites of Compton interactions in the patient. It is clearly not desirable to have these scattered photons contribute to the final image, as they significantly degrade resolution and contrast. It is important to note that a large percentage of photons striking the crystal have been scattered in the patient.

Fortunately, photons that have been Compton scattered can be distinguished from those that arrive directly by analyzing the energy deposited in the crystal via the Z-pulse, whose height is proportional to the total energy deposited in the crystal. In practice, we know the energy of gamma photons emitted from the radioactive source since we choose the radionuclide to use in a given imaging session and we understand its radioactive decay properties. If this amount of energy is deposited in the crystal, the gamma photon could not have Compton scattered in the patient since it would have lost some of its energy in the process [as determined in (4.8)]. Of course, a gamma photon could emerge from the patient with all its energy (i.e., without having interacted in the patient) and only deposit part of its energy in the crystal (via a Compton scattering interaction). We cannot distinguish between this event and one in which the gamma photon Compton scattered in the patient and then deposited all its remaining energy in the crystal via a subsequent photoelectric event. In practice, we can only be certain that the gamma photon has not scattered if it deposits in the crystal the full amount of energy with which it was emitted, and these are the only events we allow to contribute to image formation.

Scintillation spectrometry, or *pulse height analysis*, refers to the use of a scintillation counting system to obtain an energy spectrum from a radioactive source. This energy spectrum is a plot of the number of pulses with a given pulse height as a function of the pulse height, as illustrated in Figure 8.5(a). The measured spectrum is a function of the energies of gamma rays emitted by the source and the interactions of these photons in the source and any intervening material (e.g., the body), and in the crystal. It should be noted that the pulse heights on the x-axis generally have arbitrary units (AU), although they could be scaled to units of energy (keV) if the camera were properly calibrated.

A pulse height spectrum, as shown in Figure 8.5(a), has two main features: a broad plateau called the *Compton plateau* and a peak at the highest pulse heights, which is called the *photopeak*. The broad plateau represents Compton scatter interactions in the patient and/or crystal, broadly distributed in energy



(a) (b)

Figure 8.5
Pulse height spectrum for technetium-99m, plotting number of pulses or count rate as a function of pulse height or energy.

due to the random nature of the Compton scatter angle. The rightmost limit of this plateau, called the *Compton edge*, represents Compton interactions in which the incoming (unscattered) gamma photon is backscattered 180° in the crystal, thus depositing the maximum energy possible in a single Compton interaction. The photopeak represents gamma photons that come directly from the source without interacting and losing any energy and deposit all of their energy in either a single photoelectric interaction or one or more Compton interactions followed by a photoelectric interaction. Because a gamma photon cannot lose all its energy in a single Compton scatter event, there is a separation between the Compton plateau and the photopeak.

Of importance, scintillation detectors have imperfect energy resolution, as discussed later in the chapter. As a result, the photopeak, which should be an infinitely thin vertical line, is spread. The degree of spreading is related to energy resolution in exactly the same way as the point spread function is related to spatial resolution.

EXAMPLE 8.2

The pulse height analyzer is used to reject photons that have been Compton scattered.

Question What is the maximum angle through which a 140-keV photon can be scattered and still be accepted within a 20 percent pulse height window?

Answer Note that 20 percent pulse height window is 10 percent on either side. Therefore,

$$140 \text{ keV} \times 0.1 = 14 \text{ keV},$$

$$140 \text{ keV} - 14 \text{ keV} = 126 \text{ keV}.$$

Since

$$h\nu' = \frac{h\nu}{1 + \frac{h\nu}{m_0 c^2}(1 - \cos\theta)},$$

we have

$$126 \text{ keV} = \frac{140 \text{ keV}}{1 + \frac{140 \text{ keV}}{511 \text{ keV}}(1 - \cos\theta)}.$$

Solving for θ yields $\theta = 53.54°$. ■

8.2.6 Gating Circuit

The *Z*-pulse is proportional to the total energy deposited in the crystal by the gamma ray photon and is used by the pulse height analyzer to discriminate against Compton-scattered photons. The pulse height analyzer is used to set an acceptance *window* around the photopeak (the dominant energy for the particular radiotracer being used), as shown in Figure 8.5(b). Its lower threshold is set to discriminate against Compton events, which have lower energy than a non-Compton event. Its upper threshold is set to discriminate against multiple (simultaneous, summed) events, which have more energy than a single event (and for which "position" is a meaningless concept). Because scintillation detectors have imperfect energy resolution, it is not desirable to use an extremely narrow window, because desirable events (i.e., true photopeak events) would be excluded. Because the window has a finite width, some scattered photons may still be accepted (those which scatter through a small angle, and thus retain most of their energy). For example, 140-keV photons can scatter by as much as 50° and still be accepted by the often-used 20 percent window (i.e., a window of ±10 percent around the photopeak energy). In practice, proper window setting is vital, as a window that is not centered around the photopeak (an *offset window*) can degrade field uniformity of response for many cameras; this is typically a result of the slightly better light collection efficiency directly under each photomultiplier tube.

Current cameras, with microprocessor-based correction circuitry (which will be described later), generally maintain good uniformity even with offset pulse height windows. Such cameras may be purposely peaked to the high side of the photopeak to further reduce scatter, by eliminating any Compton-scattered photons that show up in the lower half of the photopeak (due to every camera's less than perfect energy resolution). Some cameras have two or three separate pulse height windows to simultaneously image the multiple emissions of some radionuclides (e.g., those from gallium-67). In this way, counts are acquired in a shorter amount of time as the multiple energy emissions are utilized.

8.2.7 Image Capture

Together, the positioning logic and pulse height analyzer (with its gating circuit) yield an estimated position (X, Y) of each scintillation event that falls within the energy discrimination window. Earlier Anger cameras were analog and relied on photographic image formation. In these systems, the X- and Y-pulses associated with each Z-pulse that passed through the pulse height analyzer were used to position a finely focused dot of light on a cathode ray tube face. A collection of these light dots over time would produce an image. Since it took a period of time for a complete picture to be obtained, perhaps up to several minutes to obtain several hundred thousand counts, an integrating medium was used to record the image. The most frequently used media were various types of photographic film. A photographic camera was mounted on a cathode ray tube, and the shutter was left open during the entire image acquisition period. The film was developed and an image of the distribution of radioactivity was obtained.

It is standard today for an Anger camera to be interfaced to a computer. During the transition period from analog image capture, as described above, to digital capture via computer, the (X, Y, Z) signals were digitized (typically with 16-bit quantization) and used by software to create images. In today's cameras (see Figure 8.1), the output of each PMT is digitized by its own analog-to-digital converter, and the positioning logic and image formation described below are implemented in a fully digital fashion. As a result, the computer and its software become an integral and necessary part of the planar scintigraphy image formation process, as they do in computed tomography. We describe the computer-assisted image formation process in Section 8.3.

8.2.8 Solid State and Other New Cameras

While Anger scintillation cameras remain the dominant nuclear medicine imaging device, newer detectors are rapidly appearing. Of these, the *solid state camera* is fast becoming an important competitor to the Anger camera. Solid state cameras use semiconductors as the detector. Of note, semiconductors directly produce current when gamma photons deposit energy in them, and thus a semiconductor detector combines the functions of the crystal and PMTs. The most commonly used semiconductor is cadmium-zinc-telluride (usually abbreviated as CZT); CZT has better energy resolution and count-rate capability than typical scintillation detectors. The basic signal in a semiconductor detector is produced by *electron-hole pairs* created along the path as radiation passes through the detector and produces ionization. Each electron-hole pair is somewhat analogous to an ion pair. The motion of electron-hole pairs in an applied electric field generates the electrical signal. In many cases, the semiconductor is *pixelated*, meaning that it is partially cut (in the thickness direction) into an array of 1 to 3 mm pixels. This pixelation facilitates the algorithms used for event positioning.

A hybrid approach between a conventional scintillation detector with PMTs and a semiconductor detector is the use of a scintillation crystal (typically, pixelated cesium iodide instead of continuous sodium iodide) with photodiodes instead of photomultiplier tubes. This is essentially the same approach as discussed in Section 6.2.4 for CT. While such systems are not "solid state" in the same way as a semiconductor, they are often labeled this way in the marketplace. Whether a continuous or pixelated detector is used and whether the detection system is solid state, the general approach to image formation described below is equally valid.

Despite these advances, the conventional gamma camera remains the "workhorse" of nuclear imaging. This is because its performance-to-cost ratio is very high, particularly at larger fields of view—for example, for general purpose and SPECT imaging.

8.3 Image Formation

The primary mechanism for creating images in planar scintigraphy is to detect and estimate the position of individual scintillation events on the face of an Anger camera. We now describe how the locations of these events are estimated, how these locations are combined to produce images, and what these images represent mathematically in terms of the distribution of radioactivity in the body.

8.3.1 Event Position Estimation

A single scintillation event produces a flash of light that yields a response from many, if not all, PMTs attached to the Anger camera. The amplitude or height of each tube's response is proportional to its distance from the scintillation event, and this is how the position (X, Y) of the scintillation event is encoded. We start our discussion of event position estimation with a general mathematical framework and then specialize the presentation to both analog and digital cameras.

Consider a coordinate system with the origin $(0, 0)$ at the center of the crystal. Let the PMTs be indexed by $k = 1, \ldots, K$, and their positions be given by (x_k, y_k). Suppose the amplitudes of their response to a scintillation event are given by a_k, $k = 1, \ldots, K$. These heights are viewed as representing samples of a 2-D distribution of light, and the location of the maximum of this distribution represents the position (X, Y) of the event.

We could set (X, Y) equal to the location of the PMT having the largest amplitude, but that would provide only a very crude estimate of the location, yielding a resolution on the order of the size of a tube. Instead, scintillation cameras implement a center of mass calculation to estimate (X, Y). To calculate the center of mass of the tube responses, we first form[1]

$$Z = \sum_{k=1}^{K} a_k \,, \tag{8.1}$$

which represents the *mass* of the event's light distribution. This signal is also proportional to the total light output from the scintillation event (called the *Z-pulse*) and is used in the pulse height analysis circuit, as previously discussed. Given the total mass Z, the components of the center of mass (X, Y) are calculated as

$$X = \frac{1}{Z} \sum_{k=1}^{K} x_k a_k \,, \tag{8.2}$$

$$Y = \frac{1}{Z} \sum_{k=1}^{K} y_k a_k \,. \tag{8.3}$$

These three values (X, Y, Z) represent the "heart" of an Anger camera. Together, they provide the location (X, Y) of the scintillation event and the total light output Z—or, equivalently, the total energy—of the event.

It is worth making a couple of comments about this process. First, it should be understood that the a_k's are actually waveforms, short pulses arising from the scintillation event causing the cascading electrons in each PMT. The individual numbers X, Y, and Z represent values obtained at the peak of such pulses or by integrating over a short time frame. Second, small signals occurring in PMTs that are distant from the event can degrade the position estimates. Therefore, a threshold circuit is added to permit only signals above a certain level to contribute the estimates of X and Y. This *discriminator circuit* does not apply to the estimate of Z, which still uses the output of all PMTs. Do not confuse this discriminator circuit with the pulse height gating circuit, which determines whether the Z signal is in a range sufficient to qualify the event as both a non-Compton and nonmultiple event.

In a modern Anger camera, the event position estimation is implemented digitally. In most camera designs, the output of each photomultiplier tube goes directly to a separate analog-to-digital converter, typically with at least 16-bit precision. Then, software embedded in programmable read-only memory

---

[1] In this and the following two equations, we use discrete summation instead of integration to acknowledge the relatively small number of PMTs, photodiodes, or other signal outputs from a camera.

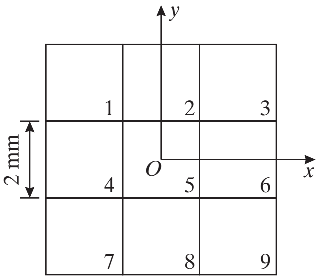


Figure 8.6 Arrangement of nine square PMTs of an Anger camera.

(PROM) chips determines the event position utilizing the center-of-mass calculation described above. An advantage of digital position estimation systems is that alternative algorithms can be easily implemented.

EXAMPLE 8.3

Consider an Anger camera with nine square PMTs arranged in a 3 × 3 array, as shown in Figure 8.6. Each tube is numbered, and the recorded pulse heights of the tubes are: 5, 10, 15, 10, 40, 30, 0, 5, 10, in the order of the tube numbers.

Question What is the value of the Z-pulse? What is the position of the event?

Answer The value of the Z-pulse is the summation of the pulse height recorded by each tube:

$$Z = 5 + 10 + 15 + 10 + 40 + 30 + 0 + 5 + 10 = 125\,.$$

The centers of the tubes are

$$\begin{aligned} c_1 &= (-2,2), \quad & c_2 &= (0,2), \quad & c_3 &= (2,2)\,, \\ c_4 &= (-2,0), \quad & c_5 &= (0,0), \quad & c_6 &= (2,0)\,, \\ c_7 &= (-2,-2), \quad & c_8 &= (0,-2), \quad & c_9 &= (2,-2). \end{aligned}$$

The estimated position of the event is

$$X = \frac{1}{Z}\sum_{k=1}^{K} x_k a_k = 80/125 = 0.64 \text{ mm},$$

$$Y = \frac{1}{Z}\sum_{k=1}^{K} y_k a_k = 30/125 = 0.24 \text{ mm}\,.$$

■

8.3.2 Acquisition Modes

For each scintillation event whose pulse height falls within the photopeak energy window, an Anger camera provides the estimated position (X, Y) of the event, the time of the event, and the peak height of the Z-pulse. In *list mode acquisition*, the (X, Y) signals are transferred directly to computer memory in the form of

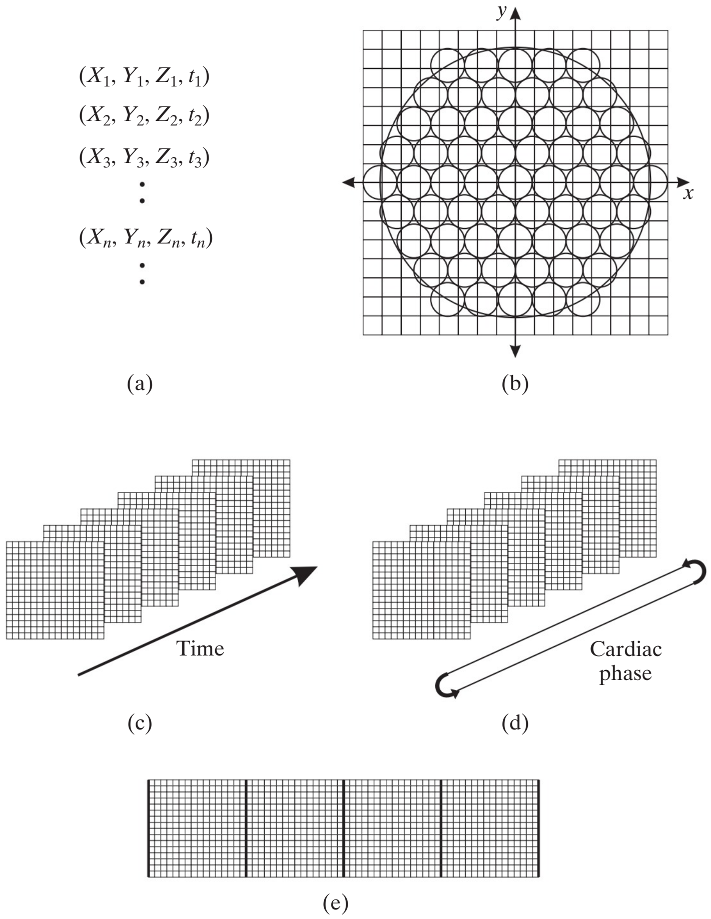


Figure 8.7
Anger camera acquisition modes: (a) list mode; (b) static frame mode; (c) dynamic frame mode; (d) multiple-gated acquisition; and (e) whole body.

a list of (X, Y) coordinates, as shown in Figure 8.7(a). In addition to the (X, Y) signals, typically the Z-pulse value is also recorded, and timing markers at evenly spaced intervals (e.g., every 10 ms) are also included. Physiological trigger marks, such as the occurrence of the R-wave from an electrocardiogram (ECG) monitor of the patient's heart, can also be inserted. In list mode acquisition, no matrices or images are formed within computer memory during acquisition. Although no images are produced immediately, list mode acquisition is useful because the (X, Y) signals are permanently recorded in computer memory, allowing flexible control over subsequent formatting into digital matrices. The precision of the event position is limited only by the number of bits used to store the X and Y positions. The downside is that a large amount of storage space is required even for a single study.

Static frame mode acquisition, as shown in Figure 8.7(b), represents the face of the Anger camera as a computer matrix—that is, a finite collection of discrete bins, or *pixels*, as introduced in Section 2.2. Accordingly, each pixel corresponds to a certain area of the camera face and is designated by a specific range of (X, Y) signal values. An image is created by starting with a zero image at the beginning of the scan and then incrementing the pixel containing (X, Y) with each (accepted) scintillation event. Finally, if there are 400,000 accepted events, then the sum of all the pixel values in the final image will be 400,000—that is, the *total counts* for the scan.

Dynamic frame mode acquisition, as shown in Figure 8.7(c) is a temporal succession of frame mode images. After the first image is created and stored, its buffer (matrix) is zeroed and another image is accumulated in identical fashion. This mode allows the study of transient physiological processes such as radiotracer uptake, washout, or redistribution. Movies of the temporal progression of a dynamic process can be created by appending the frames and replaying them after acquisition is complete. Generally, the number of pixels in dynamic frame mode images is smaller in order to yield higher counts (and correspondingly lower noise) in each pixel. [Increasing pixel size to reduce noise has already been discussed in the context of projection radiography; see (5.37).]

A further extension of frame mode acquisition is *multiple-gated acquisition* (or MUGA), as depicted in Figure 8.7(d). In this mode, the data from the camera are distributed to a series of matrices in computer memory. A trigger signal (usually a physiological trigger such as the R wave of the ECG) controls the distribution of data among the matrices. Immediately after the trigger, data from the camera are placed in the first matrix for a fixed time interval. When the time interval has elapsed, data are then placed in the second matrix for the same time interval. This process continues until the occurrence of a new trigger signal, at which time data distribution restarts at the first frame, or until all the assigned matrices in computer memory are used, in which case no data are acquired until the occurrence of a new trigger signal.

Multiple-gated acquisition is used to study a repetitive (cyclic) dynamic process. For example, in a cardiac-gated blood pool study, in which the circulating blood is labeled with radioactive atoms and the beating chambers of the heart are examined, the data from the corresponding phases of many heartbeats are superimposed during acquisition, resulting in a series of images representing one "average" cardiac cycle. Typically, the cardiac cycle is divided into 16–64 frames, with each frame representing 1/16–1/64 of the cycle.

Whole-body acquisition is yet another variation of static frame mode acquisition. In this mode, the body is divided into a matrix of pixels, as depicted in Figure 8.7(e) and a series of static frames are acquired to cover the body, in a "step and shoot" sequence. As an alternative, the camera or bed can continuously move. The position of the patient relative to the camera is maintained so that the camera can be scanned over the whole patient, thereby creating a whole-body image.

The most common (2-D) matrix sizes used in nuclear medicine are 64×64, 128×128, and 256×256, although matrices up to $1{,}024 \times 1{,}024$ are available. The larger the matrix size, the better the digital spatial resolution in the image, up to the limit imposed by the intrinsic capabilities of the camera. The digital sampling requirements necessary to preserve the spatial resolution in the analog image are given by the Nyquist theorem in Section 3.6. This theorem states that, to accurately portray a signal, the sampling frequency must be twice the highest frequency present in the signal. Thus, pixel dimensions should be smaller than one-half of the spatial resolution of the Anger camera. In practice, pixel dimensions range from 2 to 6 mm.

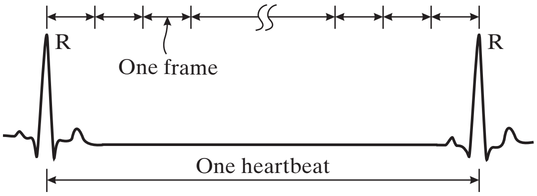


Figure 8.8
One cycle of heartbeat is divided into frames. The R wave of the ECG signal is used to trigger acquisition in multiple-gated frame mode.

EXAMPLE 8.4

The multiple-gated acquisition mode can be used to study beating hearts. In this mode, each cycle of the acquisition is triggered by the R wave of the ECG, as shown in Figure 8.8.

Question Suppose the heart rate is 50 bpm (beats per minute). We want each frame to last for 75 ms. What is the total number of frames?

Answer The heart rate is 50 bpm, so each heartbeat lasts for 60/50 = 1.2 s. Each frame lasts for 75 ms. Therefore, a heartbeat is divided into

$$N = \frac{1.2 \text{ second}}{75 \text{ ms}} = 16 \text{ frames.}$$

■

8.3.3 Anger Camera Imaging Equation

Consider what happens to a photon emitted from within the body. Statistically, the photon has equal probability to propagate in any direction. As shown in Figure 8.9, sometimes the photon will get absorbed (a), sometimes it will be scattered and then absorbed (b), and sometimes it will leave the body (with or without being scattered) (c, d, e). If it leaves the body, sometimes it will hit the Anger camera (d, e); more often it will not (c). When a gamma ray hits the camera, it will usually be absorbed by the lead in the collimator because it will either be traveling in an improper direction or will miss a collimator hole. We will ignore Compton scattering in our development of an imaging equation; hence, photons are assumed to travel in straight lines.

After biodistribution, the radioactive atoms constituting a radiotracer are distributed throughout the body (or sample) as a spatially varying concentration

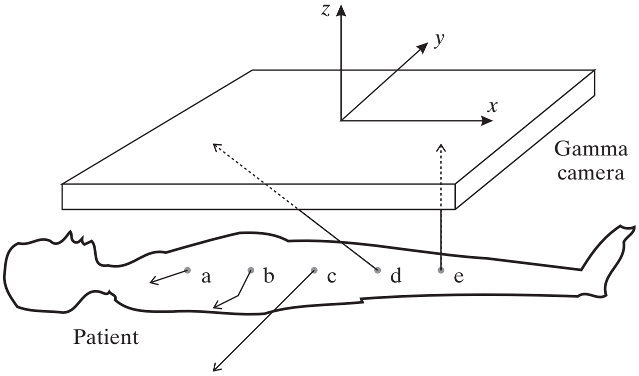


Figure 8.9
Imaging geometry showing the paths of five photons (a, b, c, d, and e) and a camera-centric coordinate system.

given by the number of radioactive atoms per unit volume. From (7.4), we see that multiplying this quantity by the radioactive decay constant λ gives the radioactivity per unit volume. It is customary in nuclear medicine to view this quantity, the radiotracer *activity concentration* function, denoted here by $f(x, y, z)$ and having, for example, units of Bq/m^3, as the fundamental quantity that is observed. Given this definition, the radioactivity within a differential volume element $dxdydz$ is given by

$$dA(x, y, z) = f(x, y, z)dxdydz\,. \tag{8.4}$$

A gamma camera counts individual photons hitting the detector, so unlike CT where intensity I is the appropriate raw measurement, in nuclear medicine the photon fluence rate ϕ [see (4.14)] is the correct raw measurement. Accordingly, incorporating both the inverse square law and attenuation along the path the differential photon fluence rate $d\phi(x_d, y_d)$ at the detector position $(x_d, y_d, 0)$ due to the differential source at (x, y, z) is

$$d\phi(x_d, y_d) = \frac{dA}{4\pi r^2} \exp\left\{-\int_0^r \mu(s; E)\, ds\right\}. \tag{8.5}$$

Here it is assumed that the origin ($s = 0$) of the line is at the point (x, y, z) and the end point of the line is at $(x_d, y_d, 0)$, which means that the total distance a photon travels from the source to the target is

$$r = \sqrt{(x - x_d)^2 + (y - y_d)^2 + z^2}\,.$$

In order to find the total photon fluence rate, we must integrate over all differential radioactivity elements that are physically able to contribute to the fluence at $(x_d, y_d, 0)$ on the detector. This implies that the collimator must be considered since it is designed to restrict the view of the detector face.

For simplicity, we restrict our attention to a parallel-hole collimator and consider the response at a detector point residing above (and within the projection of) a hole (see Figure 8.9). We further simplify our analysis by assuming that the only part of the object that is "seen" by this point is formed by a cylindrical tube created by extending the collimator hole along the z-axis into the object. Then the desired integral is given by

$$\phi(x_d, y_d) = \iiint_{\text{Tube}} \frac{f(x, y, z)}{4\pi r^2} \exp\left\{-\int_0^r \mu(s; E)\, ds\right\} dx\, dy\, dz\,. \tag{8.6}$$

This integral can be worked out in principle, but it suits our purpose better—in order to get at the essence of the situation—to make additional approximations to achieve further simplification. First, we assume that both the activity concentration $f(x, y, z)$ and linear attenuation coefficient $\mu(x, y, z)$ are constant within the tube at any given depth z. Second, we assume that $r \approx |z|$ for any differential radioactive element within the same tube. These assumptions allow us to substantially simplify the volume integral, yielding

$$\phi(x_d, y_d) = A_h \int_{-\infty}^0 \frac{f(x_d, y_d, z)}{4\pi z^2} \exp\left\{-\int_z^0 \mu(x_d, y_d, z'; E)\, dz'\right\} dz\,, \tag{8.7}$$

where A_h is the area of the collimator hole. The geometry in Figure 8.9, where the camera itself contains the laboratory origin, is used to establish the limits of integration of the integrals.

Equation (8.7) specifies the photon fluence rate at any position on the detector (given the stated assumptions); therefore, the expected number of counts in a planar scintigraphy image pixel can now be determined. First, the photon fluence can be computed by multiplying the photon fluence rate ϕ by the image acquisition time T. Second, the (mean) number of photons hitting a given pixel on the camera face is the integral of the photon fluence over the region defining the pixel. Making the assumption that the photon fluence is constant within the pixel means that we get the photon count by simply multiplying the photon fluence by the area of the pixel. Putting these facts together yields the mean photon count in pixel k as

$$\begin{aligned}\overline{n}_k &= \epsilon T A_k \phi(x_k, y_k) && (8.8)\\ &= \epsilon T A_k A_h \int_{-\infty}^{0} \frac{f(x_k, y_k, z)}{4\pi z^2} \exp\left\{-\int_{z}^{0} \mu(x_k, y_k, z'; E)\, dz'\right\} dz\,, && (8.9)\end{aligned}$$

where (x_k, y_k) and A_k are the position and area of pixel k, respectively. A detector efficiency factor ϵ has been introduced to account for losses of expected counts due to many factors such as collimator and scintillator losses.

We see from (8.9) that the pixel counts in a planar scintigraphy image are acquired by a type of projection of the underlying activity concentration. This is analogous to projection radiography—for example, (5.3)—but there are important differences. On the one hand, this equation is simpler because there is (typically) only one energy to consider; therefore, there is no integral over an energy spectrum as there is in radiography and CT. On the other hand, (8.9) is not a simple line integral of $f(x, y, z)$ since there are two sources of depth-dependent signal loss, one arising from the inverse square law and the other from object-dependent attenuation between the radiation source and the detector. Because of these effects, we can expect that radioactivity close to the camera will contribute more counts than an equivalent radioactivity that is farther away from the camera. This means that images from one side of the body will generally be very different than images acquired from the opposite side—that is, they will not just be mirror images. This is especially true for low-energy radiotracers since the body's attenuation is larger. Like the reading of x-ray films, where magnification distorts sizes of objects that are near versus far, we can get accustomed to reading planar images in nuclear medicine, knowing about object-dependent attenuation and the anatomy of the body. It will turn out this object-dependent attenuation causes problems for SPECT, and we will require a simpler imaging model for computational tractability, in analogy to the effective energy model employed in CT.

Planar Sources We saw above that the image formed on an Anger camera is a projection of the radioactivity within the body with depth-dependent effects. As in projection radiography, it is useful to consider a planar source in order to see what the image would be like without depth-dependent effects. This concept is particularly useful when we begin incorporating other effects such as blurring and noise in the next section.

The source or object in nuclear medicine imaging is the activity concentration $f(x,y,z)$. Let us define a planar source $f_{z_0}(x,y)$ which has radioactivity restricted to the plane defined by $z=z_0$; in other words,

$$f(x,y,z) = f_{z_0}(x,y)\delta(z-z_0)\,. \tag{8.10}$$

Because the units of $\delta(z)$ are inverse length (e.g., m^{-1}), we recognize that the new source $f_{z_0}(x,y)$ has units of planar activity concentration (e.g., Bq/m^2). Of course, $z_0 < 0$ in order to conform to the geometry of Figure 8.9. Substituting (8.10) into (8.7) and applying the sifting property of the impulse function [see (2.6)] yields

$$\phi(x_d,y_d) = A_h f_{z_0}(x_d,y_d)\frac{1}{4\pi z_0^2}\exp\left\{-\int_{z_0}^{0}\mu(x_d,y_d,z';E)\,dz'\right\}\,. \tag{8.11}$$

Using (8.8) we see that the mean pixel count in pixel k is

$$\overline{n}_k = \epsilon T A_k A_h f_{z_0}(x_k,y_k)\frac{1}{4\pi z_0^2}\exp\left\{-\int_{z_0}^{0}\mu(x_k,y_k,z';E)\,dz'\right\}\,. \tag{8.12}$$

Equation (8.12) shows that the mean pixel count in an image of a planar source represents the planar activity concentration within the source plane scaled by three factors. This first factor $\epsilon T A_k A_h$ is image acquisition or instrument design variables that can be controlled to some extent. To get higher counts, we can increase the efficiency, acquisition time, pixel size, or collimator hole size. This all makes perfect sense, but by this point in the book we know there will be tradeoffs in image quality that must be considered when taking any one of these measures. The second factor $1/(4\pi z_0^2)$ represents a signal loss due to the inverse square law. The signal loss, while dependent on z_0, is uniform in the plane—that is, independent of x and y—so it does not affect the relative amplitudes of the radioactive sources as they appear on the detector plane. The second factor is the effect of the integrated attenuation of photons as they travel from the plane $z=z_0$ to the plane $z=0$. Unless $\mu(x,y,z;E)$ is itself independent of x and y in the range $z \in [z_0, 0]$—which is not likely in realistic scenarios—then this term is not uniform, and it causes a variation of the relative source amplitudes across the image plane. This emphasizes the inherent difficulty, even for ideal planar sources, in comparing relative intensities between sources located at different points within a planar scintigraphic image.

8.4 Image Quality

There are many factors that affect the performance of Anger cameras, including spatial resolution, sensitivity, and field uniformity. Modern Anger cameras contain correction circuitry to improve performance in these areas as much as possible. We now explore these and other factors affecting image quality in planar scintigraphy.

8.4.1 Resolution

The ability of an Anger camera to localize a photon interaction is characterized by its *resolution*, defined by a full width at half maximum (FWHM) value, as discussed in Section 3.3. Two factors affecting resolution are most important: *collimator resolution* and *intrinsic resolution*. These two factors combine to yield the system resolution of the Anger camera. Since scintigraphic scans can be lengthy, patient motion enters as another factor affecting resolution, but we will not develop a model for patient motion here.

Collimator Resolution Since a parallel hole collimator is most commonly used, we focus on its performance here. With reference to the geometry shown in Figure 8.10, the collimator resolution is given by (see Problem 8.14)

$$R_C = \frac{d}{l}(l + b + |z|)\,, \tag{8.13}$$

where d is the collimator hole diameter, l is the collimator hole length, b is the (effective) scintillator depth, and $|z|$ is the collimator surface-to-patient distance. The most important feature of this equation is that the FWHM is dependent on target range $|z|$—that is, $R_C = R_C(|z|)$. In particular, collimator resolution degrades as $|z|$ increases; therefore, targets farther away are blurred more.

EXAMPLE 8.5

Making the collimator hole longer will reject more Compton scattered photons. This also has an impact on the resolution of the imaging system.

Question If we double the hole length to reject more Compton-scattered photons, what is the collimator resolution? Is there a limit on the collimator resolution?

Answer Equation (8.13) gives

$$R_C = \frac{d}{l}(l + b + r) = d + \frac{b + r}{l}d.$$

When the collimator hole length is doubled, the collimator resolution becomes

$$R'_C = \frac{d}{2l}(2l + b + r) = d + \frac{b + r}{2l}d < R_C.$$

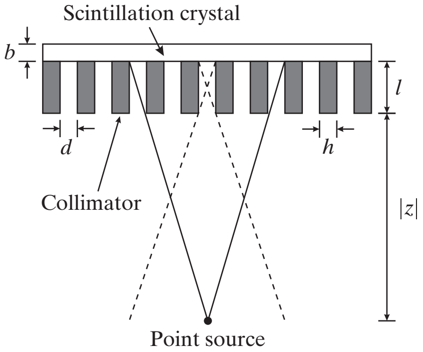


Figure 8.10 Collimator geometry demonstrating depth-dependent resolution.

So by doubling the hole length, the collimator resolution is also improved. If we make the hole length go to infinity, the collimator resolution tends to be the hole diameter, d, which is the limit to the resolution. We will see later, by using longer holes, we degrade the collimator sensitivity. ■

We can factor collimator resolution into our imaging equation for planar sources quite readily. Equation (8.13) does not give an explicit formula for the collimator PSF, but it is common to assume that the PSF is Gaussian with FWHM equal to R_C. For a Gaussian function (see Examples 3.4 and 3.5),

$$\text{FWHM} = 2\sigma\sqrt{2\ln 2}\,.$$

Therefore, since $\text{FWHM} = R_C$, a common approximation for the collimator PSF is

$$h_C(x, y; |z|) = \exp\left\{-4(x^2 + y^2)\ln 2/R_C^2(|z|)\right\}, \tag{8.14}$$

where we have made the dependence on range $|z|$ explicit.

Consider our planar source $f(x, y, z) = f_{z_0}(x, y)\delta(z - z_0)$ from the previous section. By design, all the sources are at range $r = |z_0|$ from the collimator. Therefore, putting (8.14) together with (8.11) yields

$$\phi(x, y) = A_b f_{z_0}(x, y)\frac{1}{4\pi z_0^2}\exp\left\{-\int_{z_0}^{0}\mu(x, y, z'; E)\,dz'\right\} * h_C(x, y; |z_0|)\,. \tag{8.15}$$

The PSF blurring is a convolution in this case because the collimator is spatially the same in the x-y plane. We could not write a 3-D convolution for a 3-D source, however, since there is depth dependency in both the basic imaging equation and the collimator PSF.

Equation (8.13) is based on the assumption that photons are always absorbed in the septa. In reality, some fraction of photons will penetrate the septa, which causes R_C to be somewhat larger than what is predicted above. A simple way to account for septal penetration (see Problem 8.5) is to assume in (8.13) that the hole length is slightly shorter than reality. The *effective hole length* (l_e) is defined as

$$l_e = l - 2\mu^{-1}\,, \tag{8.16}$$

where μ is the linear attenuation coefficient of the collimator material, typically lead, at the photon energy under consideration. The effective hole length is shorter than the actual length and is shortest at the highest photon energies because these photons are more penetrating (recall that μ is smallest at the highest photon energies). Using this definition, the adjusted collimator resolution becomes

$$R_C = \frac{d}{l_e}(l_e + b + r)\,. \tag{8.17}$$

This definition more accurately reflects the fact that a given collimator has a different (effective) spatial resolution depending on the energy of the radiotracer used. In particular, a collimator designed for low-energy radiotracers will have poor resolution when used for high-energy radiotracers (because photons can penetrate the septa without being absorbed), and is generally an ill-advised practice for that reason. It should be noted that b, the effective scintillation crystal thickness, is also energy-dependent. Higher-energy photons will penetrate the crystal farther, on average, before being absorbed, which contributes an additional degradation of resolution (R_C becomes larger) with high-energy photons.

Intrinsic Resolution The concept of collimator resolution incorporates both the geometry of the collimator and the energy-dependent septal penetration of photons. Additional blurring takes place in the scintillator itself, however, and this process is characterized by the *intrinsic resolution* of the Anger camera.

The concept of intrinsic resolution is not quite the same as that of resolution in an intensifying screen in radiography. In both projection radiography and planar scintigraphy, there is a flash of light in the scintillator (or phosphor) when a photon is absorbed. In projection radiography, the spread of the light is permanently recorded on the (film or solid state) detector. The amount that the light spreads represents the blurring of the image. In planar scintigraphy, the spreading of the flash of light is used to estimate the position of the single absorption event. It is inaccuracy in the estimation of (X, Y) that accounts for imperfect intrinsic resolution in planar scintigraphy.

There are two primary reasons for inaccuracy in the estimation of (X, Y) in an Anger camera. The first reason is related to the path of the absorbed photon. As discussed in Chapter 4, a photon that is absorbed by the photoelectric effect may have experienced some Compton scattering events prior to the final absorbing event. Those Compton events occurring in the scintillator generate flashes of light, smaller in intensity than the final event but spread out in space. These flashes are detected in the array of photomultiplier tubes (as though they came from a single event) and they contribute to the error in estimating (X, Y) by skewing the peak light distribution along the photon path.

The second reason for inaccuracy in the estimation of (X, Y) is noise or statistical fluctuation. Consider a photon that goes straight through a collimator hole and is absorbed promptly in a single photoelectric event. This is an ideal event, devoid of geometric or Compton scattering problems and should be localized perfectly if we lived in a perfect world. In fact, the physics and electronics that follow are beset with inherent statistical fluctuations that render the result less than perfect. First, both the number and the spatial distribution of scintillation light photons resulting from the photoelectric event are random. Although detailed analysis of the physics of the scintillator can provide good models for the mean numbers of these photons, the exact number cannot be known perfectly. Thus, the number and spatial distribution of light entering into any PMT are also random. The situation gets progressively worse within a PMT, where the electron cascade process is also random. All this leads to statistical variations in the peak pulse heights a_k, $k = 1, \ldots, K$ within the photomultiplier array. These variations lead to a statistical error variance associated with the estimation of (X, Y), which causes additional blurring in the localization of radioactive sources in space.

Intrinsic resolution of an Anger camera can be characterized by another FWHM, R_I, which encompasses both sources of variation in the scintillation process. It is common to model the PSF associated with this process by a

Gaussian function. Accordingly [see (3.23)], we define

$$h_I(x, y) = \exp\left\{-4(x^2 + y^2)\ln 2/R_I^2\right\}\,, \tag{8.18}$$

the intrinsic PSF of an Anger camera. Unlike collimator resolution, intrinsic resolution is not dependent on the target or its range. We can factor intrinsic resolution into our imaging equation for a planar source as follows:

$$\phi(x, y) = A_h f_{z_0}(x, y) \frac{1}{4\pi z_0^2} \exp\left\{-\int_{z_0}^{0} \mu(x, y, z'; E)\, dz'\right\} * h_C(x, y; |z_0|) * h_I(x, y)\,. \tag{8.19}$$

Generally, intrinsic resolution is much better than collimator resolution (i.e., $R_I \ll R_C$) at typical imaging depths, so the collimator's geometric characteristics dominate effective resolution. The effective resolution is the result of the subsystem cascade of R_I and R_C, as discussed in Section 3.3.4.

8.4.2 Sensitivity

As noted above, not every gamma ray leaving the body will be directed at the camera. If possible, we would like to detect those that are directed at the camera in the right direction. There are two major factors that may prevent such detection. First, the photon may be absorbed in the collimator; second, the photon may pass through both the collimator and the scintillation crystal. A very high sensitivity (or efficiency) camera will detect most photons; a very low sensitivity (or efficiency) camera will reject or miss most photons. Most cameras are somewhere in-between; we now characterize sensitivity in more detail.

Collimator Sensitivity From (8.13), we can see that longer collimator holes yield better resolution (smaller R_C). When the holes are made longer, however, fewer gamma ray photons that impinge upon the collimator will actually pass through it and reach the scintillator crystal. This means that the collimator is less efficient or less sensitive. The *collimator efficiency* ϵ, also called *collimator sensitivity*, is given by

$$\epsilon = \left(\frac{Kd^2}{l_e(d + h)}\right)^2\,, \tag{8.20}$$

where $K \approx 0.25$. The quantity ϵ reflects the fraction of photons (on average) that pass through the collimator for each emitted photon directed at the camera. All else being equal, we would like to choose collimator parameters that make ϵ as large as possible.

The resolution of a parallel hole collimator is best at the collimator surface and degrades with distance $|z|$, as shown in (8.13). On the other hand, (8.20) shows that the sensitivity is independent of the distance between the source and the collimator. While this seems at first to contradict the inverse-square law, it really does not. The field of view of each hole increases with increasing distance. This means that each hole “sees” a larger area at a greater distance from the collimator (see Figure 8.10). Looking at it another way, more holes see the same source if it is farther away from the collimator. As a radioactive source is moved away from the face of the collimator, the count rate through the central hole decreases due to the inverse-square law. However, more and more holes see the source, and the total count rate remains constant. Since more holes see the source, its image is spread over a larger area of the crystal face (i.e., the image is progressively smeared out). Thus, the resolution gets worse with increasing distance. The reason why the overall sensitivity is not affected in the same way is because the counts from the source are not lost but instead distributed to other parts of the camera.

Consider a fixed source range $|z|$ satisfying $|z| \gg l_e + b$. Then,

$$R_C \approx \frac{d}{l_e}|z|\,, \tag{8.21}$$

and by substitution,

$$\epsilon \approx \left(\frac{R_C K d}{|z|(d+h)}\right)^2 . \tag{8.22}$$

This relationship shows that there is a tradeoff between resolution and sensitivity. Thus, while we already understand the need for separate high-energy, medium-energy, and low-energy collimators to account for differences in septal penetration, there is also the need to consider different collimators for the same energy range that trade off resolution and sensitivity. Table 8.1, for example, shows several low energy (i.e., for 140 keV) collimators and their resolutions and sensitivities.

TABLE 8.1

Resolution and Sensitivity for Several Collimators

Collimator	d (mm)	l (mm)	h (mm)	Resolution @ 10 cm (mm)	Relative sensitivity
LEUHR	1.5	38	0.20	5.4	12.1
LEHR	1.9	38	0.20	6.9	20.5
LEAP	1.9	32	0.20	7.8	28.9
LEHS	2.3	32	0.20	9.5	43.7

LEUHR = low energy ultra-high resolution
LEHR = low energy high resolution
LEAP = low energy all purpose
LEHS = low energy high sensitivity

EXAMPLE 8.6

We saw in Example 8.5 that use of a collimator with long holes can help to reject Compton scattered photons and improve collimator resolution.

Question What is the change in sensitivity if we double the hole length and keep other parameters fixed?

Answer For simplicity, assume $l_e = l$. Then the sensitivity of a collimator is given by (8.20):

$$\epsilon = \left(\frac{Kd^2}{l_e(d+h)} \right)^2 .$$

If the collimator hole length is doubled, we have

$$\epsilon' = \left(\frac{Kd^2}{2l_e(d+h)} \right)^2 = \epsilon/4 .$$

So by doubling the collimator hole length, we reduce the sensitivity by 75 percent. ■

Detector Efficiency A gamma ray may enter the detector crystal, fail to interact with the crystal, and exit without depositing any energy. If no energy is deposited, obviously no pulse will be generated. Scintillation detectors are between 10 and 50 percent efficient for electromagnetic radiation. The wide difference in efficiencies is mostly the result of different designs trading off efficiency, resolution, and target gamma ray energy.

The crystals used in Anger cameras vary from 10 to 25 inches in diameter and from 1/4 to 1 inch thick. The thicker the crystal, the higher the probability that an incoming photon will interact, deposit its energy, and be detected; thus, the higher the sensitivity of the camera. The thicker the crystal, however, the poorer the spatial resolution, due to the complex interaction between the crystal, the PMTs, and the light pipe that is generally used to optically couple the two. One-quarter-inch-thick crystals have about 1 mm better intrinsic resolution than one-half-inch-thick crystals. When counting low-energy radionuclides such as thallium-201, there is no difference in sensitivity. However, when counting technetium-99m, one-quarter-inch-thick crystals have 15 percent less sensitivity than one-half-inch-thick crystals. At higher energies, the difference in sensitivity is even more significant.

8.4.3 Uniformity

Field uniformity is the ability of the camera to depict a uniform distribution of activity as uniform. At one time, it was thought that nonuniform response arose from changes in sensitivity across the crystal. To correct the nonuniformity, a *uniform flood* or sheet source of radioactivity was imaged and recorded, and used as a reference. Clinical images were corrected during acquisition by either adding counts to the image in areas where the flood had too few counts relative to the other areas or by subtracting (i.e., purposely not recording) counts in areas with too many counts.

It is now well-known that the majority of the nonuniformity in a camera occurs as a result of *spatial distortion*; that is, the mispositioning of events (errors in determining the (X, Y) location). To correct this distortion, reference images are acquired, and digital correction maps generated and stored. Each map contains values that represent (X, Y) correction shifts. Sophisticated microprocessor circuitry is used to reposition each count in real time during acquisition using these shifts. With many current cameras, it is best to acquire correction maps with the same radionuclide that is used for patient imaging. In some cameras, several sets of corrections maps are stored on the computer, representing all the radionuclides used in the nuclear medicine department, and the appropriate set chosen for a given patient study.

Variation in the position of a pulse from different areas of the camera within the pulse height window can also produce nonuniformities. This spatially dependent energy variation may also be corrected by microprocessor circuitry. The combination of energy variation and spatial distortion is responsible for loss of spatial resolution and imperfect linearity and uniformity. In today's cameras, explicit uniformity correction is typically carried out with multiplicative factors, only after spatial distortion (and spatially dependent energy response) corrections.

8.4.4 Energy Resolution

Pulse height analysis is critical for rejection of scattered photons, whose inclusion in the image would reduce contrast. Thus, the performance of the detector system, and especially its energy resolution, is critical. In an ideal detector, the photopeak would be a single vertical line at that pulse height representing the energy of the emitted x-ray or gamma ray. In reality, the statistical nature of the light emission and the finite energy resolution of the detector system smears out this line, producing a bell-shaped photopeak. In essence, the observed pulse height spectrum can be treated as the convolution of the ideal spectrum with the energy impulse response function of the system.

The worse the energy resolution of the detector system, the broader the photopeak. Energy resolution can be quantified as the FWHM of the photopeak in a fashion similar to the use of FWHM to characterize spatial resolution. This is measured by first determining the counts at the peak of the photopeak and locating the points on either side of the peak where the counts are half of the peak counts. The width of the photopeak in pulse height units is obtained by subtracting the lower pulse height from the upper. Finally, this width is divided by the pulse height at the apex of the photopeak to yield a percentage of energy resolution measurement. The smaller the number, the better the energy resolution; typical scintillation detector systems average 8–12 percent FWHM. It is important to note that the time resolution of the scintillation detector system also plays a role in energy resolution, in that two scintillation events occurring within the time resolution will produce a single, sum pulse. There is clearly a relation between energy resolution and the optimum width of the pulse height window discussed in Section 8.2.6 and illustrated in Figure 8.5.

In order to calibrate a pulse height analyzer to establish the quantitative relationship between energy deposited and pulse height, either the applied voltage across the PMT or the amplification in the electronics is adjusted until the photopeak from a known energy source falls at the desired pulse height. Changing the applied voltage across the PMT changes the size of pulses coming out of the detector by changing the amplification in the tube. If the voltage across each dynode is increased, electrons liberated from the previous dynode gain more kinetic energy. Upon striking the next dynode, a larger number of electrons are liberated. This leads to a larger signal from the PMT. Changing the amplification in the electronics directly changes the size of each pulse.

8.4.5 Noise

The entire imaging process, from gamma ray emission to electron cascade in a PMT is governed by the Poisson probability law (as introduced in Section 3.4.3). In particular, the number of detected photons (per unit area) is a Poisson random variable (by way of analogy to the initial discussion in Section 6.4.2). In a Poisson process, the variance is equal to the mean, which we have used to simplify the analysis of noise in projection radiography. We now study the implication of this fact for planar scintigraphy.

Signal-to-Noise Ratio Because both projection radiography and planar scintigraphy are ultimately limited by the *total* number of photons per unit area contributing to image formation, the calculation of SNR is essentially the same for the two modalities. In Chapter 5, we related intensity to the total number of photons [see (5.35)] in projection radiography; this same relation appears in (4.16), where we used the term *energy fluence rate* instead of *intensity*. By way of analogy to Section 5.4.1, we will focus on $\overline{N}$, the mean total number of acquired photons.

The intrinsic SNR of the camera (for this acquisition) is [see (3.59)]

$$\mathrm{SNR} = \frac{\overline{N}}{\sqrt{\overline{N}}} = \sqrt{\overline{N}}\,. \tag{8.23}$$

Therefore, the intrinsic SNR of the image is made larger by increasing the number of detected photons. This situation is identical to that of projection radiography.

We know that in frame mode, the face of the Anger camera is divided into a rectilinear grid of pixels. Suppose there are $J \times J$ pixels arranged in a square grid covering the camera. Then, without any prior knowledge about the source, we would expect to see the $\overline{N}$ detected photons spread evenly over all J^2 pixels. Accordingly, the intrinsic SNR per pixel is given by

$$\mathrm{SNR}_p = \sqrt{\frac{\overline{N}}{J^2}} = \frac{\sqrt{\overline{N}}}{J}\,. \tag{8.24}$$

An obvious, and important, implication of this equation is that the per-pixel SNR decreases as the pixel size becomes smaller. This should help to explain why it is not always advantageous to use a 512 × 512 matrix for a planar scintigraphic image. Aside from the fact that the camera may not have sufficient resolution to justify the implied pixel resolution, the SNR may drop precipitously, yielding a severely degraded image.

It is often very useful to characterize our ability to discern a hot spot or cold spot from the background. This requires the concept of contrast C, as introduced in Section 3.2. The contrast here is the fractional change in counts in a pixel or set of pixels from the background. If $\overline{N}_b$ is the mean background count and $\overline{N}_t$ is the mean target count, then the contrast is defined [in a fashion analogous to (3.12)] as

$$C = \frac{\overline{N}_t - \overline{N}_b}{\overline{N}_b} . \tag{8.25}$$

The local SNR is defined as

$$\mathrm{SNR}_l = \frac{\overline{N}_t - \overline{N}_b}{\sqrt{\overline{N}_b}} , \tag{8.26}$$

which is readily simplified [see (5.38)] to yield

$$\mathrm{SNR}_l = C\sqrt{\overline{N}_b} . \tag{8.27}$$

As a practical matter, the mean count is not known, anywhere, since it is a parameter of an unknown probability mass function. Instead, averages are used to approximate means. For example, if t_n, $n = 1, \ldots, N$ represents N pixel counts within a target and b_m, $m = 1, \ldots, M$ represents M pixel counts within the background (e.g., surrounding the target), then the means are estimated as

$$\hat{\overline{N}}_t = \frac{1}{N}\sum_{n=1}^{N} t_n , \tag{8.28}$$

$$\hat{\overline{N}}_b = \frac{1}{M}\sum_{m=1}^{M} b_m . \tag{8.29}$$

These estimates are then used in (8.27) to provide an estimate of the local SNR. This kind of calculation is used in practice to evaluate the detectability of objects, providing guidance about the choices of collimators, imaging times, pixel dimensions, etc.

8.4.6 Factors Affecting Count Rate

It is clear from the previous section that increasing the total number of detected and correctly positioned events will improve the SNR in images created using an Anger camera. The simplest approach to achieve this would be to increase the image acquisition duration, but this is usually undesirable in nuclear medicine for several reasons. The longer the acquisition, the greater the chance the patient will move, introducing motion blur in the image. As well, the longer the acquisition, the greater the chance that the radiotracer's distribution will change (because of the underlying physiological and biochemical processes). Finally, the longer the acquisition, the fewer the number of patients who can be imaged in a workday and the greater the discomfort and inconvenience each patient experiences.

If we are to keep the image acquisition duration to a feasible value, we must increase count *rate* in order to increase total counts. This situation is analogous to that in projection radiography, and some of the solutions are apparently the same—for example, improve detector efficiency or increase dose. But there are some practical considerations that must be addressed in planar scintigraphy that are largely missing in projection radiography. We know, for example, that it is straightforward to improve collimator efficiency, but it is at the expense of spatial resolution. So, for a given desired resolution, perhaps geared toward a particular clinical goal, we cannot improve collimator efficiency beyond a certain value.

We also know that we could improve detector efficiency in order to increase the number of detected photons. If that efficiency is improved by changes to the detector material, providing higher gamma ray stopping power with the same geometry, then that is good. A great deal of research has gone into this endeavor, leading to very efficient detector materials today, and it is not clear how much farther this can be taken. The obvious solution is to increase the detector's thickness, but we know that this will degrade resolution.

As in projection radiography, we can increase the dose, in this case by injecting a larger amount of the radiotracer (i.e., more radioactivity) into the patient. Aside from the obvious, undesirable properties associated with increasing the dose of ionizing radiation to the patient, planar scintigraphy (unlike projection radiography) has a problem handling the larger numbers of photons per unit time with increasing activity incident upon the camera. In particular, after the deposition of energy in the detector, it takes a certain amount of time for the scintillation photons to be given off, and for the electronics to register a "count." During this time, called the *dead time* or *resolving time* τ of the Anger camera, the detector is only partially responsive to the deposition of additional energy. The inverse of the resolving time is the maximum counting rate of the camera:

$$\text{Maximum counting rate} = \frac{1}{\tau} \,. \tag{8.30}$$

In clinical situations, resolving times are typically between 10 and 15 μs, and this leads to maximum counting rates of between 60K and 100K counts per second. Furthermore, Anger cameras are *paralyzable*, which means that the observed count rate does not simply plateau at $\frac{1}{\tau}$, but actually begins to progressively decrease (the camera "loses counts"), as the radioactivity increases.

Furthermore, exceeding the maximum counting rate may cause *pulse pileup*, a situation that occurs when two or more photons are absorbed in the crystal too close together to be counted as separate events. The photons may be treated as a single event whose energy is the sum of the energies deposited by the two (or more) photons; in this case, both (or more) are rejected due to energy window considerations. Alternatively, only the second may be rejected, due to the time required to count the first. Either case represents a loss of counts and an excess, unnecessary dose to the patient.

8.5 Summary and Key Concepts

Nuclear medicine produces images of function (body physiology and biochemistry). Planar scintigraphy is the nuclear medicine technique that produces 2-D projection images of radiotracer distribution using an Anger scintillation camera. In this chapter, we presented the following key concepts that you should now understand:

1. *Planar scintigraphy* is the nuclear medicine analog of projection radiography.
2. Planar scintigraphy makes use of an *Anger scintillation camera*, which consists of a *collimator*, a *scintillation crystal*, *photomultiplier tubes*, *positioning logic*, a *pulse height analyzer*, and an *image capture device* (today, a computer).
3. Typical *collimators* include *parallel-hole*, *converging*, *diverging*, and *pinhole*.
4. *Event positioning* is based on a *center-of-mass calculation*; unlike radiographic image formation, this takes place on a photon-by-photon basis.
5. The *basic imaging equation* includes terms for both *activity* (the desired parameter) and *attenuation* (an undesired, but extremely important, additional factor); these two terms are not separable.
6. *Image quality* depends on *resolution* (governed by both the collimator and the intrinsic resolution of the camera), *noise* (governed by the sensitivity of the system, the injected activity, and the acquisition time), and *count rate factors*.

Further Reading

Cherry, S.R., Sorenson, J.A., and Phelps, M.E. *Physics in Nuclear Medicine*, 4th ed. Philadelphia, PA: W. B. Saunders, 2012.

Christian, P.E. and Waterstram-Rich, K.M. *Nuclear Medicine and PET/CT: Technology and Techniques*, 7th ed. New York: Elsevier/Mosby, 2012.

Rollo, F.D. *Nuclear Medicine Physics: Instrumentation and Agents*. St. Louis, MO: C. V. Mosby, 1977.

Problems

Instrumentation

8.1 **(a)** Explain with diagrams the design and operation of an Anger gamma camera.

(b) What are some of the physical considerations when choosing which radionuclide to use for imaging.

8.2 Suppose you are designing a collimator to be used in an Anger camera. Assume the radiotracer used emits 140 keV gamma rays and the collimator surface-to-patient distance r is 10 cm.

(a) What is the minimum septal thickness h required for the septal penetration to be less than 60% for gamma rays incident at 45°?

(b) Suppose the resolution and sensitivity of the collimator are given by:

$$\text{Resolution} = d\left(1 + \frac{r}{l}\right), \quad \text{Sensitivity} = \frac{d^4}{l^2(d+h)^2},$$

where d is the collimator hole diameter and l is the collimator hole length. Assume $h = 0.2$ mm. What values of l and d should be chosen for the resolution and sensitivity of the collimator to be 7.8 mm and 28.9×10^{-4}, respectively?

(c) Describe the advantage(s) and disadvantage(s) of making the holes longer (with no other changes to the design).

(d) Can you compensate for the longer holes by changing the scintillator in any way?

8.3 A photon that has undergone Compton scattering has both changed its direction of flight and lost some energy. Pulse height analysis is used to reject these scattered photons during nuclear imaging. A pulse height window is symmetrically set around the photopeak of the pulse height spectrum. The full width of this window is usually expressed as a percent of the photopeak energy.

(a) Calculate the maximum angle through which a 140 keV photon can be scattered and still be accepted within a 20% offset window centered at 150 keV.

(b) What are the maximum acceptable scattering angles for 140 keV and 364 keV photons if 20% windows centered at photopeak are used.

(c) Draw a conclusion about the "directional selectivity" of an Anger camera as a function of energy.

8.4 Consider an Anger camera with only one parallel collimator hole. The measured intensity is the energy deposited on the camera per unit time per unit area. Suppose the hole diameter is d, and a point source with radioactivity of A is at a distance of r from the camera, directly below the hole.

(a) What is the measured intensity?

(b) What is the measured intensity if we double the hole diameter?

(c) What is the measured intensity if we double the source–camera distance?

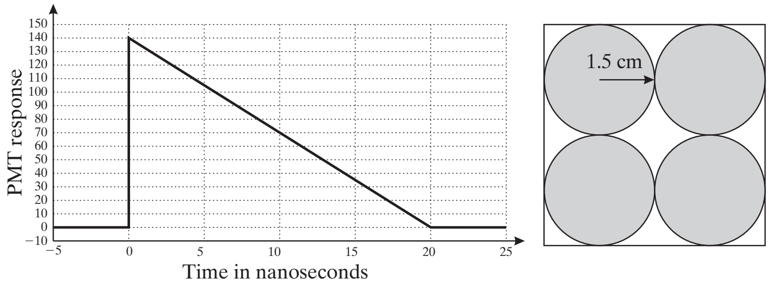
8.5 Photons can penetrate (i.e., go through) a collimator's septa, which means that a photon entering one collimator hole can sometimes be detected under another hole. Here we consider the effect of septal penetration on collimator design and collimator resolution. Let the hole diameter be d, the septal thickness be h, and the hole length be l. Ignore Compton scattering—that is, photons go straight—and ignore the scintillator thickness.

(a) What is the minimum distance w through a septum that the photon can pass under the condition that it enters one hole and is detected in the adjacent one? Simplify your expression by assuming that $l \gg 2d + h$.

(b) Suppose we desire septal thickness to be large enough to stop 95% of the photons that pass through the minimum distance w. What inequality should h satisfy? [Use the simplified expression from part (a)].

(c) How does septal penetration influence collimator resolution?

8.6 A radioactive pellet containing a 140 keV gamma ray emitting substance with activity A mCi is placed 1 m away from an Anger camera such that one-fourth of its total disintegrations hit the detector. Assume $t_{1/2} = 6$ hours. Assume that the Z-pulse height—that is, the photopeak—is equal to the energy of the incident photon.



(a) (b)

Figure P8.1
See Problem 8.6.

(**a**) What should be the acceptance window (in percentage) around the photopeak so that photons that are Compton-scattered more than 30° are rejected.

For the remainder of this problem, consider only non-Compton-scattered photons. Assume that the time response of the camera's Z-pulse to an absorbed photon hitting the camera at time $t = 0$ is given by the plot shown in Figure P8.1(a). Assume that the response of camera is additive in nature so that its net response to multiple absorbed photons at different time instants is the shifted addition of their individual responses.

(**b**) Plot and label the net response of camera's Z-pulse for two photons absorbed at time $t = 0$ ns and $t = 5$ ns.

Assume that the photon detection circuit and position logic circuit is turned on when the Z-pulse exceeds 80% of the photopeak and is turned off when the response falls below 80% of that height. A photon is considered to be accepted as an event if the maximum of the Z-pulse when the detection circuit is on falls within a 20% acceptance window around the photopeak.

(**c**) What is the minimum time interval between two detected photons such that they are accepted as separate events.

(**d**) What should be the activity of the radioactive pellet so that the chances of having at least one disintegration in the time interval given in part (c) is 50%.

Consider the arrangement of PMT's shown in Figure P8.1(b), where the origin is at the exact center of the square. Let the responses from the four PMT tubes to a detected photon be given by 80, 30, 20, and 5 (top-left, top-right, bottom-left, and bottom-right).

(**e**) Compute the height of the Z-pulse and the x and y location of the event.

Image Formation

8.7 (**a**) Derive an imaging equation for an Anger camera having a converging collimator.

(**b**) Derive an imaging equation for an Anger camera having a diverging collimator.

8.8 Consider nine 1 cm × 1 cm PMTs arranged in 3 × 3 array (Figure P8.2). The output of each PMT to a scintillation event is modeled as

$$a_i = 20 \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{5}\right), \tag{P8.1}$$

where (x, y) is the location of the scintillation event, and (x_i, y_i) is the center of the i-th PMT.

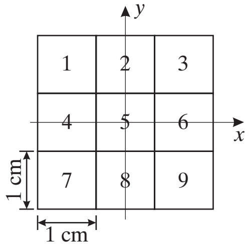


Figure P8.2
See Problem 8.8.

(**a**) Find the output of each PMT for a scintillation event that occurs at $(-0.5, 0.5)$ cm.

(**b**) What is the estimated position of the scintillation event?

(**c**) Is the estimated position the same as its true position? If not, explain why.

8.9 A series of scintillation events produce responses from the PMTs in an Anger camera. The recorded pulses have a photopeak centered at pulse height 180. The pulse height is proportional to the energy deposited into the crystal, which is shown in Figure P8.3(a). (This is not a pulse height spectrum.)

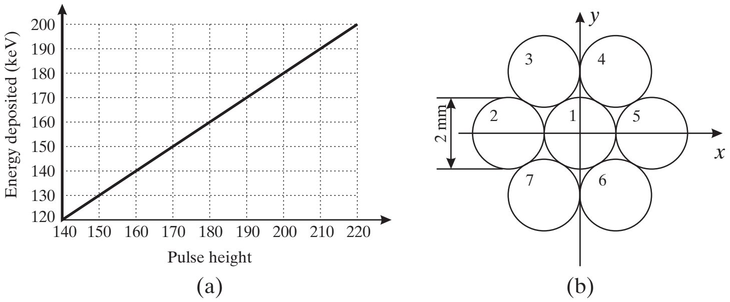


Figure P8.3
See Problem 8.9.

(**a**) We want to set an acceptance window around the photopeak to reject those the photons that have undergone Compton scattering angles larger than 50 degrees. What is the range of photon energies that will be accepted by the acceptance window?

Consider seven PMTs located around the origin of the x-y coordinates on the face of the Anger camera shown in Figure P8.3(b). The diameter of the tubes is 2 mm and each tube has been assigned a number. A single scintillation event yields a response from the PMTs. All other tubes recorded zero height except these 7 tubes, which recorded heights 40, 5, 15, 15, 20, 45, 30 in the order of tube numbers.

(**b**) Compute the Z-pulse. Will this pulse be accepted by the acceptance window?

(**c**) Estimate the position of the event (X, Y).

(**d**) Explain why we do not set (X, Y) equal to the location of the PMT having the largest amplitude.

8.10 You will be imaging a radioactive source using an Anger camera.

(**a**) Sketch and label the pulse height spectrum. Make the photopeak at 180 AU (for 140 keV photons).

(**b**) Determine the maximum scatter angle such that the photons will be within a 10% pulse height window.

For parts (c) and (d), assume the PMT arrangement and corresponding pulse heights shown in Figure P8.4.

(**c**) Compute the Z-pulse.

(**d**) Computer the position (X, Y) of the event.

(**e**) What are the causes of event localization error?

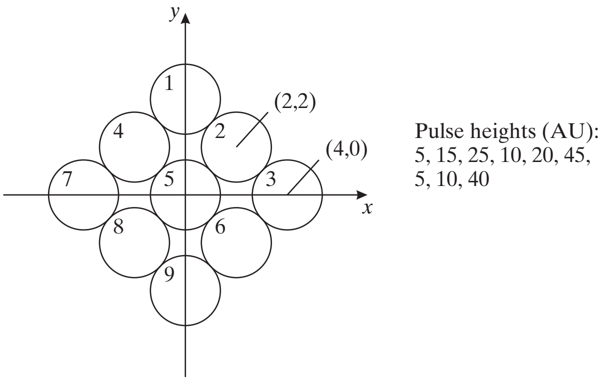


Figure P8.4
See Problem 8.10.

Image Quality

8.11 In practice the spatial resolution in a clinical nuclear medicine image is determined by three factors: the intrinsic resolution of the camera (I), the collimator resolution (C), and patient effects (P). Suppose the impulse response could be modeled as arising from the cascade of three linear shift-invariant systems, one for each factor.

(**a**) Suppose each system has an impulse response given by a rect function, and suppose their widths are r_I, r_C, and r_P, respectively. Find the FWHM of the cascade system.

(**b**) Suppose each system has an impulse response given by a Gaussian, and suppose their standard deviations are σ_I, σ_C, and σ_P, respectively. Find the FWHM of the cascade system.

8.12 We are conducting a study that requires us to acquire a number of images over a 2-hour period of time. Each acquisition lasts for 10 minutes and technetium-99m is used as the radiotracer. The images are recorded in frame mode with a matrix size of 128 × 128. In order to get satisfactory image quality, we need at least 2 million counts for each image.

(**a**) In order for the last image to have enough counts, what is the total count in the first minute of the study?

(**b**) What is the SNR per pixel as a function of the image number?

- (c) Suppose there is a tumor with 10% contrast, what is the approximate count for the first image so that the tumor has a 5 dB local signal-to-noise ratio in the last image.

8.13 Consider the cardiac study described in Example 8.4, and suppose the image for each frame is a 64 × 64 matrix. We need 1K counts for each pixel on average when the study is done. We can detect at most 128K photons per second on the entire camera.

- (a) How many heartbeats are needed?
- (b) How long does it take to complete the study?
- (c) What is the SNR for each pixel?
- (d) If we want to double the SNR by prolonging the study, how long will it take to complete the study?

8.14 Using geometric arguments, derive the collimator resolution formula given in (8.13).

8.15 A planar scintigraphy imaging system (assume a 2-D system) is set up with the collimator as shown in Figure P8.5.

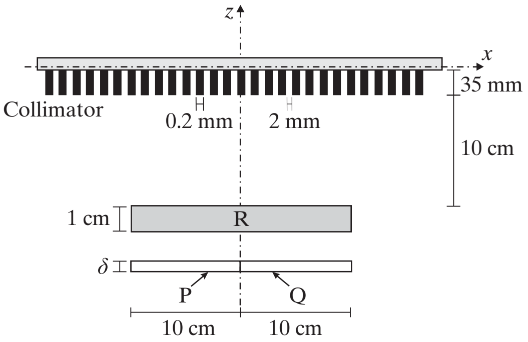


Figure P8.5
System setup for Problem 8.15.

Three compartments P, Q, and R mimick three different tissues, and have attenuations $\mu_P = 0$, $\mu_Q = 0$, and $\mu_R = 0.1\ \text{mm}^{-1}$, respectively. Assume $\delta \approx 0$ mm. N_0 atoms of radionuclide A ($t^A_{1/2} = 3$ hours) were put uniformly in compartment P, and $\frac{1}{2}N_0$ atoms of radionuclide B ($t^B_{1/2} = 6$ hours) were put uniformly in compartment Q.

- (a) Compute the projection $\phi(x)$ [units of photons $\text{s}^{-1}\ \text{cm}^{-1}$] at 1) time $t = 0$ and 2) time $t = t^A_{1/2}$.
- (b) Determine the time $t_{\max}$ at which the absolute signal difference $|\phi(5) - \phi(-5)|$ is maximum.
- (c) What is the collimator efficiency and resolution?
- (d) What is the absolute signal difference at time $t = 3$ hours incorporating collimator efficiency.
- (e) What is the width of compartment P in the final image? (Assume the system resolution is dominated by the collimator resolution.)

8.16 An Anger camera collimator has the following dimensions: septal thickness $h = 6$ mm, septal height $l = 10$ cm, and hole diameter $d = 3$ mm. The detector thickness is $b = 2.5$ cm, and the range of the object $z = 0.5$ m.

(a) Find the collimator resolution at the given range.

(b) Suppose the intrinsic resolution of the Anger camera is 0.2 mm. Write an expression for the overall PSF of the Anger camera at the given range, assuming both h_C and h_I are Gaussian.

Photons sometimes penetrate the collimator's septa, which means that a photon entering one collimator hole can sometimes be detected under another hole, as shown in Figure P8.6.

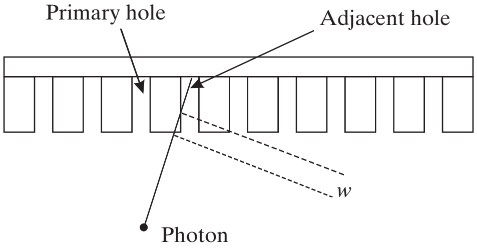


Figure P8.6
Septal penetration by a photon. See Problem 8.16.

In parts (c), (d), and (e), we consider the effect of septal penetration on collimator design and collimator resolution. Assume the linear attenuation coefficient for lead is μ.

(c) Assume a photon enters one hole, penetrates an adjacent septum and is detected in the adjacent hole. What is the minimum length w of the path through the septum as shown in the figure? (The result should be in units of mm.)

(d) Suppose we desire septal thickness to be large enough to stop 95% of the photons that pass through the minimum distance w. For fixed l and d, what inequality should h satisfy?

(e) How does septal penetration influence collimator resolution?

(f) Can we use the acceptance window to reject photons that penetrate septa and are detected in adjacent holes? Explain.

(g) A photon can be rejected if its Z-pulse is too small. What physical event(s) could cause this low Z-pulse?

8.17 Assume a 2-D scenario in which a pellet filled with radiotracer is located at the origin. The pellet emits gamma rays in the x-y plane and its radioactivity is $A = 0.54$ mCi as shown in Figure P8.7(a). The detector is made of a material with $\mu = 0.64\ \text{cm}^{-1}$ and thickness of $b = 2$ cm.

(a) What is the average rate of photons per second hitting the detector (in photons per second)?

(b) Find the efficiency of the detector at its center (that is, at $x = 0$).

A parallel hole collimator with septal thickness $h = 2.5$ mm, septal height $l = 8$ cm, and hole diameter $d = 5$ mm is added onto the detector, as shown in Figure P8.7(b).

(c) What is the collimator resolution R_C?

(d) Assume the intrinsic resolution of the detector is 1 mm. What is the overall PSF of the camera knowing that both h_C and h_I are Gaussian?

(e) Given the geometry, how many holes receive radiation that can hit the detector?

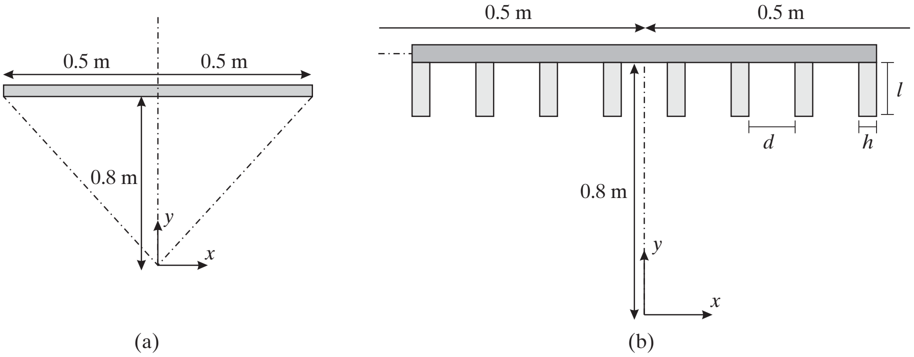


Figure P8.7 (a) Detector without collimator. (b) Detector with collimator. See Problem 8.17.

(f) If l gets larger, at some point radiation will only be able to hit the detector through one hole. Find the smallest $l = l_0$ when this happens. What is R_c in this case?

(g) If the camera sensitivity is ϵ when $l = 8$ cm, what is its sensitivity when $l = l_0$? (For simplicity assume that the effective height of a collimator hole is equal to l.)

Applications

8.18 Suppose you replaced the standard collimator on an Anger camera with one that has only one hole with diameter d positioned at the center of the camera—i.e., at $x = 0$ and $y = 0$. Let the object be a small radioactive point source that is placed 20 cm away from the camera face in direct alignment with the center of the camera.

(a) When a gamma ray photon goes through the collimator hole, the camera produces a response. Explain in words what happens to this photon and what sequence of events takes place in order to produce the X and Y signals and the Z-pulse.

(b) Suppose the X signal is held constant until the next photon hits the camera (so it is a piecewise constant signal $X(t)$ in time.) Draw $X(t)$ when the hole has a small diameter [call this $X_1(t)$] and then draw $X(t)$ when the hole has a larger diameter [call this $X_2(t)$]. Account for all effects that change when you increase the hole diameter and draw the signals on the same vertical and horizontal scales with clear labels.

(c) Would you expect the sequence of Z-pulses to change when the hole diameter (of this single hole) is increased? Explain your answer.

(d) If you double the hole diameter, what must you do to the hole length in order for the camera to have the same sensitivity?

8.19 A planar scintigraphy calibration experiment is being performed using a uniform flood phantom. With this phantom, each pixel out of 64×64 pixels on an Anger camera is struck by four gamma photons per second on average.

(a) If a total count of 2,000,000 across the whole camera is considered to be a complete experiment, how long will it take (on average) to complete the calibration scan?

The burst of light from a scintillator—the Z-pulse—lasts for a brief moment every time a gamma ray hits it. Let us model this burst as a triangle with peak height A (e.g., in volts), dropping linearly to zero in 250 μs, as shown in Figure P8.8. The pulse height analyzer detects peaks and records their peak voltage. For the pulse shown in Figure P8.8, therefore, the pulse height analyzer records the value A. Assume that the response of two or more gamma rays is the sum of their Z-pulses, delayed in time according to their time-of-arrival.

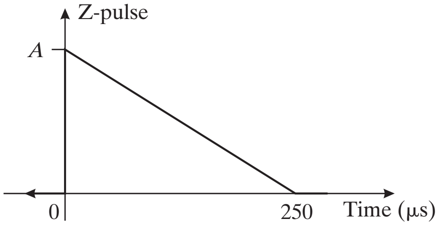


Figure P8.8
Z-pulse waveform. See Problem 8.19.

(**b**) Draw the combined Z-pulse arising from two photons, one striking the camera at $t = 0$ and the other at $t = 100$ μs. What will the output of the pulse height analyzer be? (It should be a sequence of heights at different times.)

Assume the photopeak is set at A and a ±20% discriminator window is used.

(**c**) Because of the discriminator, the second photon may not be detected if it arrives too soon after the first. Determine the time separation required in order for the second photon to be detected as a separate event.

(**d**) Is it theoretically possible to complete the experiment in the time you computed in part (a)? Explain.

(**e**) Under what condition(s) will a photon be rejected because its Z-pulse height is too small? Explain.

(**f**) Assume that the experiment is concluded when 2,000,000 counts are accepted no matter how long it takes. What is the intrinsic SNR in a single pixel?

8.20 A small pellet is located at the origin and filled with a radiotracer, as shown in Figure P8.9(a). Assume a two-dimensional scenario in which the pellet emits gamma rays only in the x-y plane and the radioactivity is equal to $A = 0.027$ mCi.

(**a**) What is the average rate of photons hitting the detector (1-D Anger camera)?

Assume the detector is made of NaI(Tl) ($\mu = 0.644\ \text{cm}^{-1}$) and its thickness is $b = 2.5$ cm.

(**b**) What is the detector efficiency at the center of the detector?

(**c**) If the Anger camera is rotated stepwise around the origin and requires 2×10^5 detected counts at each orientation, how many orientations can be captured in 10 seconds? (Ignore rotation time.)

A parallel hole collimator with septal thickness $h = 5$ mm, septal height $l = 12$ cm, and hole diameter $d = 5$ mm, is mounted on the camera, as shown in Figure P8.9(b).

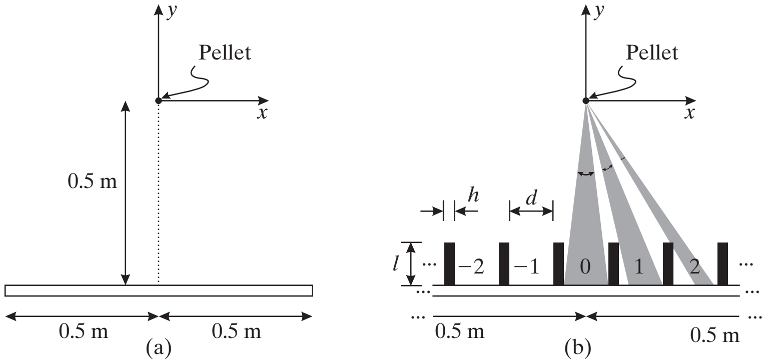


Figure P8.9
See Problem 8.20.

(d) What is the collimator resolution R_c at the range of the pellet?

(e) Assume that there is no Compton scattering and that the lead septa always absorb gamma rays incident upon them. Find the rate of photons hitting the detector in the three central holes. Plot this as a function of hole number for $n = -1, 0, +1$.

(f) What is the (approximate) resolution $\hat{R}_c$ that you would deduce from your result in part (e)? Compare with R_c and speculate on any difference you may observe.

8.21 An Anger camera collimator has the following dimensions: septal thickness $h = 6$ mm, septal height $l = 10$ cm, and hole diameter $d = 3$ mm. The detector thickness is $b = 2.5$ cm. The range $z = 0.5$ m.

(a) Find the collimator resolution.

(b) Suppose the intrinsic resolution of the Anger camera is 0.2 mm. Write an expression for the intrinsic PSF assuming it is Gaussian.

(c) Find the overall resolution of the Anger camera.

The final image (static frame mode) on the Anger camera is shown in Figure P8.10.

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	3	3	3	3	3	0	0
0	0	3	8	8	8	3	0	0
0	0	3	8	8	8	3	0	0
0	0	3	8	8	8	3	0	0
0	0	3	3	3	3	3	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

$\times 10^6$

Figure P8.10
See Problem 8.21.

(d) Find the total count of the scan.

(e) Find the local contrast and SNR of the object comprising the nine pixels at the center of the camera.

Emission Computed Tomography

CHAPTER 9

Planar scintigraphy is the only direct imaging approach in nuclear medicine (i.e., an image exists directly at the end of frame mode acquisition). The other two imaging methods in nuclear medicine—single photon emission computed tomography (SPECT) and positron emission tomography (PET)—require computed tomography image reconstruction methods. Like computed tomography (CT), these modalities require the acquisition of projection data, which are then reconstructed into transverse (or transaxial) slice images. SPECT systems can be designed around a customized array of detectors but are usually based on one or more high-quality Anger cameras, which rotate around the body to acquire a series of projection images with different angular orientations. Multiple cross sections of the body can be reconstructed—in parallel, in principle—from these data. It is reasonable to say that SPECT is to planar scintigraphy as multiple-row detector CT is to projection radiography.

PET is in a class of its own, however, having no suitable analogy in other imaging modalities. PET measures the location of the line on which a positron is annihilated, an event that produces two simultaneous 511 keV gamma photons (or *annihilation photons*). After accumulating many such *coincidence lines*, line integrals of activity are produced and computed tomography reconstruction methods are directly applied.

PET is unique in both its physics and chemistry. The most commonly used PET radionuclides, carbon-11, nitrogen-13, oxygen-15, and fluorine-18, are isotopes of elements that occur naturally in organic molecules. (Fluorine usually does not, but it is a bioisosteric substitute for hydrogen.) Thus, radiopharmaceutical synthesis is simplified, and the tracer principle (which mandates as small a change in the molecule to be traced as possible) is better satisfied. Indeed, useful PET radiopharmaceuticals are now available to measure *in vivo* such important physiologic and biochemical processes as blood flow, oxygen, glucose, and free fatty acid metabolism, amino acid transport, and neuroreceptor density. The short half-lives of the radionuclides (e.g., carbon-11, 20 minutes; nitrogen-13, 10 minutes; oxygen-15, 2 minutes; fluorine-18, 110 minutes) permit the acquisition of serial studies on the same day without background activity from prior injections interfering with the measurements.

The physics of PET permits greater quantitative accuracy and precision. The use of small, high-density detector crystals improves spatial resolution (about 4 mm in the best commercial whole-body PET scanners). The lack of collimation to determine photon direction dramatically increases sensitivity. Finally, coincidence detection allows mathematically accurate attenuation correction, as we shall see.

We now explore the instrumentation, reconstruction methods, and image quality of both SPECT and PET, which together comprise *emission computed tomography* (ECT) systems.

9.1 Instrumentation

9.1.1 SPECT Instrumentation

SPECT refers to true transaxial tomography with standard nuclear medicine radiopharmaceuticals (i.e., those that emit a single photon upon decay). SPECT is performed with either specialized geometry detector systems or rotating Anger cameras. By far the most popular method of doing SPECT is with a rotating Anger camera, or more often multiple cameras, mounted on a special gantry that allows 360-degree rotation around the patient, and we mainly focus on this approach. A typical rotating Anger camera system is shown in Figure 9.1.

In rotating Anger camera SPECT, the camera (also referred to as a *head*) is usually rotated either 180° or 360° (assuming a single-head system). At each angle, a projection image (a standard planar image) is acquired so that, effectively, multiple 1-D projections from different (contiguous) cross sections of the body are acquired simultaneously. After complete rotation of the camera, each cross section is reconstructed separately with either convolution backprojection or an iterative reconstruction method, yielding a collection of transverse images,

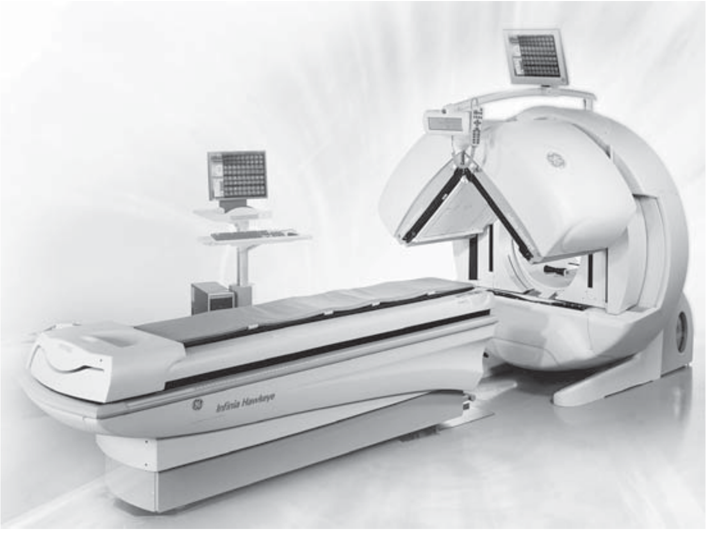


Figure 9.1
A dual-head SPECT system. Courtesy of GE Healthcare.

stacked to form a reconstructed volume. We will have much more to say about this in Section 9.2.

Two types of collimators are used in a rotating-camera SPECT system (see Section 8.2.1). Most often, a parallel-hole collimator is used; this geometry corresponds to that of parallel-beam geometry in CT. Here, the geometry is parallel both in-plane and in the axial direction. In some cases, a *fan-beam* collimator is used. This collimator converges in-plane, like a converging collimator but has parallel-hole geometry in the axial direction. With fan-beam collimation in each contiguous 1-D projection, the "stack" has fan-beam geometry, but the slice-to-slice geometry is parallel. In rare cases until now (although increasingly more often than in the past), a collimator-detector system with more complicated geometry is used, requiring more sophisticated reconstruction algorithms. Such systems are briefly considered below.

Perhaps the biggest advance in the most commonly used clinical SPECT systems is the incorporation of multiple Anger cameras—a dual-head system is shown in Figure 9.1. As shown in Table 9.1, adding camera heads increases sensitivity, all else being equal. Dual-head variable-angle systems that are capable of orienting the heads at both 90° and 180° are particularly popular as this optimizes the acquisition geometry for the two most commonly performed nuclear medicine procedures: cardiac SPECT and whole-body bone scans. These two types of studies account for at least 70 percent of all nuclear medicine imaging procedures. Cardiac SPECT is often performed with 180° acquisition (i.e., only a 180° arc of projection data around the heart is acquired), whereas many other SPECT studies use 360° of projection data. This has been a controversial issue for over two decades now. Because of photon attenuation, the "front" 180° of data and the "back" 180° of data are not identical. In some cases, it thus helps to use the full 360° of data; in others (i.e., when the organ of interest—such as the heart—is located towards the front or the back), it may be detrimental to image quality to use the full 360° of data.

The increased sensitivity provided by multihead systems can be used primarily in three ways. First, it can be used to decrease noise while using the same acquisition time as with a single-head system. Second, it can be used to decrease acquisition time to get the same counts as a single-head system; this decreased acquisition time increases patient throughput, decreases motion (and associated

TABLE 9.1

Comparison of Acquisition Times and Relative Sensitivities for Single- and Multihead Systems with Identical Camera Heads and Collimation

	360°		180°	
	Acquisition Times	Relative Sensitivities	Acquisition Times	Relative Sensitivities
Single	30	1	30	1
Double (heads@180°)	15	2	30	1
Double (heads@90°)	15	2	15	2
Triple	10	3	20	1.5

blur), and decreases tracer washout effects (which might lead to a mathematically inconsistent projection dataset). Third, it may be "traded" for higher resolution through the use of higher resolution/lower sensitivity collimators.

EXAMPLE 9.1

Suppose a single-head system requires N counts in a 30-minute scan using an all-purpose collimator. Let us imagine performing the same study with a two-head system instead, and let the two new collimators have higher resolution with just 75 percent of the sensitivity of the all-purpose collimator on the single-head system.

Question How long will it take to achieve the same counts as the single-head system, and will the image quality be equivalent?

Answer In 30 minutes, each (high-resolution) head collects $0.75 \times N$ counts (on average, of course). This corresponds to $2 \times 0.75 \times N$ counts in 30 minutes. The time t to collect just N counts satisfies

$$\frac{t}{30 \text{ minutes}} = \frac{N}{2 \times 0.75 \times N},$$

so

$$t = \frac{30 \text{ minutes}}{1.5} = 20 \text{ minutes}.$$

The image obtained in just 20 minutes will be better because, although the total counts are the same in the two scans, the resolution is better in each of the dual-head collimators.

■

Multihead systems have permitted a wide variety of SPECT scans to be performed in less than 30 minutes. This is an important time milestone because beyond 30 minutes patient motion becomes a very significant factor in image quality. Three-head systems exist, and it would seem that M-head systems (where $M > 3$) would be in the offing. However, simple geometric constraints (packing that many heads in close to the patient), quality control issues (how do you keep all those detectors in calibration?), and cost (M heads means M collimators and head components) make it impractical to go much further in this design direction. For higher sensitivity, manufacturers have gone to specialized geometries, as discussed below.

Studies have shown that Anger cameras must have significantly better performance for SPECT than for planar scintigraphy. For example, nonuniformity must be reduced to less than 1% in order to avoid reconstruction artifacts. This requires acquisition of a 30–120 million count reference image of a uniform field for subsequent computer correction of nonuniformities. (There should be around 10,000 counts per pixel in order for the relative Poisson variation or noise in the uniformity calibration to be $< 1\%$ ($\sqrt{10{,}000}/10{,}000$), so a longer reference scan is required for 128×128 acquisition versus 64×64 acquisition.) The camera image must also be mechanically properly aligned within the computer matrix, or an *axis-of-rotation correction* made. In addition, there are several other factors that must be accounted for to make SPECT truly quantitative. In the past, there was disagreement about the need for attenuation, scatter, and spatial resolution corrections for subjective visual interpretation of SPECT images, but it is now generally accepted that such corrections are desirable for both visual interpretation and quantitative analyses.

Specialized Geometry SPECT Systems Several manufacturers have released commercial SPECT systems that rely on a combination of more exotic detectors and specialized geometry in an attempt to significantly enhance imaging performance, especially detection sensitivity. Many of these systems utilize a solid-state detector, cadmium-zinc-telluride (CZT), which we mentioned in Section 8.2.8. CZT has significantly better energy resolution than NaI along with higher count-rate capability. It can be manufactured as a pixelated detector; with 1 × 1 mm pixels, intrinsic spatial resolution < 2 mm can be achieved. The specialized SPECT systems combine CZT or another detector material with unique collimation approaches; for example, multiheads and a multipinhole array. A picture of one such system is shown in Figure 9.2.

Some of these approaches involve movement of the detectors and/or collimators; others do not. In all cases, the specialized geometry yields significantly higher detection sensitivity, by a factor of as much as eight, particularly for cardiac imaging (for which many of these systems have been specifically optimized). This increased sensitivity can be used to decrease dose, imaging time, and/or image noise. All such unique approaches require specialized reconstruction methods.

Combined SPECT/CT Systems As we have emphasized throughout this book, each imaging modality yields a very different type of signal or signals, which is why each is clinically useful and none are redundant. Nuclear medicine images depict *function* (e.g., the activity of a physiological or biochemical process) rather than *structure* or *anatomy*. However, the spatial resolution in most nuclear medicine images is no better than several millimeters, and the radiotracer's distribution does not necessarily follow anatomical borders, making anatomic localization difficult. Accordingly, there has been a push in emission tomography, especially in PET (as we shall see below) and also in SPECT, to build combined systems that also incorporate CT. In such systems,

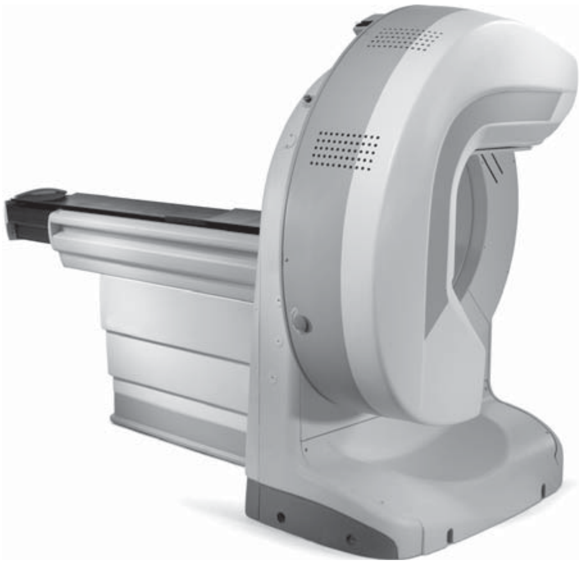


Figure 9.2
A GE Discovery NM 530c cardiac SPECT system. Courtesy of GE Healthcare.

registered SPECT or PET and CT slices are obtained, and the emission images can be superimposed on the CT slices (with the emission activity typically depicted in color, and the CT in shades of gray). This type of display makes it much easier to determine the anatomic location of any increased (or decreased) area of activity.

The first combined SPECT/CT systems used low-output x-ray tubes, producing CT images of relatively poor quality (i.e., limited spatial resolution and higher noise than conventional diagnostic CT images); they were not considered usable for diagnosis, only for anatomic localization. Most current-generation SPECT/CT systems use high-output tubes and produce CT images of routine diagnostic quality.

Of importance, the CT images can also be used to help determine the attenuation experienced by the photons emitted from the radiotracer; this attenuation effect, which confounds the measurement of activity, is intrinsic to SPECT imaging [see (9.2) and (9.6)]. The CT images provide a transmission dataset for use in reconstruction, as described in Section 9.2.1 below.

9.1.2 PET Instrumentation

Positron emitting radionuclides are used in PET; the radiopharmaceuticals make PET a functional imaging technique and in the family of nuclear medicine modalities. A typical PET system is shown in Figure 9.3. From the outside, we only see the plastic housing, and a PET system looks just like a CT or magnetic resonance imaging (MRI) system, but the hardware is quite different in PET.

Consider a positron-emitting radiopharmaceutical distributed in a patient and refer to Figure 9.4 for the following discussion. When a positron is emitted, it travels up to several millimeters in tissue depositing its kinetic energy. It then meets a free electron in the tissue, and mutual annihilation occurs (recall that a positron is an antielectron). From conservation of energy—noting that 511 keV is the energy equivalent to the rest mass of an electron or positron—two 511 keV annihilation (gamma ray) photons appear. From conservation of momentum, the two 511 keV photons are emitted back-to-back; that is, they propagate outward from the site of annihilation (almost exactly) 180° apart.

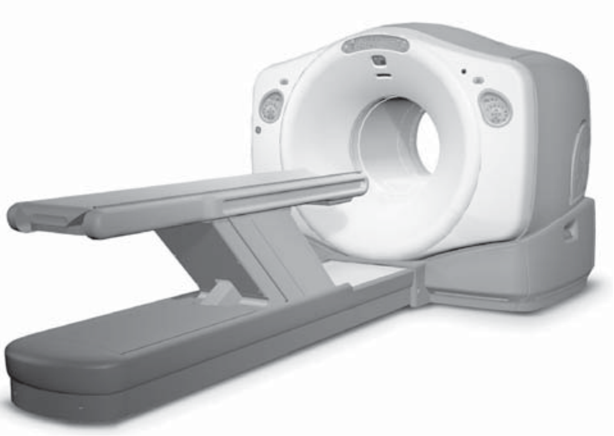


Figure 9.3
A PET system. Courtesy of GE Healthcare.

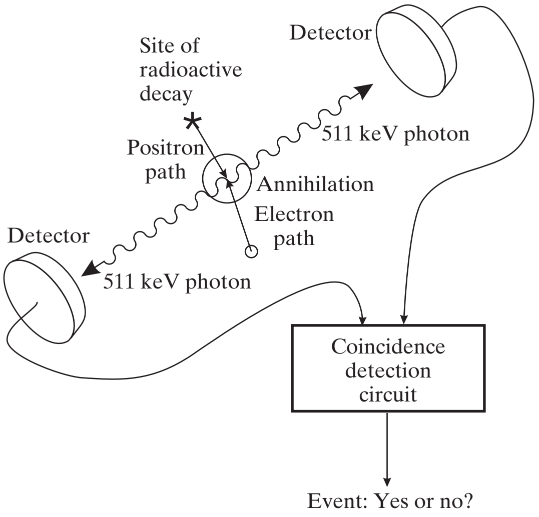


Figure 9.4
Coincidence detection due to positron decay and annihilation.

It is possible to use a single-head Anger camera to individually detect just one of the two 511 keV gamma ray photons that are emitted by positron-electron annihilation. Images could then be created using planar scintigraphy or SPECT reconstruction (if the camera rotates). With a multihead system, both photons might be detected if two of the Anger cameras were 180° apart and the photons made it through the opposing collimators. In this case, scintigraphy or SPECT techniques could be used to create an image as well.

However, it is more advantageous (and this is what distinguishes PET from SPECT) to exploit the fact that a positron annihilation produces two gamma rays at the same time. PET uses electronic circuits, as shown in Figure 9.4, to detect a pair of 511 keV photons at the same time, in what is called *annihilation coincidence detection* (ACD). With this approach, only simultaneous gamma rays are declared as events, and all other events are rejected. Whether two detected photons are declared as "simultaneous" depends on a user-set time interval, called a *time window*, which is typically 2–20 ns for modern scanners.

Of major importance, if two opposing detectors simultaneously detect the (paired) 511 keV gamma rays, the annihilation event must have occurred somewhere along the line joining the two detectors (the so-called *coincidence line*). Said another way, ACD intrinsically provides information about the direction of travel of the photons. Detector collimation is not required with this approach and is in fact undesirable because it reduces the sensitivity of the detectors and may cause undesirable directional dependencies if the detectors remain in fixed positions.

Commercial dual-head SPECT scanners could be used, together with ACD circuitry, for PET (and some were sold this way in the early days of PET). In this case, the two heads are placed 180° apart and their collimators removed (or replaced with special collimators that isolate axial planes only). The heads must be rotated in order to pick up emissions from all angles around the patient, as tomographic reconstruction methods are used to reconstruct cross-sectional images (see Section 9.2). In order to efficiently stop 511 keV photons, it is generally necessary to make the NaI(Tl) crystals (in the Anger cameras) thicker, which increases positioning uncertainty. Because of this and because of the inefficient angular acquisition, dual-head SPECT scanners are not an ideal solution for PET, and commercial PET systems based on this approach have not been available for a long time.

Full-fledged PET systems surround the patient with multiple rings of detectors, each having a diameter of about 100 cm, as shown in Figure 9.5. Each detector is electronically coupled to every other detector in a comprehensive ACD circuit. Older systems used a single scintillation crystal coupled to a PMT as the basic detector. The requirement for better resolution, however, has led to the development of so-called *detector blocks* that are based on the Anger camera principle. As an example, each detector block might comprise a single scintillation crystal backed by four PMTs, which are arranged in a 2 × 2 matrix, as illustrated in Figure 9.6. The scintillation crystal is cut into an array of smaller crystals, typically 8 × 4, 8 × 7, or 8 × 8 arrays, which partially isolate the light burst from each scintillation event to one of these subcrystals. Relative PMT signal heights are used to identify which of the subcrystals contained the event in a crude approximation of the center of mass calculation, as in a conventional Anger camera (see Problem 9.7).

With the block detector approach, a full PET scanner consists of a number of these blocks. As an example, a PET scanner might have three primary rings of detectors with 48 detector blocks in each ring for a total of 144 detector blocks (3 × 48). Each detector block has 64 subcrystals backed by four PMTs, which means that we can also think of this PET scanner as having 24 rings (3 rings × 8 subcrystals) and 384 detectors per ring (48 detector blocks × 8 subcrystals).

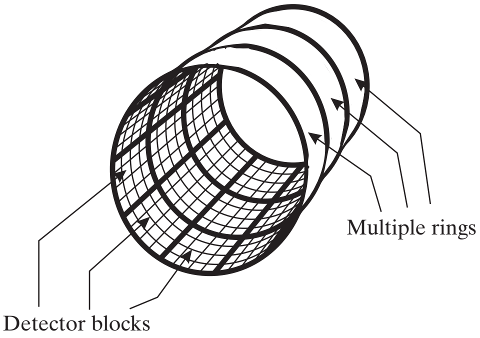


Figure 9.5
Geometry of a multiple-ring PET system.

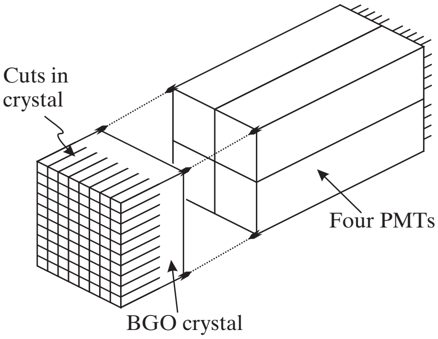


Figure 9.6
Geometry of a PET detector block.

As the quest for better and better spatial resolution continues, limitations of the current detector block approach become significant; these include concerns with efficiency, resolution, and contrast. Accordingly, cutting-edge approaches are being developed, including the use of the very same solid state approaches as discussed for planar scintigraphy in Section 8.2.8.

PET scanners have lead or tungsten *septa* between the detector rings (although this is increasingly rare because of the routine use of 3-D mode; see below). The purpose of the septa is to collimate the photons into a set of 2-D slices. In this mode, PET reconstruction methods reduce to 2-D reconstruction applied to each slice or cross section separately. This geometry reduces the contribution of scattered photons between slices and is very similar to that of SPECT. When the septa are removed, the scanner operates in 3-D mode, and coincidence events between any crystal pair are accepted. This fully 3-D ("septa-less") acquisition and reconstruction is a big advance; the use of 3-D mode provides about a factor of 5–10 increase in sensitivity. This increase in sensitivity comes with a cost—significantly higher acquired scatter events (*scatter coincidences*, in which at least one of the two annihilation photons scatters in the patient before detection)—but accurate postacquisition software algorithms now exist for scatter correction in 3-D PET.

The original commercial PET scanners used sodium iodide NaI(Tl) as the detector scintillation crystal, the same material used in planar scintigraphy and SPECT. Since 511 keV photons are harder to stop than the lower energy photons used in planar scintigraphy and SPECT, next-generation PET systems used bismuth germanate (BGO) instead of NaI(Tl) as the scintillation crystal. The linear attenuation coefficient of BGO at 511 keV is 0.964 cm^{-1} as opposed to 0.343 cm^{-1} for NaI(Tl). Therefore, BGO detectors can be about one-third the thickness of what would be required for NaI(Tl) in order to yield the same stopping power. For example, a typical BGO crystal is about 3 cm deep. Each crystal is typically "cut" into smaller subcrystals that are on the order of 6×6 mm. The downside of BGO as compared with NaI(Tl) is that it is only 12–14 percent as efficient in light conversion.

Both NaI(Tl) and BGO are relatively slow scintillators having *scintillation decay times* of 230 ns and 300 ns, respectively. (In this context, the scintillation decay time is the time it takes for the scintillation light output to decay to e^{-1} of its initial value.) Faster scintillators would be advantageous because it would permit narrower time windows to be used in ACD circuits. Narrower time windows would in turn reject more *random coincidences*, which are the detection of two photons that are deemed to be from the same annihilation event but are actually from two different annihilations. Accordingly, current-generation PET systems use other detector materials, including cerium-doped lutetium oxyorthosilicate (LSO) and gadolinium oxyorthosilicate (GSO), which are single-crystal inorganic scintillators. The decay times of LSO and GSO are on the order of 50 ns, a very significant improvement over NaI(Tl) and BGO. The attenuation coefficients are comparable, as well, so these crystals can be made relatively small for good spatial resolution and still retain good stopping power (which translates to good detection sensitivity).

Just as in planar scintigraphy, the outputs of PET detectors are analyzed by energy discrimination circuits. An energy discrimination window, centered on the photopeak, is usually desirable. In this way, photons that have been detected after Compton scattering in the patient are rejected, and multiple simultaneous detection of photons is also rejected. Sometimes, it is desirable to lower the discrimination window to include part of the Compton plateau. The reason for this is that it is common for there to be a Compton interaction within the detector, depositing a fairly large amount of energy, but allowing the photon to ultimately escape the detector without being fully absorbed. In this case, the primary Compton event is actually the correct event to detect as it represents the proper location of one of the two 511 keV photons created during positron annihilation. Setting the lower window threshold is therefore a matter of trading off the acceptance of two types of Compton events—those that occur in the patient, which we do not want to count and those that occur in the crystal, which we do want to count. A common lower energy threshold is 300–350 keV.

Time-of-Flight PET For most of the discussion in this chapter, the *difference* in arrival times (at the detectors) of the two 511 keV annihilation photons is not considered, as long as both arrivals are within the coincidence time window. However, it is important to note that the two photons only arrive at *exactly* the same time at two opposing detectors if the annihilation event is exactly halfway between the two detectors. Any offset will be reflected by one of the photons arriving sooner at its detector than the other at its detector. This difference in arrival times could be used to position the annihilation event *along* the coincidence line, potentially yielding better spatial resolution and/or higher signal-to-noise in the reconstructed image.

Bearing in mind that the two annihilation photons are traveling at the speed of light, very high temporal resolution is required to take advantage of this so-called *time-of-flight* (TOF) information in PET. In TOF-PET, the difference in arrival times is recorded and used in the reconstruction process. This arrival time difference increases as the annihilation event is farther from the midpoint of the coincidence line. Of importance, in order to have TOF information exactly locate the event, thus eliminating the need for reconstruction per se, the scanner's temporal resolution would have to be on the order of 25 picoseconds. Unfortunately, modern scanners have temporal resolution of a few hundred picoseconds (equivalent to spatial resolution of a few centimeters if we only use TOF information to position the annihilation event), so some sort of reconstruction is still required. However, the TOF information can be used to constrain the reconstruction algorithm.

The approximate improvement in SNR over that obtained with non-TOF PET is:

$$\mathrm{SNR}_{\mathrm{TOF}} = \sqrt{\frac{2D}{c\Delta t}}\,\mathrm{SNR}_{\text{non-TOF}}\,, \tag{9.1}$$

where D is the diameter of the object being imaged, c is the speed of light, and Δt is the temporal resolution of the scanner. All else being equal, the improvement in SNR is greatest for larger-sized objects (e.g., the human body as opposed to the brain) since larger-sized objects lead to greater differences in arrival times. In order to improve Δt, very new detectors with a short scintillation decay time, such as lanthanum bromide ($\mathrm{LaBr_3}$), are used.

Combined PET/CT Systems As noted above, in discussing nuclear medicine we always emphasize the "functional" nature of the image data. There is now ample evidence that clinical interpretation of nuclear medicine studies, especially in PET, can benefit from reference to anatomic images from CT or MRI. In the past, a physician would review the PET images on a computer display and the CT films on a light box and then perform mental "image fusion." This prompted the development of software for postacquisition image registration and fusion. This process requires the images from both procedures to be on a common computer (typically via transfer of the CT exam to a nuclear medicine computer workstation); the PET images are interpolated to a common voxel size and registered with the CT exam, and the fused or overlaid images are displayed.

At least in theory, the best way of registering PET and CT is if both studies are acquired via a combined system with a common gantry and patient table. Accordingly, virtually all PET equipment manufacturers have developed combined PET/CT systems, and these represent the vast majority of PET systems sold today.

In a combined PET/CT system, both PET and CT exams may be performed without moving the patient relative to the table between the two exams. This provides functional and anatomical images that can be registered and fused with a higher degree of positional accuracy than is permitted by imaging on two separate scanners, especially because of differences in body shape and position introduced by the use of two different tables. The improved registration accuracy with a combined PET/CT system thus mainly results from the patient being in the same position on one table. However, physiologic motion such as breathing and abdominal motion cannot be avoided. These motions decrease the registration accuracy and may produce artifacts in attenuation-corrected PET images using the CT map (see below). Notice that the CT exam is typically taken during a breath-hold, but the PET exam requires breathing. In addition to improved positional accuracy, there is easier availability of both the PET and CT data for the fusion software with a combined PET/CT system since they are combined within one system.

As with combined SPECT/CT systems, in addition to their use for anatomic localization, the CT images may be used as attenuation maps for performing attenuation correction on the PET images. As noted above, one of the theoretical advantages of PET over SPECT is that measured tissue attenuation values can be used to exactly correct for the effects of attenuation in the emission data, as discussed below in Section 9.2.2. In this regard, the attraction of using CT is that it provides high resolution, high contrast, and low noise attenuation maps obtained quickly using a fast helical scan. As with SPECT, note that using x-ray CT images for attenuation correction requires converting attenuation values measured with a low-energy, polychromatic x-ray beam into 511 keV, monoenergetic attenuation values.

Combined PET/MRI Systems As we will subsequently discuss in forthcoming chapters, MRI offers certain advantages over CT. Accordingly, it is not surprising that manufacturers have been developing combined PET/MRI systems, and these are now becoming commercially available. However, combining PET and MRI is technically challenging. The high static magnetic field, quickly changing gradient fields, and radio frequency signals from the MRI portion of the system interfere with PMTs and some of the PET electronics, which adversely affect the PET images. In addition, the PET detectors may produce inhomogeneities in the MRI magnetic field, as well as electromagnetic interference, both of which adversely affect the MR images.

Several approaches are being used to address these challenges, including using two separate scanners with a common patient bed (akin to the conventional PET/CT approach, which yields a very long gantry), integrating the PET detectors in the MRI scanner and guiding the scintillation light to an enclosure outside the MRI system proper, using a custom magnet architecture, and using new approaches to PET detectors, including the substitution of avalanche photodiodes for PMTs.

9.2 Image Formation

In considering reconstruction in emission tomography, we focus on the reconstruction of a 2-D cross-sectional image from observations of line integrals within the cross section. This approach is consistent with SPECT (with suitable approximation) and 2-D PET. It is completely analogous to our development of CT reconstruction methods in Chapter 6, upon which we draw heavily.

9.2.1 SPECT Image Formation

Coordinate Systems Our first task is to set up a convenient (and conventional) coordinate system. Since the Anger camera rotates in SPECT, it is necessary to have both a laboratory (fixed) coordinate system and a rotating coordinate system; these are both shown in Figure 9.7. It is important to realize that because of the requirement for rotation and to conform with our previously defined notation for tomographic reconstruction, the notation shown in this figure differs from the "camera-centric" notation that we used for planar scintigraphy in Chapter 8.

Applying the same approximations that we used in developing the imaging equation for planar scintigraphy (see Section 8.3), we assume that the photon

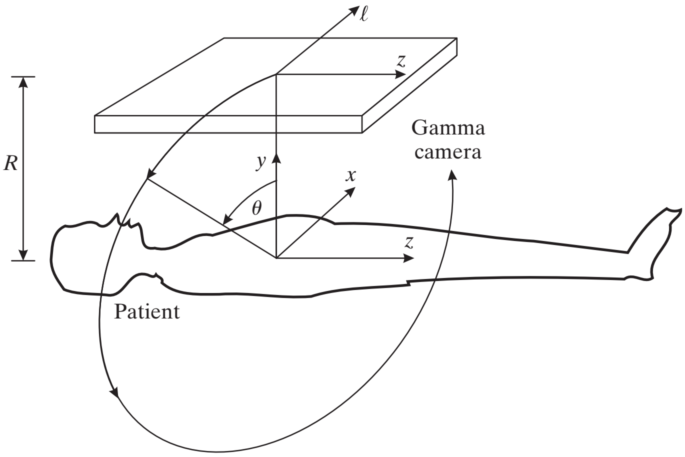


Figure 9.7
Geometry of a SPECT system.

fluence rate is constant within each pixel on the gamma camera. Therefore, given the geometry in Figure 9.7, a cross-sectional image acquired at a fixed position z—referred to as an *axial cross-section*—must use pixels located at position z on the camera. These pixels will "see" an underlying radiotracer activity concentration $f(x, y, z)$ only within the slice at z, and so we can identify and denote the image to be reconstructed as $f_z(x, y)$. When this image is viewed in a conventional orientation (positive x to the right and positive y pointing up), it is as if we are standing at the subject's feet and looking at a slice of their body. This is a conventional radiographic view of the cross-sectional image.

In principle, the imaging equation for SPECT (using an Anger camera) would be identical to the imaging equation for planar scintigraphy [see (8.9)]. However, because of our choice of a laboratory coordinate system, it is necessary to use a coordinate system on the face of the Anger camera that is different from the one that is used in Chapter 8, and a somewhat different imaging equation will emerge below. As shown in Figure 9.7, we use the variable ℓ to coincide with the x-axis when the rotation angle θ is zero. The other coordinate on the camera is z, coinciding with the laboratory coordinate z regardless of orientation. One further point is that the Anger camera is assumed to remain at a fixed distance R from the origin. With this choice, the coordinate ℓ will represent the lateral position of a line integral, and the geometry used in the reconstruction of a cross section in SPECT corresponds exactly to that in CT.

Imaging Equation We assume parallel-hole collimators, which implies parallel-ray geometry. When the camera is located in its "home" position $\theta = 0$, we can map the coordinate system used in planar scintigraphy (Chapter 8) into that of SPECT. Letting $x \to z$, $y \to \ell$, and $z \to y$ (compare Figure 8.9 and Figure 9.7), we can rewrite (8.7) as

$$\phi(z, \ell) = A_b \int_{-\infty}^{R} \frac{f(x, y, z)}{4\pi (y - R)^2} \exp\left\{ - \int_{y}^{R} \mu(x, y', z; E)\, dy' \right\} dy\,. \tag{9.2}$$

Notice that there is a slight abuse of notation in keeping the arguments of f and μ the same as in the laboratory frame; the present functions refer to the activity concentration and linear attenuation coefficients defined on the laboratory frame in Figure 9.7. Although the camera acquires data for all z positions while in any fixed rotation, the two-dimensional reconstruction process of any slice is the same. In what follows, we drop the explicit dependence on z, but do not forget that there is a separate reconstruction problem (which can be implemented on a parallel computer, in fact) for each z coordinate on the camera.

When the camera is rotated, the lines of integration are no longer in the y direction. By design of this coordinate system, the lines of integration are the same as those in CT [see (6.6)],

$$L(\ell, \theta) = \{(x, y) \mid x \cos\theta + y \sin\theta = \ell\}\,. \tag{9.3}$$

Because of the exponential term in (9.2), we cannot use the set form of the line integral (i.e., using the impulse function). Instead, we use the parameterization

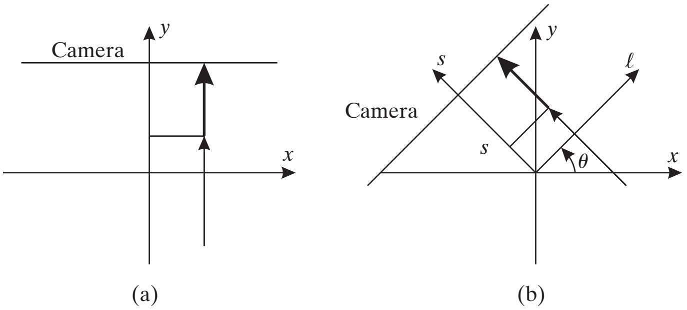


Figure 9.8
Integration geometry for SPECT imaging equation.

[introduced in (6.8) and (6.9)]

$$x(s) = \ell \cos\theta - s \sin\theta\,, \tag{9.4}$$

$$y(s) = \ell \sin\theta + s \cos\theta\,. \tag{9.5}$$

Examination of the geometry shown in Figure 9.8, recognizing that the mean pixel count is related to the underlying photon fluence rate by $\overline{n}_k = \epsilon T A_k \phi(x_k, y_k)$ [see (8.8)], and careful substitution yields

$$\overline{n}_k(\ell,\theta) = \epsilon T A_k A_b \int_{-\infty}^{R} \frac{f(x(s),y(s))}{4\pi(s-R)^2} \exp\left\{ - \int_{s}^{R} \mu(x(s'),y(s');E)\,ds' \right\} ds\,, \tag{9.6}$$

where $\overline{n}_k(\ell,\theta)$ denotes the mean pixel count at lateral position ℓ and angle θ.

Equation (9.6) is the starting point for a SPECT imaging equation. Unfortunately, there are two complications that make (9.6) generally intractable as an imaging equation. The first complication has to do with the presence of a second unknown function, $\mu(x,y)$, the linear attenuation coefficient. There has been a considerable amount of research into this problem; even when μ can be measured, inferred, or approximated, the presence of the additional term in the integrand leaves a relatively difficult reconstruction problem at hand—one with no closed-form analytic solution. This is essentially one side of the second complication, which arises due to the position-dependent terms—the inverse-square law and the photon attenuation—in the integrand. Also, be aware that the same collimator sensitivity considerations that were discussed in Section 8.4.2 apply here as well.

Instead of trying to solve (9.6) directly, we will develop a simplified form for image reconstruction—one that was used for many years in practice—by ignoring both the inverse square law and the attenuation terms in the integrand. Using this bold approximation yields

$$\overline{n}_k(\ell,\theta) = \epsilon T A_k A_b \int_{-\infty}^{\infty} f(x(s),y(s))\,ds\,, \tag{9.7}$$

where the upper limit of integration has been set to ∞ since $f = 0$ behind the collimator. We can now divide by the constant terms to form

$$g(\ell,\theta) = \frac{\overline{n}_k(\ell,\theta)}{T A_k A_b \epsilon}\,, \tag{9.8}$$

which we recognize from (9.7) as the line integral of activity concentration. Therefore, using the impulse function [see (2.3)] and referring to (9.7), (9.8) can be written as

$$g(\ell, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(x \cos\theta + y \sin\theta - \ell)\, dx\, dy\,, \tag{9.9}$$

which is precisely the form of the standard CT imaging equation [see (6.10)]. One important difference between the CT and SPECT cases is that the SPECT imaging equation uses the actual photon counts with a simple scaling factor in arriving at a Radon transform relation, while CT requires a logarithmic transformation of the observed x-ray intensities.

Reconstruction Because of approximations made above, we can use the same CBP reconstruction formula in SPECT as in CT [see (6.26)]

$$\hat{f}(x, y) = \int_{0}^{\pi} \int_{-\infty}^{\infty} g(\ell, \theta) \tilde{c}(x \cos\theta + y \sin\theta - \ell)\, d\ell d\theta\,. \tag{9.10}$$

This inverse formula is commonly used in SPECT even if only as a starting point for iterative reconstruction (see Section 9.2.3).

In (9.10), $\tilde{c}$ is an approximate ramp filter, defined in (6.27) and repeated here for convenience:

$$\tilde{c}(\ell) = \mathcal{F}_{\mathrm{1D}}^{-1}\{|\varrho| W(\varrho)\}\,. \tag{9.11}$$

In CT, the definition of the ramp windowing filter $W(\varrho)$ is not very critical. As long as its cutoff frequency ϱ_0 is high enough, there will be very little degradation in image quality. In SPECT, however, the windowing filter is very critical to image quality. In particular, if it is chosen to allow too many high frequencies through, then the SPECT images will be dominated by noise due to the high-frequency emphasis of the ramp filter. If, on the other hand, $W(\varrho)$ is chosen to pass only the very lowest spatial frequencies, then image detail will be absent. It is common to use a ramp combined with some sort of low-pass filter, such as a Butterworth filter, which is defined as:

$$W(\varrho) = \frac{1}{\sqrt{1 + \left(\frac{\varrho}{\varrho_c}\right)^{2n}}}\,, \tag{9.12}$$

where ϱ_c is the filter's critical or cut-off frequency and n is the order of the filter.

Attenuation Correction Attenuation influences the relative, as well as absolute, apparent distribution of activity within an image; it is thus important to try to correct for this effect even if only subjective visual interpretation—rather than absolute quantification—of the images is used. We used a simplified imaging equation (9.7), which does not include the effects of attenuation, in order to derive the convolution backprojection formula for SPECT that is given in (9.10). This was necessary because the influences of the activity concentration f and the attenuation factors are not separable in (9.6), and therefore no closed form solution exists to find f from the observed pixel counts. However, once an approximate image of f is reconstructed using (9.10), it is straightforward to deduce a first-order *attenuation correction factor* $a(x, y)$ that can be applied to improve the image appearance and quantification.

Suppose we knew the linear attenuation coefficient $\mu(x, y; E)$ on the plane of interest (perhaps by measuring it using CT) and imagine that we could distribute radioactivity uniformly—that is, $f(x, y) = 1$—throughout this object. From (9.6) we can then compute the expected measurements $\overline{n}_k(\ell, \theta)$ that would be theoretically produced by this object and then reconstruct an estimate of this object using (9.9). Let us denote the reconstructed image of this artificial uniform field by $a(x, y)$. We know that the values of $a(x, y)$ will be less than unity since attenuation has been applied numerically. It will also be "cup shaped," having lower values in the deep interior due to the higher overall attenuation. Since we have incorporated the attenuation factors (inverse square law and linear attenuation coefficient) in computing the measurements, the reconstructed image $a(x, y)$ will be degraded by the same factors as the true activity distribution that we are trying to reconstruct. In the case of the simulation, we know the "truth" $f(x, y) = 1$ can be achieved by simply dividing $a(x, y)$ by itself.

As a first-order correction, we can therefore apply the same correction factor to our actual reconstructed image, yielding the attenuation compensated image

$$\hat{f}_c(x, y) = \frac{\hat{f}(x, y)}{a(x, y)} . \tag{9.13}$$

Although (9.13) looks on the surface to be a valid correction method—and it is better than nothing—it is only an approximation because the inverse Radon transform is not theoretically exact for this problem since the observation equation (without approximation) is not a pure line integral. The notion of computing the expected observations is a very powerful concept, however, and it forms the basis of iterative reconstruction methods (see Section 9.2.3) in which attenuation correction can be incorporated iteratively and with theoretical justification.

Any attenuation correction approach requires knowledge of the distribution of the linear attenuation coefficients. This distribution is usually obtained from a transmission dataset. That dataset may come from either a CT (e.g., in a combined SPECT/CT system) or from scanning line sources (e.g., long-lived radioactive sources arrayed as lines that scan across the patient, integrated into the SPECT system). With CT, the measured Hounsfield units must be converted to attenuation coefficients at the emission energy of the radiotracer (e.g., 140 keV for Tc-99m). This conversion is typically a bi- or tri-linear function. The big advantages of CT are speed, low noise, and—in the case of combined SPECT/CT systems—perfect registration between the emission and transmission images. With scanning line sources, there is usually less need for any numerical conversion (because a radionuclide with approximately the same energy photons as the emission radiotracer is chosen), but transmission images from scanning lines sources are significantly noisier than those from CT, and the noise propagates into the attenuation-corrected emission images. (The spatial resolution of transmission images from scanning line sources is poorer than that from CT, but this is not as big a problem because the final (corrected) SPECT image has SPECT, not CT, resolution anyway.)

9.2.2 PET Image Formation

Coordinate System As noted above, modern commercial PET systems usually operate in 3-D mode. For simplicity's sake, here we consider only 2-D PET, the reconstruction of axial cross sections from data collected within isolated axial detector rings. In the case of a fully 3-D projection dataset, it is possible (and, indeed, common) to reorient the data into contiguous 2-D sinograms, for example, via *Fourier rebinning*. If we start with contiguous 2-D datasets, we need to consider only one detector ring, which can be taken to be in a fixed z plane. Geometry within this plane is considered relative to a conventional cross-sectional x-y coordinate system, as in CT and SPECT.

Lines of Response As with planar scintigraphy and SPECT, PET counts the number of detected photons, but these are now based on *annihilation coincidence detections* (ACDs) rather than single gamma ray detections. As shown in Figures 9.9 and 9.10, the line joining any two opposing detectors for which an ACD could be identified is called a *line of response* (LOR) or *coincidence line*. Over time, regions of high activity will generate more coincidence lines or

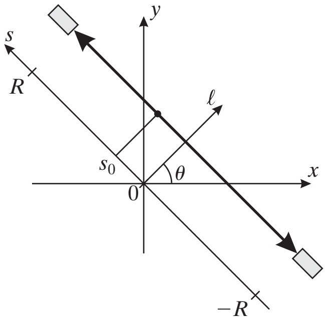


Figure 9.9
Integration geometry for PET imaging equation.

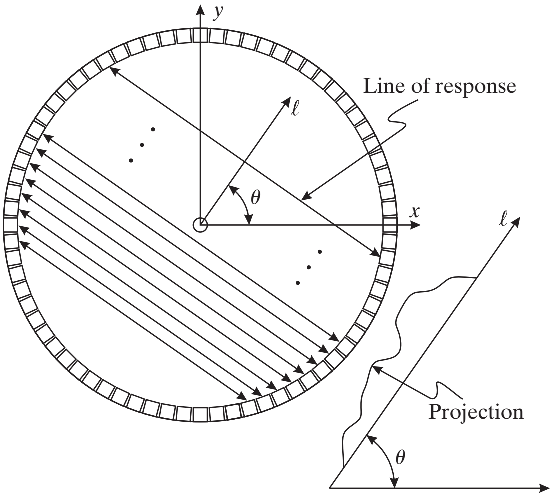


Figure 9.10
Lines of response organized into a parallel-ray projection.

coincidence events or counts along LORs passing through them than will regions of low activity. The total number of coincidence counts along any given LOR represents an integration of the total activity along that line. We note that it is not possible to know the position along an LOR where the annihilation took place (except for TOF-PET, which is not part of our development here). Therefore, the measured number of counts on a particular line represents an integration of the activity on that line. That is, like SPECT, PET measures line integrals of activity.

Imaging Equation Consider what happens at the site of a positron annihilation. Two gamma rays are generated, and they fly off in opposite directions. As in SPECT, the gamma rays will have potentially different experiences. A gamma ray can be absorbed in the body, be scattered, miss the detector ring, or hit a detector in the ring (and we consider only 2-D PET here). As in the development of our imaging equation for planar scintigraphy (which was also used in SPECT), we ignore scattering and therefore assume that photons travel in straight lines. Unlike planar scintigraphy and SPECT, each PET detector has a fixed spatial extent, and must be considered in pairs.

Consider a distribution of radioactivity within the body (or sample) described by a radiotracer activity concentration $f(x, y, z)$. Consider a differential volume *dxdydz* lying halfway between the two detectors on a line connecting the center of the detectors. For simplicity, we assume each detector is at a distance R from this point. Since the radioactivity of this differential source is given by (8.4), the photon fluence rate at a point (x_d, y_d) on either detector (ignoring attenuation for now) due to this differential source $dA = f(x, y, z)\, dx\, dy\, dz$ is

$$d\phi(x_d, y_d) = \frac{dA}{4\pi R^2}\,. \tag{9.14}$$

Since photons in PET come in pairs, for each photon that hits one detector we can expect a simultaneous hit on the opposite detector. But if either photon fails to reach the detector because of attenuation the event will not be counted by the PET system. This corresponds to a reduction in photon fluence rate due to attenuation.

To determine the factor associated with this reduction in fluence, we can carry out a brief "thought experiment." Since for each annihilation both photons must reach their target in order to be counted, the attenuation going both ways matters. That is, the attenuation from $s = 0$ to $s = R$ and $s = 0$ to $s = -R$ both can contribute to a reduction in the count. In fact, since the two paths are colinear, the total attenuation seen by this pair of photons is the same that would be experienced by a single photon originating at $s = -R$ on a trajectory through $s = 0$ and then on to $s = +R$. The photon fluence rate with attenuation included can therefore be written as

$$d\phi(x_d, y_d) = \frac{dA}{4\pi R^2} \exp\left\{-\int_{-R}^{R} \mu(s; E)\,ds\right\}\,. \tag{9.15}$$

It is the attenuation along the entire path that determines the loss factor for coincidence detection. That is, it does not matter where on this line the differential source is placed; the effect of attenuation on coincident counts will be the same.

Now consider moving the differential source to position $s \neq 0$ within the imaginary tube connecting the detectors (but still remaining on the central axis between the detectors). Now one detector is farther away from the differential source and will experience lower fluence simply due to the inverse square law. This leads to a slightly modified equation

$$d\phi(x_d, y_d) = \frac{dA}{4\pi(R+|s|)^2} \exp\left\{-\int_{-R}^{R} \mu(s'; E)\,ds'\right\}, \tag{9.16}$$

where (x_d, y_d) must be selected such that the line passing through (x_d, y_d) and the source position will also hit the opposite detector.

Now consider moving the source back to $s = 0$ but then move it off the central axis. Simple geometric considerations dictate that, for this source position, lines of coincidence between these two detectors cannot exist across the whole face of either detector. This amounts to a further reduction in effective photon fluence rate between these detectors from this shifted source. It is complicated to find this reduction factor for even simple detector geometries and not necessary for our purposes. Rather, we simply acknowledge that if we model the activity concentration as constant across the tube connecting the two detectors, its effective activity cannot be the product of the activity at the central point times the cross-sectional area of the tube; it will be less than that. For simplicity, we will use the notation $\tilde{A}_h$ for effective cross-sectional area of the tube lying between the two detector. With this assumption, the differential radioactivity under consideration then becomes a function of position along the tube

$$dA(s) = \tilde{A}_h f(x(s), y(s), z)\,ds, \tag{9.17}$$

where ds is a differential element in the direction of the line parameterized by $(x(s), y(s))$ connecting the detectors [see (9.4) and (9.5)]. Then the total fluence at (x_d, y_d) is given by the integral over the tube by combining (9.17) and (9.16), yielding

$$\phi(x_d, y_d) = \tilde{A}_h \int_{-R}^{R} \frac{f(x(s), y(s))}{4\pi(R+|s|)^2} \exp\left\{-\int_{-R}^{R} \mu(s', E)\,ds'\right\} ds, \tag{9.18}$$

where we have dropped the explicit dependence of f on z since this 2D analysis can be carried out for each z independently.

Considering the area A_h of the detector and image acquisition time T, we can write the expected number of coincidence detections for the LOR at (ℓ, θ) as

$$\overline{n}(\ell, \theta) = \epsilon T A_h \tilde{A}_h \int_{-R}^{+R} \frac{f(x(s), y(s))}{4\pi(R+|s|)^2} \exp\left\{-\int_{-R}^{R} \mu(x(s'), y(s'); E)\,ds'\right\} ds, \tag{9.19}$$

where we have also incorporated an additional (unitless) detector efficiency term ϵ. The exponential factor in (9.19) does not depend on s, so we can remove it from the integral over s. We will also make the approximation that the field of view containing radioactivity is very small with respect to R, which implies that $|s| \ll R$, and the inverse square law can also be removed the integrand. This leads to our final imaging equation for PET:

$$\overline{n}(\ell,\theta) = \frac{\epsilon T A_h \tilde{A}_h}{4\pi R^2} \int_{-R}^{R} f(x(s),y(s))ds \exp\left\{ - \int_{-R}^{R} \mu(x(s'),y(s');E)ds' \right\} . \quad (9.20)$$

Equation (9.20) is similar to (9.6) for SPECT, but the fact that the integrals of activity concentration and linear attenuation are separate provides an opportunity for attenuation correction in PET that is not possible in SPECT, as we will see below.

Image Reconstruction For our development here, we focus on 2-D reconstruction. In practice, much PET data are acquired in 3-D mode (as noted above) because of its higher sensitivity. 3-D data can be reconstructed via 3-D approaches, but it is common to reorient the 3-D projection data into 2-D projection datasets (via a variety of techniques, including *Fourier rebinning*). In 2-D, coincidence counts can be organized into parallel-ray or fan-beam projections. Here, we consider only parallel-ray projections, as shown in Figure 9.10. From (9.20), we see that although we desire to reconstruct activity concentration $f(x,y)$, the attenuation $\mu(x,y)$ also appears in the PET imaging equation.

Let us start by ignoring attenuation, as we did in SPECT. Accordingly, we have an approximate imaging equation given by

$$\overline{n}(\ell,\theta) \approx K \int_{-R}^{R} f(x(s),y(s))\, ds\,, \quad (9.21)$$

where

$$K = \frac{\epsilon T A_h \tilde{A}_h}{4\pi R^2} \exp\left\{ - \int_{-R}^{R} \mu(x(s'),y(s');E)ds' \right\} . \quad (9.22)$$

So, if we define the scaled measurements

$$g(\ell,\theta) = \frac{\overline{n}(\ell,\theta)}{K}\,, \quad (9.23)$$

then we see that a PET scanner measures the Radon transform of the radiotracer activity concentration $f(x,y)$,

$$g(\ell,\theta) \approx \int_{-R}^{R} f(x(s),y(s))\, ds\,. \quad (9.24)$$

Therefore, we can use the reconstruction methods such as CBP from Chapter 6 to obtain an approximate reconstruction of $f(x,y)$.

Attenuation Correction Unlike SPECT, PET has a straightforward approach for handling attenuation. Suppose $\mu(x, y)$ is exactly known. Then, using (9.20), we can form attenuation-corrected measurements as

$$g_c(\ell, \theta) = \frac{\overline{n}(\ell, \theta)}{K} \exp\left\{ \int_{-R}^{R} \mu(x(s'), y(s'); 511\ \mathrm{keV})\, ds' \right\} . \tag{9.25}$$

We see by combining this equation with (9.20) that $g_c(\ell, \theta)$ is exactly the integral of activity along the line $L(\ell, \theta)$

$$g_c(\ell, \theta) = \int_{-R}^{R} f(x(s), y(s))\, ds . \tag{9.26}$$

Corrected measurements can now be used to obtain a more accurate image of activity concentration. Here is the CBP formula that could be used, for example:

$$\hat{f}(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} g_c(\ell, \theta) \tilde{c}(x \cos\theta + y \sin\theta - \ell)\, d\ell d\theta . \tag{9.27}$$

As in both CT and SPECT, $\tilde{c}(\ell)$ is an approximate ramp filter, defined using an appropriate window function [see (9.11)]. Although CBP ignores both the blurring caused by the physical size of the detectors (larger than x-ray CT) and the statistical noise made prominent by the relatively low photon counts, it remained the routine method in PET reconstruction for many years.

We see from (9.25) that it is only the line integrals of $\mu(x, y)$ that are required to correct the PET counts, not an actual image of $\mu(x, y)$. So actual reconstruction of $\mu(x, y)$ need not be carried out; instead, it is sufficient to simply measure the line integrals of attenuation on each LOR in order to be able to compute the corrected measurements $g_c(\ell, \theta)$. As noted above, it is now standard to acquire CT data along with the PET data, using a combined PET/CT system. The CT data are converted to $\mu(x, y)$ (at 511 keV) via a nonlinear transformation, from which the line integrals of $\mu(x, y)$ (at 511 keV) are obtained. We note that in PET/MRI systems, the MR data cannot be readily converted to $\mu(x, y)$. Instead, current approaches rely on preexisting databases that convert MR-identified anatomic structures to $\mu(x, y)$ (at 511 keV) via segmentation and atlas-based methods.

9.2.3 Iterative Reconstruction

CBP was for many years the most commonly used reconstruction algorithm in SPECT and PET, and it is still often used. It can be efficiently implemented and is relatively robust to those physical factors present in ECT imaging that corrupt the imaging process—for example, limited detector resolution, Compton scatter, attenuation (at least in SPECT), and noise. However, use of CBP is optimal (and, indeed, analytically correct) only if the imaging equation is an accurate depiction of reality, which we know from the preceding discussion for both PET and (especially) SPECT is not the case because of the corrupting physical processes. In both SPECT and PET, it is possible to model these physical processes in an attempt to reflect reality. Given these accurate, and elaborate, imaging models, a number of algorithms have been developed to produce better reconstructions than CBP. Broadly speaking, these algorithms fall into a class of algorithms called *iterative reconstruction algorithms*.

In an iterative reconstruction algorithm, an initial "guess" of the distribution of radioactivity is made, quite frequently by using CBP. Let us denote this initial guess by $f^{(0)}(x,y)$. If this were the true activity concentration in the body (or sample), then we should expect to see specific projection values as a result, say $g^0(\ell,\theta)$. These expected measurements can be computed given a description of the imaging geometry and instrumentation, for example, our imaging equation (9.20) with (9.23). Computing the expected measurements given a source object is called *solving the forward problem* or *forward projection*. This new "forward-estimated" sinogram can now be compared with the measured sinogram $g(\ell,\theta)$. If the acquisition and reconstruction process is correctly modeled, and if the initial guess were correct, then these sinograms would match. Otherwise, it is assumed that there are errors in the initial guess and that it should be modified, yielding a second guess $f^{(1)}(x,y)$. This concept of iteratively improved object estimates is the essence of iterative reconstruction.

There are many different iterative reconstruction algorithms that have been developed and put into practice over the years. They differ primarily in how they modify the estimated distribution in response to differences between the estimated and observed projection values. Regardless of the details, in the end each algorithm produces a sequence of estimated cross sections, $f^{(i)}(x,y)$, $i = 0,1,2,\ldots$, the final one of which (after some iteration stopping rule is satisfied) is the reconstruction that is used clinically. In essence, iterative reconstruction is based on successive estimates, with each successive estimate yielding a forward-estimated sinogram that is closer to the measured sinogram. We now present a very basic algorithm, called the *Algebraic Reconstruction Technique* (ART), which uses the Kaczmarz method for solving matrix equations, to illustrate this concept.

Algebraic Reconstruction Technique (ART) Suppose the activity concentration (object) is represented by pixel values f_j, $j = 1,\ldots,m$. Assuming that $p_j(x,y)$ $j = 1,\ldots,m$ are indicator functions describing each pixel location and shape, then the object can be represented by

$$f(x,y) = \sum_{j=1}^{m} f_j p_j(x,y)\,. \tag{9.28}$$

Now we ask what will be the observations given just the radioactivity in pixel j? To find the answer we only need to substitute $f_j p_j(x,y)$ as $f(x,y)$ into (9.6) and (9.8) for SPECT or (9.20) and (9.23) for PET to determine determine $g(\ell,\theta)$ for this source. For a particular line $L(\ell_i,\theta_i)$, the observation can be written $g_i = a_{ij} f_j$, where a_{ij} represents the response to a unit activity concentration located in pixel j. If there is activity in more than one pixel, then each pixel will contribute by superposition to the measurements, yielding the discrete observation equation

$$g_i = \sum_{j=1}^{m} a_{ij} f_j \, . \tag{9.29}$$

The general solution is easier to see if we represent each collection of values as vectors. Accordingly, we set $\mathbf{g} = (g_1, \ldots, g_n)$ and $\mathbf{f} = (f_1, \ldots, f_m)$ and $\mathbf{a}_i = (a_{i1}, \ldots, a_{im})$. Then

$$g_i = \mathbf{a}_i \cdot \mathbf{f} \, , \tag{9.30}$$

where · denotes inner product, and

$$\mathbf{g} = A\mathbf{f} \, , \tag{9.31}$$

where $A = [a_{ij}]$ is a matrix that concisely describes how to transform all the image pixel values into all the measurements. We note that (9.30) is a mathematical representation of the forward projection process mentioned above.

Given this notation and conceptual setup, we observe that there is a straightforward solution if the matrix A is invertible, that is,

$$\mathbf{f} = A^{-1}\mathbf{g} \, . \tag{9.32}$$

However, the number of measurements is generally larger than the number of unknowns and A cannot be assumed to be square. Given more measurements than unknowns, however, an estimate $\hat{\mathbf{f}}$ of the object can be found by the least squares approach in linear algebra, and can be written as

$$\hat{\mathbf{f}} = (A^T A)^{-1} A^T \mathbf{g} \, , \tag{9.33}$$

where the T superscript denotes matrix transpose. This approach is known as the pseudoinverse estimate of $\mathbf{f}$, and the equation is known as the normal equation. It is a practical solution only on very small problems; otherwise, there is a significant burden in computing the required matrix inverse. Here is where the iterative solution comes in handy.

The Kaczmarz method solves (9.31) using the following iteration

$$\hat{\mathbf{f}}^{(k)} = \hat{\mathbf{f}}^{(k-1)} - \frac{(\mathbf{a}_i \cdot \hat{\mathbf{f}}^{(k-1)} - g_i)}{\mathbf{a}_i \cdot \mathbf{a}_i} \mathbf{a}_i \, , \tag{9.34}$$

which is carried out for each observation g_i and then repeated until convergence.[1] This iteration uses one measurement at a time to update $\hat{\mathbf{f}}$ and does not require a matrix inverse; hence, it is computationally tractable. Notice that the quantity in the parenthesis in (9.34) includes an implementation of the forward projection process (the inner product) applied to the current image estimate, which is

---

[1] *Convergence* in iterative reconstruction is the state that is attained when further iterations do not materially improve the reconstructed image.

then compared by subtraction to the actual observation. Thus, if the computed measurements agree with the measurements for all measurements, there will be no change in the object and the solution has been found. In general, a more modest criterion, such as an iteration producing only a very small change to $\hat{\mathbf{f}}$, is a more reasonable stopping criterion.

Maximum Likelihood by Expectation Maximization (ML-EM) The ART method captures the underlying concept of iterative reconstruction in mathematics quite elegantly. However, it is not used in practice, primarily because it does not consider the random factors that are inherent in both SPECT and PET systems. We see how the matrix A can, in principle, incorporate many deterministic factors such as detector size, shape, and efficiency, and object attenuation, but it is not immediately clear how one would incorporate the Poisson nature of photon counting and such factors as scattering and random coincidences. The *maximum likelihood by expectation maximization* (ML-EM) methods address this issue; they have become the most popular class of reconstruction algorithms in clinical use in emission tomography because they are capable of incorporating all of these deterministic and probabilistic factors.

The ML-EM method starts with an imaging equation that directly expresses the number of photon counts that are expected in a detector. For example, (9.6) and (9.20) will work for SPECT and PET, respectively. We assume the same pixel basis as in (9.28) and finite number of observations as in ART above, and derive from the geometry the coefficients a_{ij} that will transform the pixel having activity concentration f_j in it to the measurement g_i. We can also add a factor to be determined from experiment (and thereafter assumed to be known) reflecting the extra counts $\overline{r}_i$ that are expected to arise in detector i due to scattering and random coincidences, leading to the discrete imaging equation

$$\overline{n}_i(\mathbf{f}) = \sum_{j=1}^{m} a_{ij} f_j + \overline{r}_i \,. \tag{9.35}$$

In Sections 9.3.2 and 9.3.3, we discuss the estimation of the extra counts $\overline{r}_i$.

We know that the quantity $\overline{n}_i(\mathbf{f})$ represents the mean number of counts, whereas the actual observed count is the realization of a Poisson random variable with this mean. The form of the Poisson probability mass function can now be used to set up a likelihood function, which assumes that the actual observations n_i, $i = 1, \ldots, n$ are known—that is, they have been observed during image acquisition—and the image values are the unknowns. The maximum likelihood (ML) estimate is defined as the collection of image pixel values $\mathbf{f}$ (or f_j, $j = 1, \ldots, m$) that together maximize the likelihood function (or the logarithm of the likelihood function). Although various algorithms have been defined to try to find the ML solution, the most popular and computationally convenient method is the ML-EM method, which is inherently iterative.

The ML-EM algorithm seeks to find the image pixel values $\mathbf{f}$ that maximize the logarithm of the Poisson likelihood objective function, which is given by

$$L(\mathbf{f}) = \sum_{i=1}^{m} n_i \ln \overline{n}_i(\mathbf{f}) - \overline{n}_i(\mathbf{f}) \,. \tag{9.36}$$

It can be shown that the following ML-EM iteration

$$f_j^{(k+1)} = \frac{f_j^{(k)}}{a_j} \sum_{i=1}^{n} a_{ji} \frac{n_i}{\overline{n}_i(\mathbf{f}^{(k)})} \tag{9.37}$$

converges to the ML solution. Here, $\mathbf{f}^{(k)}$ is the k-th estimate of the image, and

$$a_j = \sum_{i=1}^{n} a_{ij}\,, \quad j = 1, \ldots, m \tag{9.38}$$

are the *sensitivity factors* that characterize the sensitivity of each voxel to the collection of measurements. (If $a_j = 0$ for some j, then the geometry of the system cannot estimate f_j, and this pixel should be excluded from the model.) The ML-EM algorithm (9.37) makes use of the forward model through $\overline{n}_i(\mathbf{f}^{(k)})$, although the comparison with the observations is done using a ratio rather than a difference. We see that if $n_i = \overline{n}_i(\mathbf{f}^{(k)})$ for all i, then the image estimate remains unchanged because of the definition of the sensitivity factors in (9.38).

In the past, ML-EM approaches suffered from lengthy reconstruction times and difficulty in converging to a useful solution, particularly in the presence of significant noise. (Difficulties in convergence lead to increased numbers of iterations, increasing overall reconstruction time.) Hudson and Larkin made a major breakthrough when they developed an algorithm—*ordered subsets expectation maximization* (OSEM)—that utilizes "ordered subsets" to accelerate iterative reconstruction. In this approach, observations are grouped into an ordered sequence of subsets or blocks. For example, if a full set of projections consists of 360 views and there are 15 subsets, each subset would consist of 24 views. Of major importance, these subsets represent mutually exclusive and collectively exhaustive use of the projection data, and they are treated independently in parallel. One iteration of the algorithm is a single pass through all the subsets. However, the reconstructed image is updated after every subset, which accelerates convergence.

Ordered subsets are an extension of iterative reconstruction. In Hudson and Larkin's original implementation, they used Shepp and Vardi's expectation maximization approach. Like any standard EM algorithm, this consists of a backprojection–projection–backprojection sequence. At each iteration, the reconstruction from the previous iteration is the starting point. In OSEM, the standard EM algorithm is applied to each of the subsets in turn. Such an approach can be applied to both SPECT and PET imaging. In SPECT, the subsets may correspond to natural groupings of projections—for example, each head in a triple-head system. In PET, projection data may be reorganized after acquisition to define blocks or subsets in a similar fashion. OSEM has become the preferred approach to reconstruction in emission tomography (including both SPECT and PET).

9.3 Image Quality in SPECT and PET

From a physics point of view, there are five major factors that affect image quality and the quantification of absolute radioactivity in emission tomography. These include finite spatial resolution (e.g., the in-plane and axial spatial resolution of a SPECT or PET scanner); attenuation of photons by tissue; detection of scattered photons; accidental coincidence counting of *random* (unpaired) photons in coincidence, applicable only to PET imaging; and ''noise'' resulting from the statistical nature of radioactive decay.

9.3.1 Spatial Resolution

As first introduced in Chapter 3, all imaging systems have a limited ability to resolve small objects. The spatial resolution of an emission tomography system can be thought of as that distance by which two small point sources of radioactivity must be separated to be distinguished as separate in the image (see again Figure 3.6). The spatial resolution of a SPECT system is mainly determined by the resolution in the projection data, which itself is determined by the factors discussed in Section 8.4.1, and the effects of reconstruction, especially the specific filter used, since the approximate ramp filter given in (9.11) usually includes a low-pass filter $W(\varrho)$ to reduce noise. The spatial resolution of a PET system is mainly determined by the physical cross-section dimensions of each detector element and the effects of reconstruction. In addition, because PET uses the annihilation site, rather than the emission site of the positron per se, *positron range* (i.e., the distance the positron travels before annihilation) also limits resolution. In particular, positron range depends on the initial energy of the positron. Positrons are emitted with a range of energies, leading to a distribution of positron ranges. The effective spatial resolution of a PET scanner can be modeled as the convolution of the scanner's intrinsic point spread function (PSF) with the positron range distribution.

In practice, we can model SPECT and PET resolution as a cascade of LSI systems [see (2.46)]. The development is identical to that for CT resolution, as fully derived in Section 6.4.1. The final expression for the relation between the reconstructed (but blurred) image and the object, identical to (6.56), is

$$\hat{f}(x, y) = f(x, y) * \mathrm{h}(r), \tag{9.39}$$

where $r = \sqrt{x^2 + y^2}$.

Here (as in CT), $\mathrm{h}(r)$ represents the (cascaded) PSF and $\mathrm{h}(r)$ for SPECT includes (1) the (1-D, in-plane) projection PSF [i.e., (8.14), with y set to 0, and $x = \ell$] and (2) w [identical to w in (6.48)]; $\mathrm{h}(r)$ for PET includes (1) a *positron range function*, (2) a rect function for detector width [identical to s in (6.48)], and (3) w.

Finite spatial resolution in SPECT and PET results in two important, related effects. First, the image is blurred, with the degree of blurring dependent on the spatial resolution. This blurring prevents the delineation of edges of larger structures and may not allow the visualization of smaller ones as distinct objects. Second, neighboring areas are smeared and averaged together, reducing the measured value in the areas with greater radioactivity, and increasing it in the areas with lesser radioactivity (i.e., reducing contrast). Said another way, referring back to (3.1), the reduction in $f_{\max}$ and the increase in $f_{\min}$ reduces modulation (or contrast). In this fashion, finite spatial resolution produces an underestimation of radioactivity in small structures with greater radioactivity, with progressive underestimation as the structures get smaller. The effect is not eliminated until the object size is approximately three times the resolution of the imaging system. These effects, which also apply to the axial resolution (sometimes mistakenly called "slice thickness") of a tomographic scanner, are referred to as *partial volume effects*.

9.3.2 Attenuation and Scatter

The photons detected in nuclear medicine imaging are electromagnetic radiation. As such, they undergo two major types of interactions in tissue—photoelectric effect and Compton scattering—as discussed in Chapter 4. The photoelectric effect results in complete absorption of the photon, reducing observed count rate, while scattered photons are still detected. Since attenuation is a combination of absorption (by the photoelectric effect) and scatter (by Compton scattering), correction schemes can either treat these processes independently or jointly.

As first discussed in Section 5.4.3 in the context of projection radiography, large-angle scatter produces a low-level background "haze" in the image, which reduces contrast [see (5.41)]. This reduction in contrast arises from the addition of this scatter background, rather than a reduction in $f_{\max}$ or an increase in $f_{\min}$ per se, as discussed above in the context of resolution effects and in (3.1). Event C in Figure 9.11 depicts a scatter coincidence in PET.

Absorption produces a gradual, progressive underestimation of radioactivity from the edge to the center of the body, by about a factor of five to as much as 50, depending on body size, photon energy, and SPECT versus PET detection (since comparison of (9.6) and (9.20) indicates that, for SPECT, the photon only has to travel through part of the body, whereas for PET, the sum of the two annihilation photons' travel distance is always the complete body thickness). The same physics and mathematics that make "perfect" attenuation correction at least theoretically achievable in PET, but not in SPECT, make the effects of attenuation itself more dramatic in PET than in SPECT. This is counterintuitive given the difference in photon energies of the commonly used radionuclides for PET and SPECT. In PET, the main photon energy and linear attenuation coefficient (for water) are 511 keV and $0.095\ \text{cm}^{-1}$, respectively, while for SPECT they are 140 keV and $0.145\ \text{cm}^{-1}$. (Note that the difference in energy is a factor of 3.7, but the difference in attenuation coefficients is only about 50 percent.) The underlying basis for higher attenuation in PET is the fact that

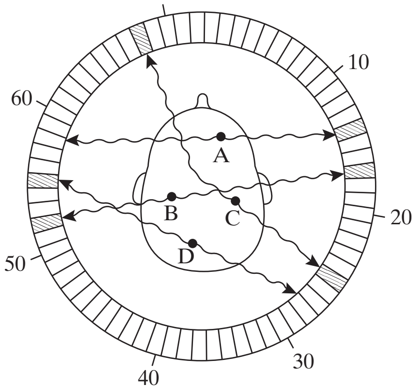


Figure 9.11 Coincidence counting in PET. B represents a true coincidence, A and D random or accidental coincidences, and C scattered coincidence.

both photons must escape the patient and be detected in coincidence, meaning that the total attenuation path length is longer for coincidence imaging than for SPECT. This distinction increases exponentially with patient size. For example, with PET, a 30 cm diameter patient requires an average attenuation correction factor of 16, whereas a 40 cm diameter patient requires a correction factor of 50. For comparison, the attenuation correction factor in cardiac SPECT imaging, where the photons pass through 20 cm of tissue, is just 20.

As discussed above and formalized in (9.25) and (9.27), in PET an exact correction for attenuation can be made, while such is not the case in SPECT. In both SPECT and PET, scatter is usually estimated at the projection level by use of a model of the scattering process that incorporates characteristics of the scanner and the patient. This estimate is used to correct for scatter in the reconstruction process.

9.3.3 Random Coincidences

In PET, random or accidental coincidences are possible. These occur when one annihilation photon from one electron-positron annihilation event, and another annihilation photon from a different electron-positron annihilation event are detected in coincidence. Events A and D in Figure 9.11 depict random or accidental coincidences.

The random coincidence count rate is given by the product of the coincidence time window and the "singles" (individual) count rates from the two opposing detectors in coincidence,

$$R = 2\tau S^{+} S^{-}\,, \tag{9.40}$$

where 2τ is the coincidence time window width and S^{+} and S^{-} are the individual count rates of opposing detectors.

As the activity in the field of view increases, the true coincidence rate increases linearly (ignoring dead-time effects), but the random coincidence rate increases as the square of the increase in activity. Sensitivity may be increased by increasing the width of the coincidence time window (up to the point when the full coincidence time spectrum has been included), but only at the expense of acquisition of more random coincidences. A great challenge in coincidence imaging is to balance sensitivity (true coincidences) with random coincidences in optimizing the time window setting. Some manufacturers specify the window in terms of the time resolution of the system; others do so in terms of twice the time resolution, which is usually the actual window width, and therefore the more traditional specification. The timing window, which typically ranges between 2 and 20 ns, is primarily determined by the "speed" of the detector material.

In practice, random or accidental coincidences are estimated from the observed projection data through use of (9.40) and then incorporated in the reconstruction process. As an alternative, an offset coincidence time window (in which the window is intentionally offset by a large time lag, such that any and all observed coincidences must be random) can be used to directly measure the random coincidence rate.

9.3.4 Contrast

In the process of considering the effects of finite spatial resolution and scatter, we noted their influence on image contrast in SPECT and PET. In both cases, our consideration of contrast followed that first introduced in (3.1), and recast specifically for planar imaging in (8.25). In SPECT and PET, the definition of contrast remains the same, but $f_{\max}$ and $f_{\min}$ are from reconstructed images. The definition of local contrast also remains the same, with the values of foreground intensity and background intensity taken from the reconstructed images at the appropriate locations.

9.3.5 Noise and Signal-to-Noise Ratio

It is important to remember that deterministic effects influence only the signal component of an image's signal-to-noise ratio; in other words, they influence only quantitative *accuracy*. Noise arises from the random statistical nature of radioactive decay, as well as potentially from the imaging hardware and certain image processing operations, and influences the noise component of the SNR; in other words, it influences *precision*. The noise in projection data in SPECT and PET is identical to the situation in planar scintigraphy and follows the development given in Section 8.4.5. However, the reconstruction process changes the noise magnitude and correlation in a fashion similar to the way that reconstruction changes the noise in CT compared with projection radiography.

In practice, estimating the noise in a reconstructed SPECT or PET image is difficult. In order to present a generic formulation, we make the same assumptions we made to derive an expression for variance in CT: The measured line integrals are statistically independent, and the object is sufficiently uniform that the line integrals are essentially equal. With these simplifying assumptions, the derivation of the variance in a reconstructed SPECT or PET image is identical to that presented in Section 6.4.2, is independent of (x, y) location, and is given by (6.72).

In PET, a powerful concept called *noise equivalent count rate* (NEC) has been developed to reflect these concepts and include PET-specific issues (such as random coincidences). The NEC is given by:

$$\mathrm{NEC} = \frac{T^2}{T + S + 2R}, \tag{9.41}$$

where T is the true coincidence count rate, S is the scatter count rate, and R is the random coincidence count rate. NEC can be thought of as the equivalent true count rate of an ideal system (i.e., one with no scatter or random coincidences) that results in the same SNR as an actual system with given scatter and random coincidence count rates.

Ultimately, key performance parameters such as lesion detectability or quantitative accuracy and precision depend on the signal-to-noise ratio and total *error*. Since both contrast and noise are defined the same way in CT and SPECT/PET, the SNR formulation presented for CT in (6.73) applies, and the final SNR formulation is the same as (6.74).

9.4 Summary and Key Concepts

Cross-sectional—i.e., tomographic—imaging in nuclear medicine is accomplished with SPECT and PET. SPECT usually utilizes a rotating Anger camera or cameras; PET utilizes a dedicated imaging system containing a ring or multiple rings of detectors. In this chapter, we presented the following key concepts that you should now understand:

1. *Emission computed tomography* includes *single photon emission computed tomography* (SPECT) and *positron emission tomography* (PET).
2. SPECT is based on an ensemble of projection images, each of which is usually a conventional planar scintigram.
3. PET has no projection analog and is based on *coincidence detection of paired gamma rays* (“annihilation photons”) following positron-electron annihilation.
4. SPECT is typically performed with rotating Anger scintillation cameras (with multicamera systems being the most common); specialized geometry systems now exist for enhanced sensitivity, especially for cardiac imaging.
5. PET is always performed with a dedicated PET scanner; such scanners are always multislice and typically operate in 3-D mode.
6. Multimodality SPECT and PET systems are common; these systems incorporate CT (or MRI) to facilitate anatomic localization and attenuation correction.
7. The *basic SPECT imaging equation* is, in essence, identical to that in planar scintigraphy, except for a change in notation that facilitates consideration of reconstruction; reconstruction can follow the approach outlined for CT.
8. The activity and attenuation terms in the *basic SPECT imaging equation* are not separable; thus, there is no closed-form analytic solution for attenuation correction in SPECT.
9. The *basic PET imaging equation* is similar to that for SPECT, with one important difference: the use of coincidence detection leads to an attenuation term whose limits of integration span the entire body; thus, PET has a closed-form solution for attenuation correction.
10. *Iterative reconstruction* is a newer, more computer-intensive approach that implicitly takes the random nature of decay into account and can incorporate models of attenuation, scatter, and blur.
11. *Image quality* in SPECT and PET is limited by *resolution*, *scatter*, and *noise*.

Further Reading

Bendriem, B. and Townsend, D.W. *The Theory and Practice of 3D PET*, Dordrecht, Netherlands: Kluwer Academic Publishers, 1998.

Cherry, S.R., Sorenson, J.A., and Phelps, M.E. *Physics in Nuclear Medicine*, 4th ed. Philadelphia, PA: W. B. Saunders, 2012.

Christian, P.E. and Waterstram-Rich, K.M. *Nuclear Medicine and PET/CT: Technology and Techniques*, 7th ed., New York, NY: Elsevier/Mosby, 2012.

Hudson, H.M. and Larkin, R.S. “Accelerated Image Reconstruction Using Ordered Subsets of Projection Data,” *IEEE Transactions on Medical Imaging*, 13(4):601–609, 1994.

Shepp, L.A. and Vardi, Y. “Maximum Likelihood Reconstruction for Emission Tomography,” *IEEE Transactions on Medical Imaging*, MI-1(2):113–122, 1982.

Problems

SPECT

9.1 Consider a 2-D cross section consisting of three separate compartments, R1, R2, and R3, as shown in Figure P9.1.

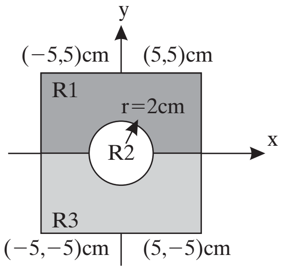


Figure P9.1 Geometry of a radioactive cross section. See Problem 9.1.

(**a**) Suppose a solution containing a 511 keV gamma ray emitting radionuclide with concentration 0.5 mCi/cm^3 fills R2. The object is imaged using a (2-D) SPECT scanner. Find expressions for $g_{\text{SPECT}}(\ell, 0^\circ)$ and $g_{\text{SPECT}}(\ell, 180^\circ)$. (Ignore the inverse square law.)

(**b**) Suppose that the radionuclide in part (a) is replaced with a positron-emitting radionuclide with the same concentration and the object is imaged using a (2-D) PET scanner. Find expressions for $g_{\text{PET}}(\ell, 0^\circ)$ and $g_{\text{PET}}(\ell, 180^\circ)$.

For parts (c) and (d), let $\mu_1 = 0.3\ \text{cm}^{-1}$, $\mu_2 = 0.4\ \text{cm}^{-1}$, and $\mu_3 = 0.2\ \text{cm}^{-1}$.

(**c**) Calculate $g_{\text{SPECT}}(0, 0^\circ)$ and $g_{\text{SPECT}}(0, 180^\circ)$.

(**d**) Calculate $g_{\text{PET}}(0, 0^\circ)$ and $g_{\text{PET}}(0, 180^\circ)$.

9.2 Consider a slice of an object, as shown in Figure P9.2, with two compartments filled with radioactive liquids emitting gamma photons with energy 150 keV.

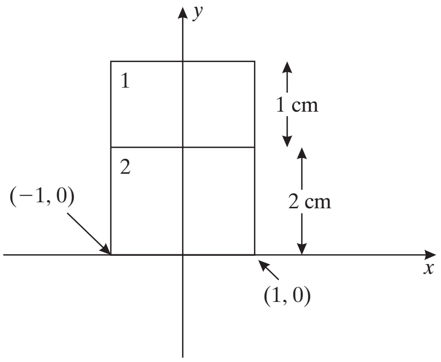


Figure P9.2 Two-compartment object. See Problem 9.2.

The linear attenuation coefficients of the liquids are given by $\mu_1(150\ \text{keV}) = 0.2\ \text{cm}^{-1}$, $\mu_2(150\ \text{keV}) = 0.4\ \text{cm}^{-1}$, $\mu_1(511\ \text{keV}) = 0.1\ \text{cm}^{-1}$, and $\mu_2(511\ \text{keV}) = 0.3\ \text{cm}^{-1}$. The concentrations f (dps per unit volume) of the liquids are given by $f_1 = 0.2\ \text{mCi/cm}^3$ and $f_2 = 0.4\ \text{mCi/cm}^3$.

The object is imaged using a SPECT scanner. Assume that there are ideal collimators having 100% efficiency.

(**a**) Find the projections $g(\ell, 180^\circ)$ and $g(\ell, 90^\circ)$. Neglect inverse square law effects.

(**b**) The liquids are replaced by two positron emitting liquids of same concentrations. If the system is now imaged by a PET scanner, find $g(\ell, 0^\circ)$, assuming collimators have 100% efficiency.

(**c**) Describe how to compensate for attenuation in PET reconstruction. Is the same approach applicable in SPECT? Explain with imaging equations.

(**d**) Explain how the answers in part (a) and (b) would change if real collimators are used instead of ideal collimators.

9.3 Suppose there are two kinds of radiotracers, P and Q. At time 0, there are $N_0 = 1 \times 10^{15}$ atoms of each of them.

(**a**) After 1 hour, there are $\frac{1}{3}N_0$ of P and $\frac{2}{3}N_0$ of Q remaining. What are the half-lives of P and Q?

(**b**) What is the radioactivity A_P and A_Q at time 1 hour in units of curies?

A square phantom with width 6 cm has a circle with radius 1 cm inside, as illustrated in Figure P9.3. The circle is filled with radiotracer Q, with the amount of radioactivity A_Q in (b) everywhere, and the square excluding the circle is filled with radiotracer P, with the amount of radioactivity A_P in (b) everywhere. The phantom is put into a SPECT system.

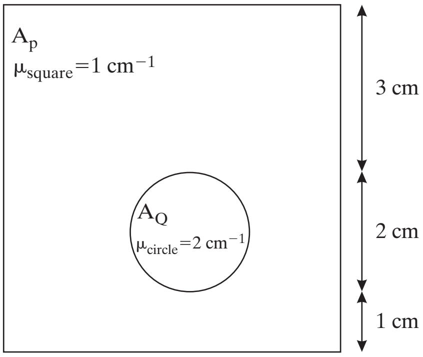


Figure P9.3
A square phantom. See Problem 9.3.

(**c**) What is the local contrast of the projection $g(\ell, 180^\circ)$ if no attenuation is present? (Let the peak value in the circle be used as the intensity of the object, which applies to (d) as well.)

(**d**) What is the answer to (c) if attenuation is present? Let the linear attenuation of the square excluding the circle be $\mu_{\text{square}} = 1\ \text{cm}^{-1}$ and that of the circle be $\mu_{\text{circle}} = 2\ \text{cm}^{-1}$.

(**e**) Take attenuation into account and estimate whether the absolute value of the local contrast ($|C|$) of the projection $g(\ell, 180^\circ)$ would be bigger or smaller than that of $g(\ell, 0^\circ)$. Explain.

9.4 Consider a 2-D object consisting of two triangle compartments, as shown in Figure P9.4. Suppose a solution containing a 511 KeV gamma ray emitting radionuclide with concentration $f = 0.5$ mCi/cm^3 fills the lower triangle. The linear attenuation coefficients in the two regions are $\mu_1 = 0.1$ cm^{-1} and $\mu_2 = 0.2$ cm^{-1}. Assume perfect detection in all cases and ignore the inverse square law effect.

(**a**) We image the radioactivity using a 2D SPECT scanner. Compute the projected radioactivities $g_{SPECT}(\ell, 0°)$ and $g_{SPECT}(\ell, 180°)$. The camera is located on the $+y$-axis looking down when $\theta = 0°$.

(**b**) Now assume the radionuclide in part (a) is replaced by a positron emitting radionuclide with the same concentration. Assume the linear attenuation coefficients in the two regions are the same as in part (a). If using a 2D PET scanner, compute $g_{PET}(\ell, 0°)$ and $g_{PET}(\ell, 180°)$.

(**c**) Explain why attenuation is not a big problem in PET.

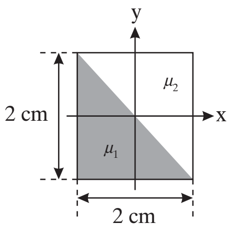


Figure P9.4
An object with triangular compartments. See Problem 9.4.

9.5 Ignoring the inverse square law and attenuation, an approximate reconstruction for SPECT imaging is given by

$$\hat{f}(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} g(\ell, \theta)\tilde{c}(x\cos\theta + y\sin\theta - \ell)d\ell\, d\theta\,,$$

where $\tilde{c}(\ell) = \mathcal{F}_{1D}^{-1}\{|\varrho|W(\varrho)\}$ and $W(\varrho)$ is a rectangular windowing filter that cuts off at $\varrho = \varrho_0$. Suppose we use M projections $(\theta_1, \theta_2, ..., \theta_M)$ uniformly spaced over the range $[0, \pi)$, and $N + 1$ (odd) ray paths per projection. Assume the spacing between detectors is T, and $g_{ij} = g_{\theta_j}(iT)$. A discrete approximation to the reconstruction of $\hat{f}(x, y)$ can be written as

$$\hat{f}(x, y) = \frac{\pi T}{M} \sum_{j=1}^{M} \sum_{i=-N/2}^{N/2} g_{ij}\tilde{c}(x\cos\theta_j + y\sin\theta_j - iT)\,.$$

(**a**) The observation g_{ij} is proportional to the number of photons hitting the detector i at angle θ_j, N_{ij}, that is, $g_{ij} = kN_{ij}$. Assume N_{ij} is a Poisson random variable with mean $\overline{N}_{ij}$ and is independent for different i and j. Give the mean and the variance of the reconstructed image, mean$[\hat{f}(x, y)]$ and var$[\hat{f}(x, y)]$.

(**b**) Show that

$$\frac{\pi}{M} \sum_{j=1}^{M} T \sum_{i=-N/2}^{N/2} [\tilde{c}(x\cos\theta_j + y\mathrm{sin}_\mathrm{j}\theta - iT)]^2$$

can be approximated as $\frac{2\pi\varrho_0^3}{3}$.

Now, like in CT, we assume that $\overline{N}_{ij} \approx \overline{N}$.

(**c**) Find $\mathrm{var}[\hat{f}(x, y)]$ using the result in (b).

(**d**) Define $\mathrm{SNR} = \frac{\mathrm{mean}[\hat{f}(x,y)]}{\sqrt{\mathrm{var}[\hat{f}(x,y)]}}$. Assume that we double the photon counts, that is, $\overline{N}' = 2\overline{N}$. Before doubling, the SNR is SNR_1; and after doubling, the SNR is SNR_2. What is the ratio $\frac{\mathrm{SNR}_2}{\mathrm{SNR}_1}$?

PET

9.6 Suppose a PET detector comprises four square PMTs (arranged as a 2 by 2 matrix) and a single BGO crystal with slits made in such a way that it is divided into an 8 by 8 matrix of individual detectors. Assume that the PMTs and the detectors cover the exact same square area and that each PMT is 2 in by 2 in in size. This geometry is shown in Figure P9.5. The response of a PMT to an event occurring in a particular subcrystal depends on the distance from the center of the PMT to the center of the subcrystal, r, as follows

$$\text{PMT Response} = \exp[-r/\tau],$$

where $\tau = 1$ inch.

(**a**) Find a general expression for the response in $\mathrm{PMT}(i, j)$ to an event in crystal $C(k, l)$.

(**b**) Find the numerical responses in each to an event in crystal $C(4, 6)$.

(**c**) Ignoring the possibility of noise, develop a scheme to uniquely identify the crystal in which an event occurred.

(**d**) Characterize a worst-case scenario in which the smallest possible additive noise in one PMT's signal causes an error in event localization.

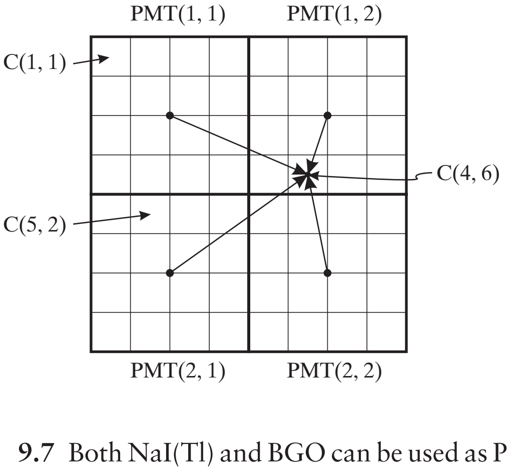


Figure P9.5
Detector geometry for Problem 9.6.

9.7 Both NaI(Tl) and BGO can be used as PET detectors. The linear attenuation coefficient of BGO at 511 keV is 0.964 cm^{-1} as opposed to 0.343 cm^{-1}

for NaI(Tl). BGO does not convert gamma ray photons to light photons as efficiently as NaI(Tl), and is in fact roughly 13% as efficient.

In this problem, suppose two detectors are designed, one from NaI(Tl) and one from BGO. Assume both are designed to stop 75% of the 511 keV photons that strike the crystal.

(**a**) Find the detector thicknesses for NaI(Tl) and BGO.

(**b**) A light burst from NaI(Tl) has a higher intrinsic SNR than that from BGO. Find the ratio of intrinsic SNRs, NaI(Tl) to BGO.

9.8 You decide to build a PET scanner from a dual-head SPECT camera you own. It has two 30-cm square Anger cameras mounted facing each other 1.5 m apart.

(**a**) Would you use the low-energy collimators, the high-energy collimators, or no collimators? Explain.

(**b**) What significant piece of electronics would you have to add to the SPECT circuitry to make this work as a PET scanner?

(**c**) Explain how you would use the X and Y signals coming out of the opposing Anger cameras. Would you use the Z-pulses? Why or why not?

(**d**) You will have to rotate the Anger cameras at some point during the scan. What is the minimum number of angular positions of the two cameras that would be required in order to get full angular coverage for PET reconstruction?

(**e**) Consider the radiotracer concentration in the body. Would you expect to require a higher, lower, or about the same dose in this make-shift PET scanner versus a real PET scanner? Explain your rationale.

9.9 Consider the two-dimensional cross section shown in Figure P9.6 consisting of three separate compartments R_1, R_2, and R_3.

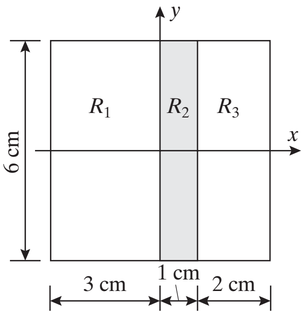


Figure P9.6
Object geometry for Problem 9.9.

(**a**) Suppose a solution containing a 511 keV gamma ray emitting radionuclide with concentration 0.3 mCi/cm^3 fills only R_2; R_1 and R_3 contain nonradioactive solutions. Let the linear attenuation coefficients (at 511 keV) in the three regions be $\mu_1 = 0.2\ \text{cm}^{-1}$, $\mu_2 = 0.3\ \text{cm}^{-1}$, and $\mu_3 = 0.1\ \text{cm}^{-1}$, respectively. Suppose we image the radioactivity using a (2-D) SPECT scanner outside the object. Compute the projected radioactivities $g_{\mathrm{SPECT}}(\ell, 90^\circ)$ and $g_{\mathrm{SPECT}}(\ell, 270^\circ)$.

(b) Now assume the radionuclide in (a) is replaced by a positron emitting radionuclide with the same concentration. Assume the linear attenuation coefficients in the three regions are the same. This time the body is imaged using a (2-D) PET scanner. Compute $g_{\mathrm{PET}}(\ell, 90^\circ)$ and $g_{\mathrm{PET}}(\ell, 270^\circ)$.

(c) Explain how to compensate for attenuation in a PET scanner in order to reconstruct an accurate image of the radionuclide concentration. Can you do the same for a SPECT scanner?

9.10 Suppose a PET scanner has 1,000 detectors packed tightly around a circle of diameter of 1.5 m.

(a) What is the approximate size (width) of each detector? Explain the tradeoff between using deep (long) and shallow (short) detectors.

(b) What is the purpose of coincidence detection in PET? What is a nominal time interval defining a coincidence "event window" for PET? Explain why is it undesirable to make the event window (1) smaller or (2) longer.

(c) Assume there is no "wobbling" or "dichotomic" motion of the PET gantry. What is the line integral sampling interval (which has the symbol T in CT) in the center of the scanner? Assuming the usual sampling "rule of thumb" from CT, how many pixels would be in a typical PET image. Explain why motion of the PET gantry is desirable.

(d) Explain why resolution of the PET scanner is typically worse away from the center of the scanner.

9.11 A PET-CT system is shown in Figure P9.7. There are N_0 positron annihilations at the origin of the circle whose gamma rays travel on the line $L(0, 0^\circ)$. N^+ of these positron annihilations reach detector A and N^- reach detector B.

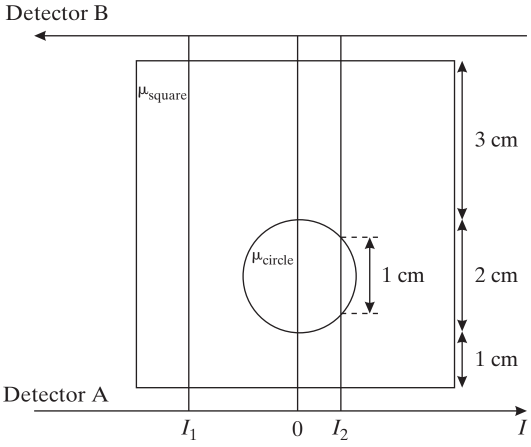


Figure P9.7
See Problem 9.11.

(**a**) Write an equation for the number of coincidence events N_c arising from positron annihilations at the center of the circle that will be detected as a function of N_0, N^+, and N^-.

For parts (b)–(d): Suppose that the radioactivity is the same everywhere and the value of $g(\ell_1, 0^\circ)$ is 2/3 of the value of $g(0, 0^\circ)$.

(**b**) Write an equation representing the relationship between μ_{square} and μ_{circle}.

(**c**) What is the ratio $g(\ell_2, 0^\circ)/g(0, 0^\circ)$?

(**d**) What is the local contrast on the 0° projection? (Take the peak of the circle as the object intensity.)

This page intentionally left blank

Ultrasound Imaging

PART IV

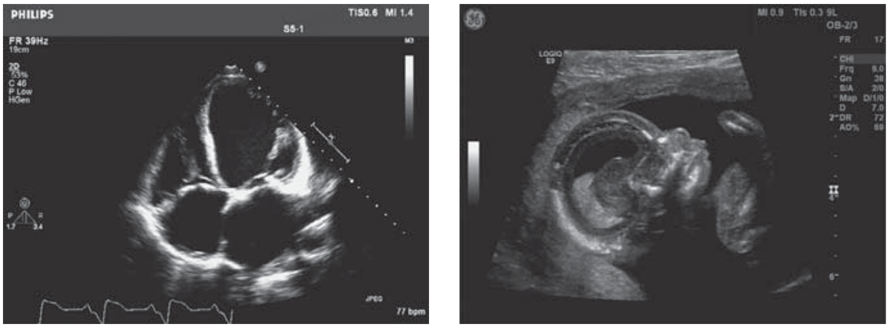
Overview

In the preceding two parts of the book, we considered those imaging modalities that rely on ionizing radiation to create the signals. We saw that these modalities produce images of structure and function through a variety of mechanisms, all based on ionizing radiation. The modalities are of high clinical utility—that is why they are so commonly used—but the use of ionizing radiation carries some risk. At the low doses used in medical imaging, the main biological risk is cancer; the probability of inducing cancer from a diagnostic imaging examination is very, very low, but it is not zero. Sometimes this is of concern, such as when imaging *in utero*.

Here, we examine a modality considered to be fully noninvasive and risk free: ultrasound imaging. In projection radiography and computed tomography, we were interested in the *transmission* of a beam of radiation through the patient; the medical signal of interest was produced by the variable attenuation of this beam by body tissues. In contrast, with ultrasound, we are interested in the *reflection* of a beam of sound by internal structures within the patient. The medical signal of interest here is produced by the variable reflectance of this sound beam by different body tissues.

A basic ultrasound imaging system consists of a transducer and associated electronics, including a display. The transducer itself is both a transmitter and receiver of ultrasound energy. In practice, the electronics ‘‘steer’’ the beam so that (typically) an arc is covered within the patient, producing a 2-D image on the display. Many modern systems have 3-D capability whereby they produce successive 2-D arcs on different planes which, when pieced together, form a 3-D volumetric image. Therefore, in its most basic form, ultrasound imaging systems can be used to directly visualize the shape and size of various anatomical structures of interest.

The most common ultrasound procedures involve imaging the developing fetus *in utero* and *echocardiography* for heart imaging. A representative echocardiogram is shown in Figure IV.1(a). Here, the transducer is positioned at the tip



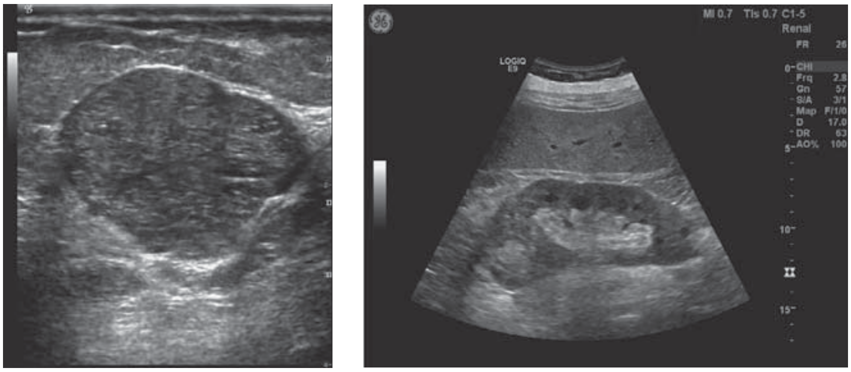
(a) (b)

Figure IV.1
Two common and easily recognizable clinical ultrasound examinations. (a) Echocardiogram, showing the four chambers of the heart. (b) Fetal ultrasound, showing a normal fetus at the third trimester of gestation. Image (a) courtesy of Philips Healthcare. Image (b) courtesy of GE Healthcare.

or *apex* of the heart, pointing in such a way as to cut through all four heart chambers. Cross-sectional images of the beating heart provide the physician with a lot of diagnostic and functional information about the health of the heart as a pump. The fact that the method is noninvasive and readily performed in a doctor's office helps explain the popularity of 2-D echocardiography. Since the transducer is held by hand during most ultrasound procedures, the demand for well-trained, experienced ultrasonographers is quite high, as the details of acquisition, including the view angle, significantly influences both the field of view and the image quality.

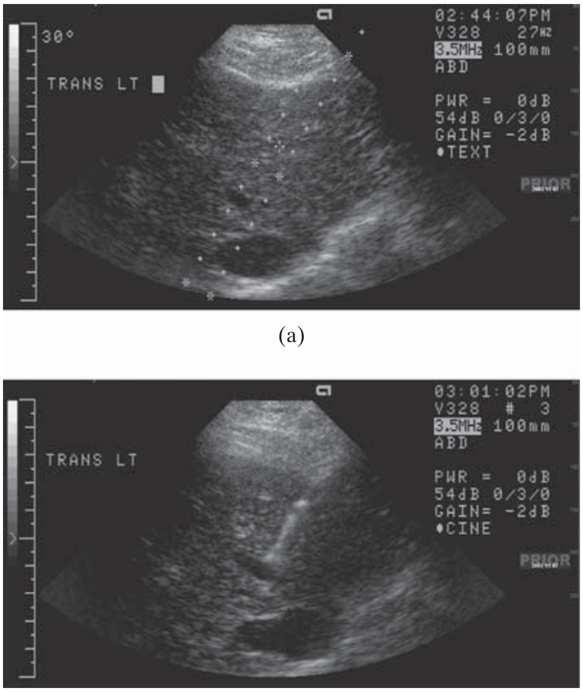
Because it is completely safe for both mother and baby, ultrasound imaging is used extensively in obstetrics. You have all probably seen images of a fetus—looking quite baby-like—produced by ultrasound; such an image is shown in Figure I.1(c). Figure IV.1(b) shows another such image, taken during the third trimester. If the angle of the view is just right, you can even tell the baby's gender.

We will learn later that a key quality of ultrasound images is its pronounced texture, which is produced by a combination of the small-scale structures within the body's tissues and a physical phenomenon associated with wave imaging called *speckle*. The two images in Figure IV.2 show these variations in texture as well as the anatomical shapes of (a) a breast mass and (b) a kidney. With training, physicians can use texture differences to characterize tissues as



(a) (b)

Figure IV.2
Detailed ultrasound examinations of (a) a breast mass and (b) a kidney. Images courtesy of GE Healthcare.



(a)

(b)

Figure IV.3 Ultrasound guidance for a liver biopsy. (a) Planning mode with a graphic to aid insertion and (b) needle visualization after insertion. Images courtesy of Dr. Avneesh Chhabra, Johns Hopkins University, Baltimore, MD.

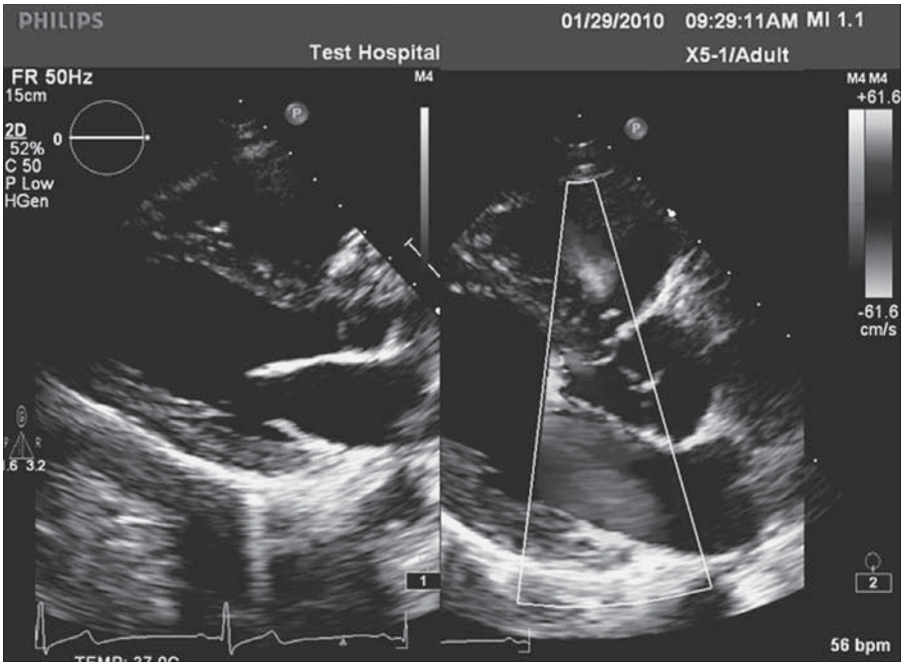


Figure IV.4 Mitral valve regurgitation depicted using Doppler ultrasound. Image courtesy of Philips Healthcare.

healthy or diseased, thereby providing another dimension beyond simple anatomical differences upon which to base a diagnosis.

Another important property of ultrasound imaging is that it is real-time in nature—and has been so since its inception. Because of this property and the fact that it is completely safe, ultrasound is widely used for image-guided surgical interventions. An example showing a simple planning graphic overlaid on an ultrasound image in preparation for a liver biopsy is shown in Figure IV.3(a). After the biopsy needle is inserted, ultrasound imaging can be used to verify and adjust its position as shown in Figure IV.3(b).

In addition to depicting varying sound reflectance, ultrasound is capable of portraying the velocities of moving bodies and substances, such as blood within vessels. This so-called *Doppler imaging* technique can provide a quantitative measure of blood flow. The combination of anatomical and functional information ultrasound is able to provide—in a fully noninvasive manner—makes it a popular imaging modality. An image showing the flow through and around the mitral valve of the heart is shown in Figure IV.4. The left image shows a parasternal long-axis cross section of the mitral valve and the right image shows an overlay of the blood flow velocities in that region (within the sector outlined in white). Doppler images are always shown in color where black indicates no flow, red indicates flow towards the transducer, and blue indicates flow away from the transducer. Because images are only printed in grayscale in this book, the overlaid colors within the outlined sector clearly depicting mitral valve regurgitation cannot be see here. Examples are readily found on the internet.

CHAPTER

10

The Physics of Ultrasound

10.1 Introduction

Ultrasound is sound with frequencies higher than the highest frequency that can be heard by human beings. This means that any sound above about 20 kHz is considered to be ultrasound. For reasons that we will provide in this chapter and in Chapter 11, medical ultrasound systems operate at much higher frequencies than this defining frequency, typically between 1 and 10 MHz, although there are specialized systems operating in ranges up to 70 MHz. The principles of ultrasound propagation, regardless of the exact frequency range, are the same as those of ordinary sound propagation and are defined by the theory of *acoustics*. In particular, ultrasound moves in a wavelike fashion by expansion and compression of the medium through which it is moving, and ultrasound waves travel at certain speeds, depending on the material through which they are traveling. In addition, ultrasound waves can be absorbed, refracted, focused, reflected, and scattered.

Ultrasound provides a noninvasive technique for imaging human anatomy. A *transducer*, which converts electrical signals to acoustic signals, generates pulses of ultrasound, which are sent through a patient's body. Organ boundaries and complex tissues produce echoes (by reflection or scattering) that return back to and are detected by the transducer, which converts the acoustic signal to an electrical signal. The ultrasound imaging system then processes the echoes and presents a grayscale image of human anatomy on a display. Each point in the image corresponds to the anatomic location of an echo-generating structure, and its brightness corresponds to the echo strength. The dominant diagnostic areas in which ultrasound is used are obstetrics, oncology, cardiology, and gastroenterology.

The use of ultrasound in diagnostic imaging dates back to the mid-1950s. Rapid expansion of its use began in the early 1970s with the advent of two-dimensional real-time ultrasonic scanners. Additional milestones were the appearance of phased array systems in the early 1980s, color flow-imaging systems in the mid-1980s, three-dimensional imaging systems in the 1990s, and nonlinear imaging systems in the 2000s. Manufacturers offer several types of ultrasonic medical imaging systems, each one having interchangeable transducers and the ability to operate in a variety of modes. This is because different parts of the human anatomy have different imaging requirements.

In this chapter, we study the physics of ultrasound and its interaction with biological tissues. This background is necessary for a complete understanding of medical ultrasound imaging systems, which will be discussed in the next chapter.

10.2 The Wave Equation

Acoustic waves are pressure waves that propagate through matter via compression and expansion of the matter. A wave can be generated by compressing a small volume of tissue and then releasing it. The elastic properties of the material cause the compressed volume to expand past its point of equilibrium, making neighboring volumes compress. As this process continues through successive volumes of tissue, a wave is generated.

As an acoustic wave propagates, small particles of the material—e.g., its molecules—move back and forth in order to generate the compressions and expansions of the acoustic wave. In medical ultrasound, where waves propagate only in soft tissues, these particles move back and forth in the same direction that the acoustic wave is traveling. When this happens, the acoustic wave is called a *longitudinal wave*. Harder materials also support *shear waves*, in which the particles move at right angles to the direction of propagation of the acoustic wave, but these types of waves do not need to be considered in medical ultrasound imaging.

10.2.1 Three-Dimensional Acoustic Waves

The *compressibility* κ and *density* ρ of a material, combined with the laws of conservation of mass and momentum, directly imply the existence of acoustic waves. ($B = 1/\kappa$ is called the *bulk modulus* of a material.) These waves travel at a *speed of sound* c, given by

$$c = \sqrt{\frac{1}{\kappa\rho}}\,. \tag{10.1}$$

Table 10.1 gives the speed of sound (as well as other information) for various materials and biological tissues. A good rule of thumb is that sound travels about 1,540 m/s in tissue and 330 m/s in air.

There are a variety of ways to mathematically characterize an acoustic wave. First, it is important to realize that an acoustic wave is inherently a 3-D phenomenon—that is, it has spatial dependencies in all three directions. For example, the sound generated in ordinary speech spreads out in all directions so that people can hear the speaker from many orientations. An acoustic wave also depends on time—that is, it has temporal dependencies. In ordinary speech, for example, a sudden exclamation or shout dies away momentarily. So, whatever the physical quantity that is used to describe the wave, we realize that it must depend upon three spatial variables, x, y, z, and time, t.

TABLE 10.1

Acoustic Properties of Various Materials

Material	Density, ρ [kg m^{-3}]	Speed, c [m s^{-1}]	Characteristic Impedance, Z [kg m^{-2} s^{-1}] ($\times$ 10^{6})	Absorption Coefficient, α [dB cm^{-1}] (at 1 MHz)	Approximate Frequency Dependence of α
Air at STP	1.2	330	0.0004	12	f^2
Aluminum	2,700	6,400	17	0.018	f
Brass	8,500	4,490	38	0.020	f
Castor oil	950	1,500	1.4	0.95	f^2
Mercury	13,600	1,450	20	0.00048	f^2
Polyethylene	920	2,000	1.8	4.7	$f^{1.1}$
Polymethyl-methacrylate	1,190	2,680	3.2	2.0	f
Water	1,000	1,480	1.5	0.0022	f^2
Blood	1,060	1,570	1.62	[0.15]	
Bone	1,380–1,810	4,080	3.75–7.38	[14.2–25.2]	
Brain	1,030		1.55–1.66	[0.75]	
Fat	920	1,450	1.35	[0.63]	
Kidney	1,040	1,560	1.62	—	
Liver	1,060	1,570	1.64–1.68	[1.2]	
Lung	400		0.26	[40]	
Muscle	1,070		1.65–1.74	[0.96–1.4]	
Spleen	1,060		1.65–1.67	—	
Water	1,000	1,484	1.52	[0.0022]	

As noted above, as an acoustic wave propagates in a given material, small particles of the material move back and forth in order to generate the compressions and expansions of the acoustic wave. A traditional treatment of acoustics would begin by deriving a mathematical expression for the *particle displacement* $\vec{u}(x, y, z, t)$ associated with the compression and expansion of the acoustic wave. The *particle velocity*, $\vec{v}(x, y, z, t)$, the temporal derivative of particle displacement, is another common choice for describing an acoustic wave. Since we are constraining our discussion to longitudinal waves, it is actually only necessary to describe the magnitudes of particle displacement and velocity—$u(x, y, z, t)$ and $v(x, y, z, t)$, respectively—since the orientation of these quantities is always identical to the direction of wave propagation. However, we are actually going to take this a step farther. Since the compression and expansion of a small volume are associated with a local change in the material's pressure, an acoustic wave can also be described in terms of a spatially dependent, time-varying pressure function $p(x, y, z, t)$, called the *acoustic pressure*. Acoustic pressure is zero in the absence of an acoustic wave, so it should be considered to be the variation of pressure caused by the acoustical disturbance around the ambient pressure.

In longitudinal waves, it is straightforward to relate the acoustic pressure to the underlying particle velocity. They are, in fact, linearly related by

$$p = Zv\,, \tag{10.2}$$

where

$$Z = \rho c \tag{10.3}$$

is called the *characteristic impedance*. The particle speed v and the speed of sound c are *not* the same quantity and are generally not equal. This is a common point of confusion and source of computational mistakes. The term *impedance* arises by analogy to electrical circuits or electromagnetic transmission lines. In particular, acoustic pressure is analogous to voltage, while particle speed is analogous to current. Since density has units kg/m^3 and speed has units m/s, the characteristic impedance Z has units kg/m^{2}s, which is given the name *rayls*, after Lord Rayleigh.

Throughout this chapter and the next, we use acoustic pressure to describe longitudinal acoustic waves. Where necessary (e.g., in deriving boundary conditions), we refer to the velocity or displacement. From the physical properties of matter, it turns out that acoustic pressure p must satisfy the following *three-dimensional wave equation*:

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}\,, \tag{10.4}$$

where ∇^2 is the 3-D Laplacian operator given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\,. \tag{10.5}$$

In general, a wave equation is a partial differential equation relating spatial partial derivatives of pressure to the temporal partial derivative of pressure. Solving this equation is difficult in general, but our needs will be satisfied if we consider two special cases: plane waves and spherical waves.

10.2.2 Plane Waves

If an acoustic wave varies in only one spatial direction and time, it is called a *plane wave*. For example, if $p(x, y, z, t)$ were constant for any choice of x and y given a fixed z and t, then the resulting pressure function $p(z, t) = p(x, y, z, t)$ describes a plane wave moving in the $+z$ or $-z$ direction. In this case, plugging $p(z, t)$ into (10.4) yields the *one-dimensional wave equation*:

$$\frac{\partial^2 p}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}\,. \tag{10.6}$$

Notice that a one-dimensional wave equation exists for any direction of propagation, not just the z direction. In the following, it will be convenient to think of the z direction in the sense of a rotated coordinate system—that is, z is the direction of plane wave propagation not a fixed “laboratory” coordinate direction.

The general solution to (10.6) can be written as

$$p(z, t) = \phi_f(t - c^{-1}z) + \phi_b(t + c^{-1}z)\,, \tag{10.7}$$

which can be verified by direct substitution into (10.6). The function $\phi_f(t - c^{-1}z)$ is interpreted as a *forward-traveling wave*, since the basic waveform $\phi_f(c^{-1}z)$—that is, the spatial pressure function at $t = 0$—is shifted in the positive z direction as t increases. The function $\phi_b(t + c^{-1}z)$ is interpreted as a *backward-traveling wave* for an analogous reason. The only requirement of the functions $\phi_f(\cdot)$ and $\phi_b(\cdot)$ is that they be twice differentiable, although even this assumption can be relaxed by using generalized derivatives. Since each function ϕ_f and ϕ_b satisfies the wave equation independently, one of them might be identically zero, so we can have only a forward traveling wave in a given medium, for example. In Chapter 11, we will use the plane wave to approximate the acoustic wave produced by certain types of ultrasound transducers.

An important example of a function that satisfies the one-dimensional wave equation is the *sinusoidal* function

$$p(z, t) = \cos k(z - ct)\,. \tag{10.8}$$

Viewed as a function of t only (holding z fixed), we see that the pressure around a fixed particle varies sinusoidally with (radial) frequency $\omega = kc$. This corresponds to a (cyclic) frequency of

$$f = \frac{\omega}{2\pi} = \frac{kc}{2\pi}\,, \tag{10.9}$$

which has units of Hertz (Hz) or cycles per second. Viewed as a function of z only (holding t fixed), we see that the pressure at a particular time varies sinusoidally with (radial, spatial) frequency k, a quantity that is known as the *wave number*. The *wavelength* λ of this sinusoidal wave is the spacing between crests or troughs at any given time and is given by

$$\lambda = \frac{2\pi}{k}\,, \tag{10.10}$$

which has units of length. Solving (10.9) for k and substituting the resulting expression into (10.10) yields the important relationship between wavelength, speed of sound, and frequency

$$\lambda = \frac{c}{f}\,. \tag{10.11}$$

Truly sinusoidal waves cannot exist in nature because they need infinite space and infinite time. However, medical ultrasound systems use waveforms that are approximately sinusoidal when viewed locally over a short period of time. For example, it is common for ultrasound imaging systems to operate at 3.5 MHz, which corresponds to a wavelength of 0.44 mm when $c = 1,540$ m/s. Although it cannot be made entirely clear at this point why this would be the case, the wavelength of the acoustic wave is approximately equal to the achievable resolution of the system (at least in the direction of the propagating wave). Therefore, we should understand that submillimeter resolution might be achievable in ultrasound imaging.

EXAMPLE 10.1

Suppose a steady-state sinusoidal wave with frequency 2 MHz is traveling in the $+z$ direction in the liver.

Question What is its wavelength?

Answer From Table 10.1, we find that $c = 1{,}570$ m/s in liver. Therefore, the wavelength is $\lambda = 1{,}570 \text{ ms}^{-1}/(2 \times 10^6 \text{ Hz}) = 0.785$ mm.

■

Another important example of a function that satisfies the 1-D wave equation is

$$p(z,t) = \delta(z - ct)\,, \tag{10.12}$$

where $\delta(\cdot)$ is the unit impulse (delta) function (see Section 2.2.1). Since $\delta(\cdot)$ is a generalized function, it satisfies the wave equation only through the use of generalized derivatives. Like sinusoidal waves, the impulse function cannot exist in practice. In this case, the impulse plane wave is supposed to have infinite extent orthogonal to the direction of propagation and an infinitesimally thin extent in the direction of propagation, neither of which can be achieved in practice. Nevertheless, the impulse plane wave is useful for approximating the short pulses that are used in medical ultrasound systems, in order to better understand the overall properties of these systems.

EXAMPLE 10.2

An ultrasound transducer is pointing down the $+z$ axis. Starting at time $t = 0$, it generates an acoustic pulse with form

$$\phi(t) = (1 - e^{-t/\tau_1})e^{-t/\tau_2}.$$

Question Assume the speed of the sound is $c = 1{,}540$ m/s. What is the forward traveling wave down the $+z$ axis? At what time does the leading edge of the impulse hit the interface 10 cm away from the transducer?

Answer The forward traveling wave down the $+z$ axis is

$$\phi_f(z,t) = (1 - e^{-(t-z/c)/\tau_1})e^{-(t-z/c)/\tau_2}.$$

It will take $0.1\,\text{m}/(1{,}540\,\text{m/s}) = 64.9\,\mu$s for the leading edge of the ultrasound impulse to travel 10 cm and hit the interface. ■

10.2.3 Spherical Waves

In an isotropic material, a spherical wave can be generated by a small, local disturbance in the pressure. A spherical wave depends only on time t and the radius $r = \sqrt{x^2 + y^2 + z^2}$ from the source of the disturbance, which is assumed to be at $(0, 0, 0)$. Using (10.4), evaluating the Laplacian (viewing r as a function of x, y, and z), and manipulating the resulting equation (see Problem 10.4), the pressure $p(r,t)$ of a wave traveling in the radial direction can be shown to satisfy

$$\frac{1}{r}\frac{\partial^2}{\partial r^2}(rp) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}\,, \tag{10.13}$$

which is known as the *spherical wave equation*.

The general solution to the spherical wave equation is

$$p(r,t) = \frac{1}{r}\phi_o(t - c^{-1}r) + \frac{1}{r}\phi_i(t + c^{-1}r)\,, \tag{10.14}$$

which can be verified by direct substitution. The wave $\phi_o(t - c^{-1}r)$ is an *outward-traveling wave*, and $\phi_i(t + c^{-1}r)$ is an *inward-traveling wave*. Generally, there is no source that will create an inward-traveling wave, so the most general solution we need to consider here is

$$p(r,t) = \frac{1}{r}\phi_o(t - c^{-1}r)\,. \tag{10.15}$$

As in the 1-D plane wave solution, $\phi_o(\cdot)$ should be twice differentiable, but otherwise it is an arbitrary function. This solution looks very much like the forward-traveling wave in the solution to the 1-D wave equation (10.7), except for the factor of $1/r$, which causes the outward spherical wave to lose amplitude. This loss is simply due to the increasing overall surface area of the wave as it propagates radially outward from its source.

10.3 Wave Propagation

10.3.1 Acoustic Energy and Intensity

An acoustic wave carries energy with it. Particles in motion have kinetic energy, and those that are poised for motion have potential energy. To characterize these energies in a wave, we define the energy per unit volume, leading to the *kinetic energy density*

$$w_k = \frac{1}{2}\rho v^2\,, \tag{10.16}$$

and the *potential energy density*

$$w_p = \frac{1}{2}\kappa p^2\,. \tag{10.17}$$

The *acoustic energy density* is defined as the sum of these two:

$$w = w_k + w_p\,. \tag{10.18}$$

This quantity captures the idea of a change in energy at a point in space as a wave passes through it.

To capture the idea of energy that moves with the wave, we define the *acoustic intensity* as

$$I = pv\,, \tag{10.19}$$

which is also called the *acoustic energy flux*. Notice that by using the analogy to electric circuits, I is analogous to electric power, which gives a strong intuitive connection to the present application. In particular, by substituting (10.2) into (10.19), we find that

$$I = \frac{p^2}{Z}\,, \tag{10.20}$$

which is analogous to the expression in electric circuits, V^2/R, that gives the power in a resistor. There is power in acoustic waves. Unlike the case of the resistor, which converts electrical power to thermal energy, the acoustic wave simply carries the power along with it as it propagates (unless there is acoustic attenuation, as described in Section 10.3.4). In fact, it can be shown that the acoustic energy density and the acoustic intensity are related by the equation of energy conservation

$$\frac{\partial I}{\partial x} + \frac{\partial w}{\partial t} = 0\,, \tag{10.21}$$

which describes the propagation of power and energy in an acoustic wave.

10.3.2 Reflection and Refraction at Plane Interfaces

Figure 10.1 depicts a plane wave incident upon a plane interface with incidence angle θ_i (measured from the nearest surface normal). Assuming that the wavelength of the sound is small with respect to the spatial extent of the interface, the reflected and transmitted wave directions will obey the laws of geometric optics:

$$\theta_i = \theta_r \tag{10.22}$$

and

$$\frac{\sin\theta_i}{\sin\theta_t} = \frac{c_1}{c_2}\,, \tag{10.23}$$

where c_1 and c_2 are the speeds of sound in medium 1 and medium 2, respectively. Equation (10.23) is known as *Snell's law*.

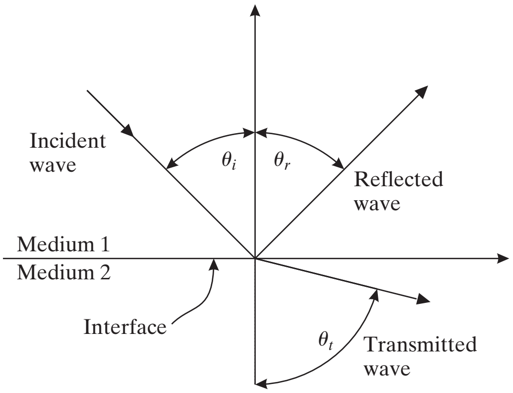


Figure 10.1
An incident plane wave reflects some energy and transmits the rest at a plane interface.

EXAMPLE 10.3

Suppose medium 1 is fat and medium 2 is liver, and a plane wave is incident upon their interface with incidence angle $\theta_i = 45^\circ$.

Question What are the reflection and transmission angles θ_r and θ_t, respectively?

Answer $\theta_r = \theta_i = 45^\circ$. Since $c_1 = 1{,}450$ m/s and $c_2 = 1{,}570$ m/s,

$$\sin\theta_t = \frac{1{,}570 \sin 45^\circ}{1{,}450} = 0.7656\,.$$

Solving this yields $\theta_t = 49.96^\circ$. It makes intuitive sense that the transmission angle should be larger than the incidence or reflection angle, since $c_2 > c_1$. ■

It may be that the quantity $c_2 \sin\theta_i / c_1$ is greater than 1.0. In this case, the inverse sine does not exist, and we conclude that all the energy is reflected. For a given pair of materials with speeds c_1 and c_2 in which $c_2 > c_1$, all incident angles above the so-called *critical angle* θ_c, given by

$$\theta_c = \sin^{-1}(c_1/c_2) \quad \text{for } c_2 > c_1, \tag{10.24}$$

will result in total reflection.

10.3.3 Transmission and Reflection Coefficients at Plane Interfaces

In the previous section, we examined the geometric properties of plane waves at plane intersections. Now we consider the energy characteristics in this same situation. Since all three waves—that is, the incident, reflected, and transmitted waves—meet at the interface, the tangential particle motion caused by the incident wave must coincide with the sum of the tangential particle motions of transmitted and reflected waves. That is,

$$v_i \cos\theta_i = v_r \cos\theta_r + v_t \cos\theta_t\,.$$

Since $p = Zv$ [see (10.2)], we may replace the velocities in the above expression to get

$$\frac{\cos\theta_t}{Z_2} p_t + \frac{\cos\theta_r}{Z_1} p_r = \frac{\cos\theta_i}{Z_1} p_i\,. \tag{10.25}$$

Since pressure must be continuous across the interface, we also have

$$p_t - p_r = p_i\,. \tag{10.26}$$

Equations (10.25) and (10.26) are simultaneous linear equations with unknowns p_t and p_r, which when solved yield

$$R = \frac{p_r}{p_i} = \frac{Z_2\cos\theta_i - Z_1\cos\theta_t}{Z_2\cos\theta_i + Z_1\cos\theta_t}\,, \tag{10.27}$$

$$T = \frac{p_t}{p_i} = \frac{2Z_2\cos\theta_i}{Z_2\cos\theta_i + Z_1\cos\theta_t}\,, \tag{10.28}$$

where we have used the fact that $\theta_r = \theta_i$. The quantity R is called the *pressure reflectivity*, and the quantity T is called the *pressure transmittivity*.

Using the relationship $I = p^2/Z$ [see (10.20)], the *intensity reflectivity* is given by

$$R_I = \frac{I_r}{I_i} = \left(\frac{Z_2 \cos\theta_i - Z_1 \cos\theta_t}{Z_2 \cos\theta_i + Z_1 \cos\theta_t} \right)^2 , \tag{10.29}$$

and the *intensity transmittivity* is given by

$$T_I = \frac{I_t}{I_i} = \frac{4 Z_1 Z_2 \cos^2\theta_i}{(Z_2 \cos\theta_i + Z_1 \cos\theta_t)^2} . \tag{10.30}$$

In deriving the intensity transmittivity, the fact that p_t and p_i are in different media must be taken into account.

EXAMPLE 10.4

Consider an acoustic wave encountering a fat/liver interface at normal incidence.

Question What fraction of acoustic intensity is reflected back from a fat/liver interface at normal incidence? Does it matter from which direction the incident energy arrives?

Answer At normal incidence, we have

$$R_I = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2 .$$

The acoustic impedance of fat is 1.35×10^{-6} kg m^{-2}s^{-1}. The acoustic impedance of liver is 1.66×10^{-6} kg m^{-2}s^{-1} (nominal value). For propagation from fat to liver, we have

$$R_I = \left(\frac{1.66 - 1.35}{1.66 + 1.35} \right)^2 = 0.0106 .$$

Only about 1% of the incident power is reflected back from the interface; about 99% is transmitted through. From the expression above, it does not matter from which direction the incident wave originates; the same fractions are calculated in either direction. (Is this also true for nonnormal incidence?) ■

10.3.4 Attenuation

In practice, the amplitude of a real acoustic wave decreases as the wave propagates. *Attenuation* is the term used to account for loss of wave amplitude (or "signal") due to all mechanisms, including absorption, scattering, and mode conversion. *Absorption* is the process by which the wave energy is converted to thermal energy, which is then dissipated in the medium. *Scattering* is the process by which secondary spherical waves are generated as the wave propagates. *Mode conversion* is the process by which longitudinal waves are converted to transverse shear waves (and back again). In this section, we develop a phenomenological expression for attenuation, which includes all mechanisms for signal loss.

Consider a forward-traveling plane wave $p(z,t)$ in the $+z$-direction, where $p(0,t) = A_0 f(t)$. Then, under ideal circumstances $p(z,t) = A_0 f(t - c^{-1}z)$. Due to attenuation, however, we actually have

$$p(z,t) = A_z f(t - c^{-1}z)\,, \tag{10.31}$$

where A_z is the actual amplitude of the traveling wave and is dependent on the z-position of the wave. We model the amplitude decay as

$$A_z = A_0 e^{-\mu_a z}\,, \tag{10.32}$$

where μ_a is called the *amplitude attenuation factor* and has units cm^{-1}. This model of attenuation is *phenomenological*, meaning it agrees well in practice but is not easily supported by theory. In particular, the pressure function given in (10.31) no longer satisfies the wave equation.

The unit given to the natural logarithm of an amplitude ratio is nepers (Np), pronounced “nay-pers,” and the units of μ_a are sometimes quoted in nepers/cm. This is because

$$\mu_a = -\frac{1}{z} \ln \frac{A_z}{A_0}\,. \tag{10.33}$$

Since $20 \log_{10} \frac{A_z}{A_0}$ is the amplitude gain in decibels (dB), it is useful to define the *attenuation coefficient* α as

$$\alpha = 20(\log_{10} e)\mu_a \approx 8.7\mu_a\,, \tag{10.34}$$

which has the units dB/cm. Notice that $1\ \text{Np} = 8.686\ \text{dB}$. Be careful when calculating the amplitude loss when given an attenuation coefficient α. You must first convert α to the amplitude attenuation factor μ_a [via (10.34)], and then use the amplitude loss equation (10.32).

When attenuation is due solely to the conversion of acoustic energy to thermal energy, the attenuation coefficient is called the *absorption coefficient*. Most measurements of attenuation in materials are conducted so that absorption is the dominant loss mechanism. Some typical absorption coefficients for biological and nonbiological materials are given in Table 10.1.

The absorption coefficient of a material is generally dependent on frequency f. A good model for this dependency is

$$\alpha = af^b\,, \tag{10.35}$$

where b is just slightly greater than 1 in biological tissues. For example, in homogenized liver, $a = 0.56$ and $b = 1.12$. The rough approximation that $b = 1$ is often used, leading to a linear relationship between α and frequency f. The values of a in Table 10.2 are based on the assumption that $b = 1$.

TABLE 10.2

Frequency Dependence of Various Biological Tissues

Material	$a = \alpha/f$ [dB cm^{-1} MHz^{-1}]
Fat	0.63
Skeletal muscle	
Along fibers	1.3
Across fibers	3.3
Cardiac muscle	1.8
Blood	0.18
Bone	20.0
Lung	41.0
Liver	0.94
Kidney	1.0
Brain	
White matter along fibers	2.5
White matter across fibers	1.2
Gray matter	0.5–1.0

EXAMPLE 10.5

Suppose a 5-MHz acoustic pulse travels from a transducer through 2 cm of fat, then encounters an interface with the liver at normal incidence.

Question At what time interval after the transmitted pulse will the reflected pulse—that is, the *echo*—arrive back at the transducer? Taking both attenuation and reflection losses into account, what will be the amplitude loss in decibels of the returning waveform?

Answer Table 10.1 shows that the speed of sound in fat is 1,450 m/s. The round-trip travel distance is 4 cm. Therefore, the echo will return after the interval

$$t = \frac{0.04\ \text{m}}{1{,}450\ \text{m/s}} = 27.6\ \mu\text{s}\,.$$

Table 10.2 shows that $a = 0.63$ dB cm^{-1} MHz^{-1}. Therefore the absorption coefficient at 5 MHz is

$$\begin{aligned}\alpha &= 0.63\ \text{dB cm}^{-1}\ \text{MHz}^{-1} \times 5\ \text{MHz} \\ &= 3.15\ \text{dB cm}^{-1}\,.\end{aligned}$$

The amplitude attenuation factor is $\mu_a = 3.15\ \text{dB cm}^{-1}/(8.686\ \text{dB/Np}) = 0.363$ Np/cm. The roundtrip amplitude ratio (ignoring the reflection loss) is therefore

$$\frac{A_z}{A_0} = \exp\{-0.363\ \text{Np/cm} \times 4.0\ \text{cm}\} = 0.234\,.$$

From the previous example, we know that the intensity reflectivity is 0.0106. The amplitude reflectivity is therefore $\sqrt{0.0106} = 0.103$. Putting these facts together yields

$$\text{dB loss} = 20\log_{10}\frac{A_z}{A_0} = 20\log_{10}(0.234 \times 0.103) = -32.4\ \text{dB}\,.$$

■

10.3.5 Scattering

Many targets within the body are much smaller than the acoustic wavelength. For these targets, the geometric equations for reflection and refraction do not hold. Instead, we assume that when these targets are excited by an incident acoustic plane wave, they vibrate as small spherical bodies, giving rise to spherical waves whose amplitude is some fraction of the incident wave amplitude.

Consider the plane (attenuated) wave

$$p(z,t) = A_0 e^{-\mu_a z} f(t - c^{-1}z)$$

incident upon a small pointlike target residing at $(0,0,d)$, as shown in Figure 10.2. Notice that this wave is traveling in the $+z$ direction. The small target acts as a spherical wave source, converting a fraction R of the incident plane wave into a spherical wave centered at $(0,0,d)$. Called the *reflection coefficient*, R is a property of the individual target and the embedding medium. Treating $(0,0,d)$ as the origin [which means that r is the radius from $(0,0,d)$], the resulting scattered wave is given by

$$p_s(r,t) = \frac{Re^{-\mu_a r}}{r} A_0 e^{-\mu_a d} f(t - c^{-1}d - c^{-1}r)\,. \tag{10.36}$$

Here, we have included attenuation terms for both the incident plane wave and the scattered spherical wave, in addition to the reflection coefficient and the natural $1/r$ decay of the spherical wave. The wave is also delayed in time by $c^{-1}d$ due to the propagation delay from the true origin to $(0,0,d)$. This equation is at the core of our ultrasound imaging equation, presented in Chapter 11.

10.3.6 Nonlinear Wave Propagation

Up to this point, we have assumed that the acoustic waves in an ultrasound imaging system follow the principle of linear propagation. Basically, we have assumed that a sinusoidal excitation yields a sinusoidal response—that is, as a wave travels through the body its shape remains the same. This turns out to be not quite true, even in the low power regime of diagnostic ultrasound imaging systems.[1] As noted in Chapter 2, linear systems have many desirable properties that lend themselves to analysis and system design. And most ultrasound imaging systems are fundamentally designed around the principle of linearity as we shall see in Chapter 11. However, nonlinearities are present and they should be

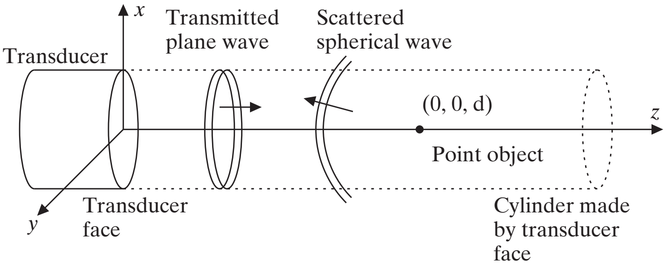


Figure 10.2 Basic pulse-echo experiment.

---

[1] Nonlinearities are most definitely present in high-power ultrasound therapy systems.

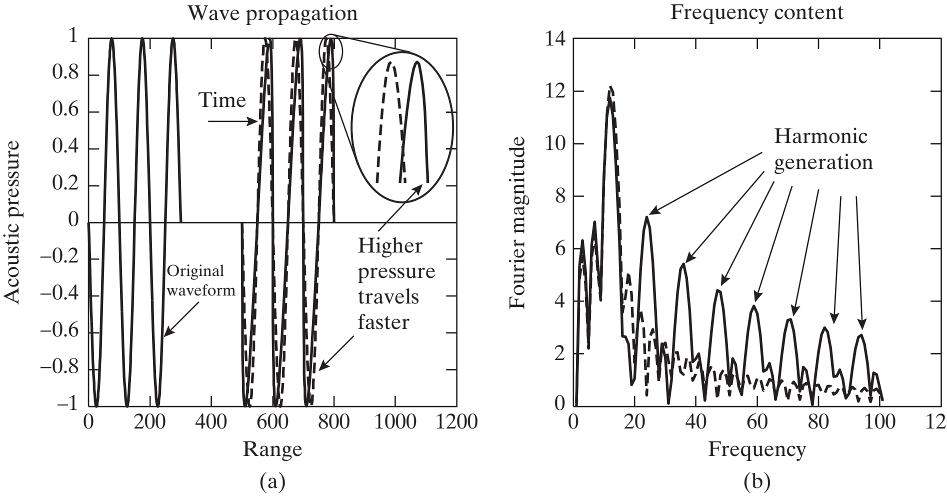
understood since they can be exploited in conjunction with ultrasound contrast agents and tissue ablation procedures, and to improve resolution.

The nonlinear characteristic of longitudinal wave propagation occurs because the speed of sound for a given material actually depends on the acoustic pressure that is present. To a good approximation, the speed of sound in the nonlinear regime is given by

$$c = c_0 + \beta \frac{p}{Z_0}\,, \tag{10.37}$$

where c_0 is the speed given in (10.1) and Z_0 is the acoustic impedance given in (10.3) (both assuming a nominal tissue density ρ_0 determined in the absence of acoustic excitation). Here, p is the acoustic pressure at a given location and time—that is, $p = p(x, y, z, t)$—in general. The constant β is called the *coefficient of nonlinearity* and it depends on the acoustic medium. In tissue, β is typically in the range 3–6, lower in water and blood and higher in fatty tissues, and it is always positive so that tissue points having higher pressure (or equivalently higher particle velocity) yields a higher acoustic wave speed.

The practical result of this acoustic nonlinearity is to distort the acoustic waveform as it propagates. The linear wave equation predicts that a wave will propagate unchanged (ignoring attenuation), as depicted by the dashed wave in Figure 10.3(a). However, when points having higher acoustic pressure send the wave on faster (due to higher c), this causes different isosurfaces of constant pressure to move at different speeds, with higher pressures arriving at distant locations first. At any given spatial point, the observed waveform is therefore distorted from its original shape. In Figure 10.3(a), the solid waveform has its higher pressures (crests) move faster than its lower pressures (troughs), which



(a) (b)

Figure 10.3
Illustration of nonlinear acoustic wave propagation. (a) An initial single cycle sinusoidal pulse (left solid curve) propagates with no shape change under the linear wave equation (right dashed curve) but gets distorted under nonlinear propagation (right solid curve curve). (b) Nonlinear propagation causes the frequency content of the original pulse (dotted curve) to gain higher frequency harmonics (solid curve) as it propagates.

causes the waveform to become more and more like a sawtooth, even though it started out as a sinusoid.

As the waveform begins to distort due to nonlinear propagation, harmonic frequencies begin to develop, as shown by the solid curve in Figure 10.3(b). If the signal were perfectly periodic then the harmonics would be characterized by the growth of coefficients in a Fourier series at perfect integer multiples of the fundamental frequency. This phenomenon is called *harmonic generation*. If the fundamental frequency were at 3 MHz, for example, over time there would be evidence of signal components at 6 MHz, 9 MHz, 12 MHz, etc. Figure 10.3(b) shows such growth at periodic multiple of the fundamental, but due to the finite character of the pulse, there is also evidence of some frequency spread around each of the multiples. Therefore, the frequencies that are present in a propagating nonlinear wave depend on the shape of the initial wave including its amplitude, its fundamental frequency, and its duration.

It is straightforward to see how nonlinear propagation can be exploited when differences in the coefficient of nonlinearity β can help to differentiate tissues or materials. For example, harmonic generation in a suspension of microbubbles is very substantial. Therefore, microbubbles make excellent contrast agents for functional imaging in conjunction with nonlinear ultrasound. As another example, in ablation procedures—e.g., using microwaves, lasers, or radio frequency electrodes—tissue is heated in order to destroy cancer cells. It turns out that β in tissue is dependent on temperature, which makes it possible to monitor temperature using harmonic (nonlinear) ultrasound imaging techniques. Finally, a somewhat unexpected benefit of the nonlinear propagation property of ultrasound is that it can be used to improve resolution. We will see how this is possible in Chapter 11.

If one thinks carefully about the development of the nonlinear waveform distortion, it becomes apparent that at some time it is possible that the propagation of the crest of a waveform could outpace a part of the waveform having lower acoustic pressure. This would seem to imply that there are spatial points having two acoustic pressures simultaneously, which is clearly impossible. In fact, when such a situation develops, the wave takes on a discontinuity and is known as a *shock wave*, which at large propagation distances, converges to a sawtooth waveform. For three reasons, it is uncommon for shock waves to develop in medical ultrasound imaging: the coefficient of nonlinearity in tissues is relatively small; the acoustic pressures in imaging systems are relatively small; and the absorption is relatively large, especially as frequency increases, which tends to fight the growth of the high frequencies that are characteristic of nonlinear propagating waveforms.

10.4 Doppler Effect

The *Doppler effect*, also known as the *Doppler shift*, is the change in frequency of sound due to the relative motion of the source and receiver. The Doppler effect we most commonly experience is that of the siren of an emergency vehicle. The pitch of the siren is higher as the vehicle approaches and becomes lower as the vehicle passes and moves away.

The idea behind the Doppler effect for a moving source and stationary observer is illustrated in Figure 10.4(a). Here, the object **O**, which is emitting sound and is therefore the source, moves to the right, away from the transducer

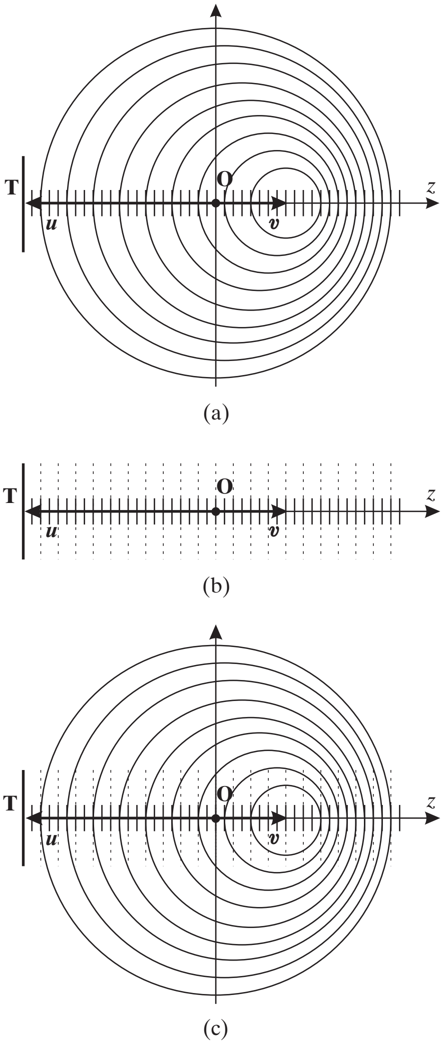


Figure 10.4
Basic ideas behind the Doppler effect: (a) a moving source **O** and stationary observer **T**, (b) a stationary source **T** and moving observer **O**, and (c) a stationary source/receiver **T** and moving scatterer **O**, which also acts as both a receiver and a source.

T, which is receiving the sound and is therefore the observer. The frequency f_T of the sound received by **T** can be derived by considering the separation of the wave crests that are emanating from the moving source. If the source is producing a sinusoid with frequency f_O, then the crests occur at the interval $1/f_O$—that is, the period. In one period, the crest will propagate a distance of c/f_O, where c is the speed of sound in the medium. In this same interval the source moves a distance of v/f_O, where v is the speed of the source, and at this moment another crest will be produced by the source. So, the physical separation of crests—that is, the wavelength—observed by **T** is $(c + v)/f_O$. Therefore, **T** will receive (i.e., hear) an acoustic wave having frequency

$$f_T = \frac{c}{c+v} f_O \,. \tag{10.38}$$

This situation is generalized for an arbitrary observer velocity by recognizing that the frequency shift depends only on the component of source velocity in the direction of the observer (relative to where the source was when it generated the sound that is being received). Accordingly, the general expression for the sound frequency f_T observed by a stationary observer from a moving source with frequency f_O is

$$f_T = \frac{c}{c - v\cos\theta} f_O \,, \tag{10.39}$$

where θ is the angle between the vector pointing from the source to the receiver and the vector pointing from the source in its direction of its motion. Specifically, if $\mathbf{x}_T$ and $\mathbf{x}_O$ give the position of the transducer and object, respectively, and if $\mathbf{v}$ is the velocity (vector) of $\mathbf{O}$, then θ is the angle between vectors $\mathbf{v}$ and the vector

$$\mathbf{u} = \mathbf{x}_T - \mathbf{x}_O \,, \tag{10.40}$$

as shown in Figure 10.4. In the situation depicted in Figure 10.4(a), $\cos\theta = -1$, and (10.39) reduces to (10.38).

The *Doppler frequency* f_D is the difference between the observed frequency and the source frequency. For the moving source situation in Figure 10.4(a), the Doppler frequency is given by

$$f_D = f_T - f_O \,. \tag{10.41}$$

Substituting (10.39) into (10.41) yields

$$f_D = \left(\frac{v\cos\theta}{c - v\cos\theta} \right) f_O \,, \tag{10.42}$$

after some algebra. Since v is ordinarily much smaller than c, this can be approximated as

$$f_D \approx \left(\frac{v\cos\theta}{c} \right) f_O \,. \tag{10.43}$$

The sign of the Doppler frequency indicates whether the source is moving toward or away from the observer. In the example shown in Figure 10.4(a), $\cos\theta = -1$, so $f_T < f_O$, and $f_D < 0$. Thus, negative Doppler frequencies mean that the source is moving away from the observer. This is consistent with our impression that the pitch of an emergency vehicle's siren is lower as the vehicle moves away.

In medical ultrasound, the Doppler effect will be observed in pulse-echo mode, which requires that the above situation be extended. In pulse-echo mode, the transducer acts as both the source of the sound and the receiver of the Doppler-shifted echo returning from the object. Often this is done with a transducer that has two separate crystals located in the same housing, each performing one of these functions. The sound that returns to the transducer has been both *received* by the moving object and *retransmitted* by the moving object, thereby forming the echo that propagates back to the transducer. Thus, the object acts as both a moving receiver and a moving source, in order to create the echo. To fully understand what is happening, we therefore need to consider a second scenario in which there is a stationary source and a moving receiver;

in this case, the transducer and object have switched their transmit and receive roles with respect to the previous scenario.

Following the preceding discussion, we now consider the transducer **T** as a stationary source, transmitting an acoustic wave with frequency f_S, and the object **O** as a moving receiver, as shown in Figure 10.4(b). The wavelength of the source in the medium is $\lambda = c/f_S$; but as far as the moving observer (the object) is concerned, the *acoustic speed* of the medium appears to be different than c because the object encounters wave crests and troughs at a lower or faster rate than a stationary observer would. In particular, the speed of sound appears to be lower if the object moves away from the transducer and higher if the object moves toward the transducer. Using the same definition of θ as before, it can be shown that the moving object observes a frequency

$$f_O = \frac{c + v\cos\theta}{c} f_S \,. \tag{10.44}$$

In Figure 10.4(b), $\cos\theta = -1$, so $f_O < f_S$, which is consistent with our intuition that the observed frequency should be less than the source frequency when the source and observer are moving apart.

Applying the fundamental definition of the Doppler frequency—the difference between the observed frequency and the source frequency—the Doppler frequency in the present scenario is given by

$$f_D = f_O - f_S \,. \tag{10.45}$$

Using (10.44), we find

$$f_D = \left(\frac{v\cos\theta}{c}\right) f_S \,. \tag{10.46}$$

The Doppler frequency is negative in the case depicted in Figure 10.4(b), but if the object is moving toward the transducer, the Doppler frequency will be positive.

In pulse-echo mode, the echo received by **T** will be shifted by *both* the effects of a moving receiver and a moving source; this will yield essentially twice the Doppler frequency than in either case alone. To see this, we start by assuming that the transducer **T** generates an acoustic wave having frequency f_S. The object **O**, moving with velocity **v** at an angle θ relative to **u**, receives a wave having frequency f_O, as in (10.44). The object reflects or scatters this received wave acting like a moving source with frequency f_O, and this yields a receive transducer frequency given by (10.39). Substituting (10.44) into (10.39) yields

$$f_T = \frac{c + v\cos\theta}{c - v\cos\theta} f_S \tag{10.47a}$$

$$= \left(1 + \frac{2v\cos\theta}{c - v\cos\theta}\right) f_S \,, \tag{10.47b}$$

where the second equation follows after some algebra.

The Doppler frequency in pulse-echo mode is

$$f_D = f_T - f_S = \left(\frac{2v\cos\theta}{c - v\cos\theta} f_S \right) . \tag{10.48}$$

Since ordinarily $c \gg v$, the Doppler frequency in this case is well-approximated by

$$f_D = \frac{2v\cos\theta}{c} f_S . \tag{10.49}$$

Again, for the example depicted in Figure 10.4(c), $\cos\theta = -1$ and therefore $f_D < 0$. To a good approximation, the Doppler frequency in the case of a pulse-echo scenario is twice that of either a moving source or moving receiver situation. In addition, a negative Doppler frequency means that the object is moving away from the transducer.

If the angle θ of movement of the scatterers were known and not equal to 90°, then it is possible in principle to measure f_D and invert equation (10.49) in order to deduce the speed v of the scatterers. This is the principle of so-called Doppler-shift velocimeters. It is also possible to simply listen to $|f_D|$ as a function of time; this is the principle behind Doppler motion monitors, which are popular in fetal monitoring. Finally, it is also possible to measure f_D as a function of spatial position and to display its magnitude and sign in an image. This is the principle of Doppler imaging.

EXAMPLE 10.6

Suppose it is known that a 5-MHz transducer axis makes an angle of 30° relative to the direction of motion of blood in a vessel.

Question If the Doppler frequency is measured to be +500 Hz, what is the velocity of the blood? Is it moving toward or away from the transducer?

Answer Since $f_D > 0$, the blood is moving toward the transducer. The velocity is given by

$$\begin{aligned} v &= \frac{c f_D}{2 f_S \cos\theta} \\ &= \frac{+500 \text{ Hz} \times 1,540 \text{ m/s}}{2 \times 5 \times 10^6 \text{ Hz} \times \cos 30^\circ} \\ &= 0.0889 \text{ m/s} . \end{aligned}$$

■

10.5 Beam Pattern Formation and Focusing

In Section 10.3.5, we used a plane wave confined to a cylinder to study scattering. As it turns out, the acoustic energy of a real transducer is not confined to a cylinder but instead tends to spread out in a cone. The spatial distribution of the acoustic intensity of a transducer undergoing steady-state sinusoidal excitation is called the *field pattern* of the transducer.

In this section, we first study the field patterns created by the sinusoidal vibrations of a flat plate. This models flat transducers, which are elements of phased arrays (see Section 11.2.2), for example. We then study the field patterns created by focused transducers, which are composed of either curved plates or flat plates with lenses.

10.5.1 Simple Field Pattern Model

Treating the acoustic wave as if it were confined to the cylinder extended from the transducer's (flat) face is called the *geometric approximation*. It is only valid—and even then only approximately—very near to the face of the transducer. At a farther distance, the so-called *Fresnel approximation* holds and beyond that the *Fraunhofer approximation* applies. Before developing a mathematical treatment of these approximations, it is useful to begin with a very simple model that captures many of the characteristics of field patterns from a geometric point of view.

Suppose a transducer has a "diameter" of D and is pointing down the z-axis, as shown in Figure 10.5. Then, the *geometric region* begins at the transducer face and extends in range out to $D^2/4\lambda$ (circular transducer) or $D^2/2\lambda$ (square transducer). The *Fresnel region* extends beyond this point out to range D^2/λ; the *Fraunhofer region* is beyond this range. In the Fraunhofer region, also called the *far field*, the beam is spreading. At range z, the beamwidth is approximately $w = \lambda z/D$. To summarize these observations, the beamwidth $w(z)$ of a flat transducer at range z is approximately given by

$$w(z) = \begin{cases} D, & z \leq D^2/\lambda\,, \\ \lambda z/D, & z > D^2/\lambda \end{cases} . \tag{10.50}$$

This approximation ignores the "waist" occurring in the Fresnel region but is still very useful.

In the following section, we will develop the mathematics leading to the beamwidth formula in the far field or Fraunhofer region. As we will see, the width $\lambda z/D$ is approximately where the signal strength drops off by a factor of two from the peak strength, which is on the z-axis. We will also see that as we go further off the z-axis, the acoustic energy continues to drop until it hits zero, and then it rises again, then drops, and so forth. The acoustic energy confined between the first zeros on either side of the z-axis is called the *main lobe*. The "spokes" of energy confined between the other zeros are called *side lobes*. Most of our discussion is directed at the main lobe. The side lobes are ignored in developing imaging concepts, but when they have significant energy, they can be a source of artifacts.

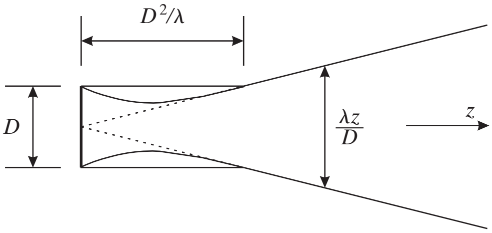


Figure 10.5 Simple field pattern geometry.

EXAMPLE 10.7

Consider a flat 1-cm transducer operating at 2 MHz in water.

Question What is the approximate beamwidth at a 5 cm range? What about at a 20 cm range?

Answer The wavelength in water at 2 MHz is $\lambda = c/f = 1,484 \text{ m/s}/2 \times 10^6 \text{ s}^{-1} = 0.742$ mm. The range at which the beam transitions from geometric to far field is $z = D^2/\lambda = (10 \text{ mm})^2/0.742 \text{ mm} = 134.8$ mm. Since 5 cm = 50 mm < 134.8 mm, the range of interest remains in the geometric region. Therefore, the beamwidth is approximately $w = D = 1$ cm. A range of 20 cm, however, exceeds the transition range. In this case, the beamwidth is approximately

$$\begin{aligned} w &= \frac{\lambda z}{D} \\ &= \frac{cz}{Df} \\ &= \frac{1,484 \text{ m/s} \times 0.2 \text{ m}}{2 \times 10^6 \text{ s}^{-1} \times 0.01 \text{ m}} \\ &= 1.48 \text{ cm}. \end{aligned}$$

■

10.5.2 Diffraction Formulation

The *diffraction formulation* is used to derive a more accurate model for the field pattern of a transducer comprising a vibrating flat plate. This formulation requires a series of simplifying acoustic approximations and also requires that we ignore the electronics, the electrical-mechanical interface to the transducer, and the mechanical-acoustical interface to the medium. Despite these simplifications, the diffraction formulation gives a reasonably accurate formula for the field pattern, especially in the far field, or Fraunhofer zone.

Narrowband Pulse It is useful to model the pressure signal of a transmitted pulse as a narrowband pulse, which can be written as

$$n(t) = \mathrm{Re}\{\tilde{n}(t)e^{-j2\pi f_0 t}\}, \tag{10.51}$$

where $\tilde{n}(t) = n_e(t)e^{j\phi}$ is the *complex envelope* and $n_e(t)$ is the *envelope*. The envelope of a narrowband pulse is the low-frequency (so-called baseband) signal that "rides" the crests of the narrowband pulse. A narrowband pulse and its envelope are shown in Figure 10.6. We require that the envelope is long with respect to the period $1/f_0$ in order that $n(t)$ be considered narrowband. This is true for the narrowband pulse depicted in Figure 10.6 and is generally a good approximation in medical ultrasound imaging systems, although the number of high-frequency signal periods within the envelope is usually much fewer (on the order of three to five, typically).

Our analysis is made easier by presuming that the input to the acoustical system is the *complex signal*

$$\mathrm{n}(t) = \tilde{n}(t)e^{-j2\pi f_0 t}. \tag{10.52}$$

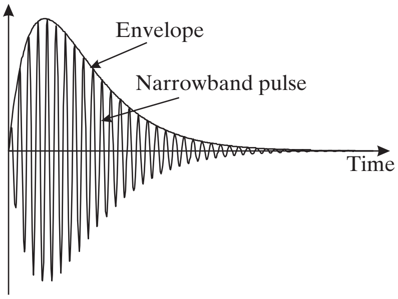


Figure 10.6 A narrowband pulse and its envelope.

With this model, the true input is given by

$$n(t) = \mathrm{Re}\{\mathrm{n}(t)\}\,, \tag{10.53}$$

and the envelope of the input is given by

$$n_e(t) = |\mathrm{n}(t)|\,, \tag{10.54}$$

where $|\cdot|$ gives the complex modulus. Since the acoustical system itself is a real physical system, the output must also be real. Therefore, the true output of the system driven by a complex input is just the real part of the complex output. Likewise, the envelope (baseband signal) of the output is just the complex modulus of the complex output. The primary reason that we use this complex formulation is that the output of an ultrasound system is typically the envelope, not the narrowband signal, and this formulation makes it easier to write the equations describing the envelope.

Received Signal with Field Pattern The geometry we consider is shown in Figure 10.7. The plate is assumed to be vibrating in the z direction, and each point $(x_0, y_0, 0)$ on the plate is assumed to produce an independent acoustic wave. The linear superposition of all these waves comprises the field pattern of the flat plate. Because the plate vibrates in the z direction, each point on the plate acts like an acoustic dipole rather than a monopole (spherical source). Accordingly, the acoustic wave produced by each point is not a spherical wave but is instead given by

$$p(x, y, z, t) = \frac{z}{r_0^2}\mathrm{n}(t - c^{-1}r_0)\,, \tag{10.55}$$

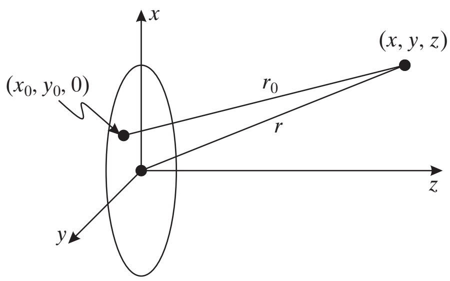


Figure 10.7 Geometry for field pattern analysis.

where $r_0 = \sqrt{(x-x_0)^2 + (y-y_0)^2 + z^2}$ is the distance between the point on the plate and (x, y, z), an arbitrary point in space. A complex narrowband excitation $\mathtt{n}(\cdot)$ is assumed; hence, the true signal is recovered by taking the real part, and the envelope is recovered by taking the modulus. In a direction parallel to the z-axis, this pressure wave has exactly the form of a spherical wave propagation pattern, but in an orthogonal direction, the pressure is zero (because $z = 0$).

The total pressure at (x, y, z) is the superposition of all (dipole) sources on the transducer face

$$p(x, y, z; t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x_0, y_0) \frac{z}{r_0^2} \mathtt{n}(t - c^{-1} r_0) dx_0 \, dy_0 \,,$$

where $s(x, y)$ is the transducer face *indicator function*, given by

$$s(x, y) = \begin{cases} 1, & (x, y) \text{ in face,} \\ 0, & \text{otherwise.} \end{cases} \tag{10.56}$$

Now suppose that a spherical wave scatterer of strength $R(x, y, z)$ is at (x, y, z). Then, a spherical wave will be sent back to the transducer, and the pressure at face point (x_0', y_0') will be

$$p_s(x_0', y_0'; t) = R(x, y, z) \frac{1}{r_0'} p(x, y, z; t - c^{-1} r_0') \,, \tag{10.57}$$

where r_0' is the distance from (x, y, z) to $(x_0', y_0', 0)$. Since the transducer is sensitive only to the dipole pattern, the response at each point on the face must be weighted by the dipole pattern. Accordingly, the received electrical (still modeled as complex) waveform due to a single scatterer at (x, y, z) is

$$\mathtt{r}(x, y, z; t) = K \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x_0', y_0') \frac{z}{r_0'} p_s(x_0', y_0'; t) dx_0' \, dy_0' \,, \tag{10.58}$$

where K is an arbitrary gain factor accounting for the transmit and receive sensitivity of the transducer and any preamplification hardware. Substituting all previous expressions gives

$$\begin{aligned} \mathtt{r}(x, y, z; t) = {} & KR(x, y, z) \\ & \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x_0', y_0') \frac{z}{r_0'^2} \\ & \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x_0, y_0) \frac{z}{r_0^2} \mathtt{n}(t - c^{-1} r_0 - c^{-1} r_0') dx_0 \, dy_0 \, dx_0' \, dy_0' \,. \end{aligned} \tag{10.59}$$

Plane Wave Approximation At this point, we use the first of a series of approximations. The *plane wave approximation* maintains that the envelope of the excitation pulse arrives at all points in a given z-plane simultaneously. This statement can be written mathematically as

$$\mathtt{n}(t - c^{-1} r_0 - c^{-1} r_0') \approx \tilde{n}(t - 2c^{-1} z) e^{-j2\pi f_0 (t - c^{-1} r_0 - c^{-1} r_0')} \,, \tag{10.60}$$

which can also be written as

$$\mathrm{n}(t - c^{-1}r_0 - c^{-1}r_0') \approx \mathrm{n}(t - 2c^{-1}z)e^{jk(r_0-z)}e^{jk(r_0'-z)} . \tag{10.61}$$

Using this approximation, the receive equation (10.59) simplifies so that the quadruple integral separates into two identical double integrals. Defining the *field pattern* as

$$q(x, y, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x_0, y_0)\frac{z}{r_0^2}e^{jk(r_0-z)}dx_0\,dy_0 , \tag{10.62}$$

the complex received signal for a single scatterer located at (x, y, z) is

$$\mathrm{r}(x, y, z; t) = KR(x, y, z)\mathrm{n}(t - 2c^{-1}z)[q(x, y, z)]^2 . \tag{10.63}$$

Assuming superposition holds, the total response for a distribution of scatterers is

$$\mathrm{r}(t) = \int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r(x, y, z; t)dx\,dy\,dz \tag{10.64}$$

$$= \int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} KR(x, y, z)\mathrm{n}(t - 2c^{-1}z)[q(x, y, z)]^2 dx\,dy\,dz . \tag{10.65}$$

For completeness, we include a round-trip attenuation factor, yielding the *basic pulse-echo signal equation*,

$$\mathrm{r}(t) = K\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x, y, z)\mathrm{n}(t - 2c^{-1}z)e^{-2\mu_a z}[q(x, y, z)]^2 dx\,dy\,dz . \tag{10.66}$$

The importance of this equation as the basis of an imaging equation will become clear in the next chapter. For now, it is important to continue examining the field pattern to see how it might be approximated in order to better understand it and to make this triple integral more tractable.

Paraxial Approximation The *paraxial approximation* assumes that we are primarily interested in the pattern near the transducer axis, in which case $r_0 \approx z$. This approximation can be applied only to the amplitude terms, not the phase terms. Accordingly, applying this approximation to (10.62) yields

$$q(x, y, z) \approx \frac{1}{z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x_0, y_0)e^{jk(r_0-z)}dx_0\,dy_0 . \tag{10.67}$$

Notice that this is the same result that we would get assuming that all the elements of the transducer acted as spherical wave generators and receivers instead of dipoles.

Fresnel Approximation The *Fresnel approximation* simplifies the phase term in (10.67) by first noting that

$$r_0 = \sqrt{(x - x_0)^2 + (y - y_0)^2 + z^2}$$

$$= z\sqrt{1 + \frac{(x - x_0)^2}{z^2} + \frac{(y - y_0)^2}{z^2}} . \tag{10.68}$$

Then, if z is large enough, two terms of the binomial expansion are sufficient to approximate the square root as

$$\begin{aligned} r_0 &\approx z\left[1+\frac{1}{2}\left(\frac{(x-x_0)^2}{z^2}+\frac{(y-y_0)^2}{z^2}\right)\right] \\ &\approx z+\frac{(x-x_0)^2}{2z}+\frac{(y-y_0)^2}{2z}\,. \end{aligned} \tag{10.69}$$

Applying this to the phase term in (10.67) yields

$$q(x,y,z) \approx \frac{1}{z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x_0,y_0)e^{jk\left(\frac{(x-x_0)^2}{2z}+\frac{(y-y_0)^2}{2z}\right)}\,dx_0\,dy_0\,. \tag{10.70}$$

This is in the form of a convolution in x and y; hence, we write the *Fresnel beam pattern* as

$$q(x,y,z) = \frac{1}{z}s(x,y) ** e^{jk(x^2+y^2)/2z}\,, \tag{10.71}$$

where $**$ denotes 2-D convolution [see (2.38) and (2.39)].

Substituting (10.71) into (10.66) yields

$$\begin{aligned} \mathtt{r}(t) = K&\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x,y,z)\mathtt{n}(t-2c^{-1}z)\frac{e^{-2\mu_a z}}{z^2} \\ &\times\left[s(x,y) ** e^{jk(x^2+y^2)/2z}\right]^2 dx\,dy\,dz\,. \end{aligned} \tag{10.72}$$

If $\mathtt{n}(t)$ is sufficiently short, both the attenuation and $1/z^2$ factors may be moved outside the integral by setting $z = ct/2$. This gives a received (complex) signal

$$\begin{aligned} \mathtt{r}(t) = K\frac{e^{-\mu_a ct}}{(ct)^2}&\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x,y,z)\mathtt{n}(t-2c^{-1}z) \\ &\times\left[s(x,y) ** e^{jk(x^2+y^2)/2z}\right]^2 dx\,dy\,dz\,, \end{aligned} \tag{10.73}$$

where the constant 4 has been absorbed by K.

EXAMPLE 10.8

An ultrasound transducer is placed in the x-y plane, pointing down in the z-axis. The transducer face is of dimension 1 mm × 10 mm, as shown in Figure 10.8. The transducer is working at 2 MHz and the speed of sound is 1,540 m/s.

Question At a range of 15 cm, where are the peaks of the first sidelobes along the x- and y-axes?

Answer The wavelength of the ultrasound wave is

$$\lambda = c/f = 1,540\,\text{m/s}/2\times 10^6\,\text{Hz} = 0.77\,\text{mm}\,.$$

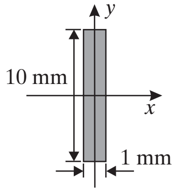


Figure 10.8
A transducer in the x-y plane, pointing in the z-axis.

For the ultrasound field in x-z plane, the transducer has width 1 mm, so the far field approximation holds for $z \geq D^2/\lambda = 1.2$ mm. For the ultrasound field in y-z plane, the transducer has width 10 mm, so the far field approximation holds for $z \geq D^2/\lambda = 129.8$ mm= 12.98 cm. So at range of 15 cm, the far field approximation holds for both directions. The widths of the main lobes along the x- and y-axes are

$$w_x = \lambda z/D_x = 11.55\,\text{cm}\,,$$
$$w_y = \lambda z/D_y = 1.16\,\text{cm}\,.$$

Therefore, the peaks of the first sidelobes along the x- and y-axes are approximately at

$$p_x = \pm 1.5 w_x = 17.32\,\text{cm}\,,$$
$$p_y = \pm 1.5 w_y = 1.73\,\text{cm}\,.$$

■

Fraunhofer Approximation The Fresnel approximation for r_0 in (10.69) can be expanded as

$$r_0 \approx z - \frac{xx_0}{z} - \frac{yy_0}{z} + \frac{x^2+y^2}{2z} + \frac{x_0^2+y_0^2}{2z}\,. \tag{10.74}$$

Using this form in (10.67) yields the following alternate expression for the Fresnel field pattern,

$$q(x,y,z) = \frac{1}{z}e^{jk(x^2+y^2)/2z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x_0,y_0)e^{jk(x_0^2+y_0^2)/2z}\, e^{-jk\left(\frac{x_0x}{z}+\frac{y_0y}{z}\right)}dx_0dy_0\,. \tag{10.75}$$

To see how to further approximate $q(x,y,z)$, we define

$$D = 2\sqrt{\max_{x_0,y_0\in\text{face}}(x_0^2+y_0^2)}\,, \tag{10.76}$$

which is essentially the diameter (or maximum lateral dimension) of the transducer. If $z \geq D^2/\lambda$, then the term $\exp[jk(x_0^2+y_0^2)/2z]$ is approximately 1. This approximation is called the *Fraunhofer approximation* and only applies in the far field or Fraunhofer region. Under this approximation, the double integral in (10.75) is revealed as the Fourier transform of $s(x,y)$, evaluated at the spatial frequencies $u = x/\lambda z$ and $v = y/\lambda z$. Therefore, the Fraunhofer approximation to the field pattern is written as

$$q(x,y,z) \approx \frac{1}{z}e^{jk(x^2+y^2)/2z}S\left(\frac{x}{\lambda z},\frac{y}{\lambda z}\right)\,, \quad z \geq D^2/\lambda\,, \tag{10.77}$$

where

$$S(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x,y)e^{-j2\pi(ux+vy)}dx\,dy\,. \tag{10.78}$$

Using (10.77) in (10.66), and making similar approximations as was done in the Fresnel approximation, yields

$$\begin{aligned} \mathtt{r}(t) = K\frac{e^{-\mu_a ct}}{(ct)^2}\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x,y,z)\mathtt{n}(t-2c^{-1}z)e^{jk(x^2+y^2)/z} \\ \times\left[S\left(\frac{x}{\lambda z},\frac{y}{\lambda z}\right)\right]^2 dx\,dy\,dz\,. \end{aligned} \tag{10.79}$$

The Fraunhofer approximation holds well for ranges greater than about D^2/λ, where D is the maximum lateral dimension of the transducer. For example, a 1-cm transducer operating at 2 MHz hits the far field at a range of about 13 cm. If the transducer is square, then the width of the beam is about $\lambda z/D$ in the far field. Therefore, at $z = D^2/\lambda$, the width of the field is D, equal to the width of the transducer. Doubling the range doubles the width of the beam, so that (in this example) at 26 cm, our beamwidth is 2 cm. Suppose we desire a lateral resolution of about 0.5 cm. Roughly, the geometric region holds until about D^2/λ—that is, up to the onset of the far field. Therefore, a transducer with $D = 0.5$ cm satisfies our criteria up to about 3 cm. Unfortunately, at 6 cm range, the beamwidth has increased to about 1 cm; at 12 cm range, the beamwidth has gone up to about 2 cm; and at 24 cm the beam width has gone to about 4 cm. Therefore, although the smaller transducer has better resolution up to its own far field than a larger transducer, the far field comes much sooner (i.e., at a distance much closer to the transducer), and the beam exceeds the dimensions of the larger transducer fairly quickly.

10.5.3 Focusing

Virtually all transducers used in medical ultrasound are focused to at least some degree. Here, *focusing* means to shape the beam into a narrower beam than is achieved using a flat vibrating plate. Focusing can be accomplished by manufacturing the transducer crystal in a curved shape, by applying a lens to a flat crystal, or by electronic focusing using multiple transducer crystals arranged either in a linear array or a set of concentric rings. Electronic focusing will be described in Chapter 11, as will the materials and construction of ultrasound transducers. In this section, we explore the physics of focusing using curved vibrators or lenses.

Consider the geometry shown in Figure 10.9. Here, the curved surface has radius d, and an acoustic pulse emitted simultaneously from the surface would arrive at the point $(0,0,d)$ simultaneously. The focal depth is therefore designed to be at the range $z = d$. In order to analyze the field pattern for this geometry, it is convenient to assume that the actual pulse (or waveform) is generated from a flat plate at $z = 0$ and that we can control the timing of the pulses as a function of position (x_0, y_0). Our goal is to generate the transmit waveform first at the farthest point $(x_m, y_m, 0)$ (or set of points) and last at the origin in such a way

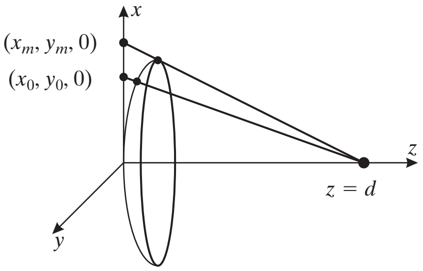


Figure 10.9 Geometry for the analysis of field patterns from focused transducers.

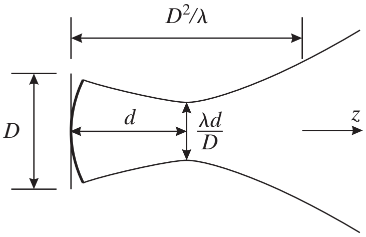


Figure 10.10 Approximate field pattern for a focused transducer.

that a round trip from the flat transducer to the focal point and back again takes the same amount of time across the entire face. This simulates the curved transducer as well as flat transducers with lenses, as we shall see.

It can be shown from the geometry (see Problem 10.16) that the time delay required for the pulse generated at point $(x_0, y_0, 0)$ is approximately

$$\tau \approx \frac{1}{dc}[r_m^2 - (x_0^2 + y_0^2)], \tag{10.80}$$

where $r_m^2 = x_m^2 + y_m^2$. Applying this time delay to the complex signal, and carrying out the plane wave and Fresnel approximations using analogous steps as above, yields the following Fresnel field pattern at the focal distance $z = d$:

$$q(x, y, d) = \frac{1}{d} e^{jk(x^2+y^2)/2d} S\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right). \tag{10.81}$$

The field pattern of (10.81) is identical to that of the Fraunhofer field for a flat transducer (10.77), except that z is replaced by d. In practical terms, this means that a relatively large transducer can produce a relatively narrow beam at the focal depth, as depicted in Figure 10.10. We will see in the next chapter that a narrower beam improves spatial resolution. The penalty for this improved resolution is that the beam diverges more rapidly than with an unfocused transducer. Therefore, although the resolution is better at or around the focal depth, it can be considerably worse at smaller and larger distances than with an unfocused transducer.

10.6 Summary and Key Concepts

Ultrasound imaging is based on the use of ultrasound, whose behavior is characterized by the 3-D wave equation. Ultrasound imaging is of interest because it is truly noninvasive, and the systems are typically quite portable.

In this chapter, we presented the following key concepts that you should now understand:

1. *Ultrasound* is sound with frequencies above 20 kHz.
2. Ultrasound systems generate and detect ultrasound via a *transducer*, which converts electrical signals to acoustic signals and vice versa.
3. Ultrasound images are based on the *reflection* and *scattering* of ultrasound by body tissues; the returning *echoes* are detected and displayed.
4. *Acoustic waves* are pressure waves that propagate through materials via *compression* and *expansion* of the material; these waves can have different geometries or patterns.
5. A *plane wave* is the most common wave pattern, which is represented by a simple *geometric approximation*, the *Fresnel approximation*, or the *Fraunhofer approximation*, depending on the distance from the transducer.
6. Plane waves are *reflected*, *refracted*, or *attenuated/transmitted* by/through *plane interfaces*.
7. Ultrasound waves can experience *nonlinear propagation* effects, which cause harmonic frequencies to develop as the wave propagates.
8. The *Doppler effect* is the change in frequency of sound due to the relative motion of the source and/or receiver.
9. The beam that is formed by a transducer can be *focused* to be narrower.

Further Reading

Cobbold, R.S.C. *Foundations of Biomedical Ultrasound*. New York, NY: Oxford University Press, USA, 2006.

Kremkau, F.W. *Diagnostic Ultrasound: Principles and Instruments*, 7th ed. Philadelphia, PA: W. B. Saunders, 2005.

Macovski, A. *Medical Imaging Systems*. Englewood Cliffs, NJ: Prentice Hall, 1983.

Pierce, A.D. *Acoustics: An Introduction to its Physical Principles and Applications*. New York, NY: McGraw-Hill, 1989.

Wells, P.N.T. *Biomedical Ultrasonics*. New York, NY: Academic Press, 1977.

Problems

The Wave Equation

10.1 Show that the function $w_1(z,t) = \xi(z - ct)$, where $\xi(\tau)$ is a twice differentiable function, is a solution of the one-dimensional wave equation (10.6). Moreover, show that the same is true for the function $w_2(z,t) = \xi(z - ct) + \xi(z + ct)$. What is the physical meaning of the wave function $w_2(z,t)$?

10.2 Show that the wavelength of a sinusoidal plane wave modeled by (10.8) is given by $\lambda = 2\pi/k$.

10.3 Consider the situation in Example 10.2 and let $\tau_1 = \tau_2 = 5\ \mu$s.

(a) At what time does the peak of the wave with maximum pressure hit the interface?

(b) After the wave hits the interface, a backward traveling wave will be generated. Write down an expression for the backward traveling wave.

(c) When will the peak of the backward traveling wave arrive at the transducer face?

10.4 Derive the spherical wave equation (10.13) for a source at the origin. Start with the full three-dimensional wave equation (10.4) and assume that the amplitude of particle displacement depends only on the radius r and on time t.

10.5 Show that the function $w(r,t) = \xi(r - ct)/r$, where $\xi(\tau)$ is a twice differentiable function, is a solution of the spherical wave equation (10.13).

10.6 Prove the validity of the general solution to the spherical wave equation

$$p(r,t) = \frac{1}{r} f(t - c^{-1}r) + \frac{1}{r} g(t + c^{-1}r),$$

by direct substitution into (10.13).

Wave Propagation

10.7 Prove that for a plane wave in which $v(0,t) = \mathrm{Re}\{Ve^{j\omega t}\}$ and $p(0,t) = \mathrm{Re}\{Pe^{j\omega t}\}$, that the average power in the acoustic wave at $x = 0$ is $I_{\mathrm{av}} = (1/2)\mathrm{Re}\{VP^*\}$ where $*$ denotes complex conjugate.

10.8 (a) Solve Equations (10.25) and (10.26) to find the expressions for pressure reflectivity R and pressure transmittivity T given by Equations (10.27) and (10.28), respectively.

(b) Derive the expressions for intensity transmittivity and reflectivity given in Equations (10.29) and (10.30), respectively.

10.9 An acoustic dipole can be modeled by two point sources (monopoles) residing close together but vibrating exactly out of phase. Suppose point source 1 is located at $(0,0,-d)$ and point source 2 is located at $(0,0,d)$, where d is very small. Further suppose that source 1 is generating the narrow-band signal $f(t)$, leading to a spherical pressure wave given by $p(r_1,t) = f(t - c^{-1}r_1)/r_1$ and source 2 is generating the narrow-band signal $-f(t)$, leading to a spherical pressure wave given by $p(r_2,t) = -f(t - c^{-1}r_2)/r_2$, where the two radii r_1 and r_2 are measured from their respective sources.

(a) Show that the measured pressure in the far field is approximately $p(r,t) = zf(t - c^{-1}r)/r^2$.

(b) Sketch isopressure lines in the x-z plane for the above field pattern.

10.10 Consider an outward propagating spherical wave.

(a) Write down a general expression of an attenuated spherical wave whose source is located at the origin.

(b) If there is a small target located at (x,y,z), derive an expression for the scattered wave.

(c) Generalize the above result to a spherical wave generated at a arbitrary location (x_0,y_0,z_0).

10.11 (**a**) Derive the intensity reflectivity and transmittivity from the pressure reflectivity and transmittivity.

(**b**) Show that $T - R = 1$.

(**c**) Explain why $T_I \neq 1 - R_I$ in general and find the correct relationship between T_I and R_I.

Doppler Effect

10.12 Consider a stationary source that generates plane wave with frequency f_0. Assuming that a receiver is moving towards the source with speed v and that the speed of sound in the medium is c, prove that the frequency observed is $f_R = (c + v)f_0/c$. What if the receiver moves away from the source?

10.13 Using the above derivation, discuss the Doppler effects in the following situations when the receiver moves towards and away from the source:

(**a**) $c > v$.

(**b**) $c = v$.

(**c**) $c < v$.

Ultrasound Field Pattern

10.14 Suppose the Fourier transform of a real signal $n_e(t)$ is $N_e(\varphi)$. Find the Fourier transform of $\tilde{n}(t) = n_e(t)e^{j\phi}$ and $n(t) = \mathrm{Re}\{\tilde{n}(t)e^{-j2\pi f_0 t}\}$.

10.15 An ultrasound imaging system is equipped with two square transducers. One operates at 5 MHz, and the other one operates at 12 MHz. The 5-MHz transducer is 2.0 cm by 2.0 cm and the 12-MHz one is 0.4 cm by 0.4 cm. The imaging system is tested in a medium having a speed of sound of 1,560 m/s at both frequencies.

(**a**) What are the absorption coefficients $\alpha_{5\text{ MHz}}$ and $\alpha_{12\text{ MHz}}$ of the medium at the frequencies of the transducers?

(**b**) What ranges are considered as far field for each transducer?

10.16 A transducer can be focused by curving the crystal to conform to the shape of a sphere, as shown in Figure P10.1. To analyze the field pattern, however, it is convenient to assume that the transducer is flat but that we can independently control the time delay of each element on its face. Assume that the point $(x_m, y_m, 0)$ is the point farthest out on the flat face.

(**a**) Show that the time delay for a point $(x_0, y_0, 0)$ on the flat face is approximately

$$\tau \approx \frac{1}{dc}[r_m^2 - (x_0^2 + y_0^2)],$$

where $r_m^2 = x_m^2 + y_m^2$ and d is the desired focal depth.

(**b**) Use the steady-state approximation and show that at $z = d$ the Fresnel field pattern is

$$q(x, y, d) = \frac{1}{d} e^{jk(x^2+y^2)/2d} S\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right),$$

where $S(u, v)$ is the 2-D Fourier transform of $s(x, y)$.

(**c**) Discuss the merits of focusing.

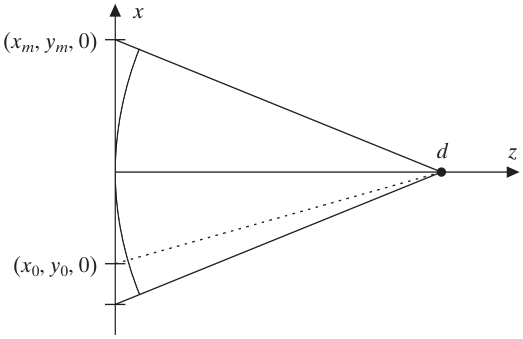


Figure P10.1 Geometry of curved transducer for Problem 10.16.

10.17 A portion of a linear array is shown in Figure P10.2. Note that $h > w$. The array is composed of five transducers (called elements), whose faces are contained in the x-y plane and vibrate in the z direction. Assume the far-field (Fraunhofer) approximation in all of the parts below.

(**a**) What is the far-field pattern $q_0(x, y, z)$ of the central element taken by itself?

(**b**) Find the range z_0 at which the first zeros of this pattern (in the x direction) coincide with $x = -s$ and $x = +s$?

(**c**) What condition must the separation s satisfy in order that z_0 is in the far-field of the central element?

(**d**) Find the far-field pattern $q(x, y, z)$ of the five elements operating in unison.

(**e**) What is the beamwidth (out to the first zeros) of this pattern at range z_0 in both the x and y directions. (Think carefully about this one before you proceed blindly ahead; it is easier than it appears at first glance.)

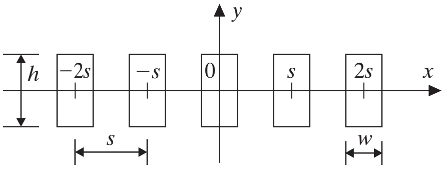


Figure P10.2 A portion of a linear array for Problem 10.17.

CHAPTER

11

Ultrasound Imaging Systems

11.1 Introduction

In this chapter, we study the principles of ultrasound imaging systems. After projection radiography systems, these are the most widely used medical imaging systems in the world. In some hospitals, ultrasound imaging accounts for one-third of all imaging procedures. In part, this is because at diagnostic intensities, ultrasound poses no known risk to the patient. Another factor is that these are among the least expensive medical imaging systems and are portable and therefore easily moved from bedside to operating room. A typical imaging system is shown in Figure 11.1.

11.2 Instrumentation

A block diagram of a typical ultrasound imaging system is shown in Figure 11.2. The vast majority of medical ultrasound imaging systems use the same transducer for both generation and reception of ultrasound; this is the so-called *pulse-echo mode* of operation. The transducer is coupled to the body using an "acoustic gel," and a brief pulse-like acoustic wave is generated. This wave propagates into the body, where it encounters reflecting surfaces and small scatterers. These objects reflect or scatter the sound, a part of which returns to the transducer. The transducer then converts the acoustic wave sensed at its face to an electrical signal that can be amplified, stored, and displayed. In this section, we will describe each of the components required to carry out this basic imaging paradigm. This sets the stage for the development of an imaging equation, the analysis of image quality, and an introduction to pulsed Doppler and three-dimensional imaging systems in subsequent sections.

11.2.1 Ultrasound Transducer

Transducer Materials Medical transducers use *piezoelectric crystals* to both generate and receive ultrasound. These crystals have the property that an induced electric field produces a *strain* (mechanical displacement), which in turn causes

Figure 11.1
A typical ultrasound imaging system. (Courtesy of GE Healthcare.)

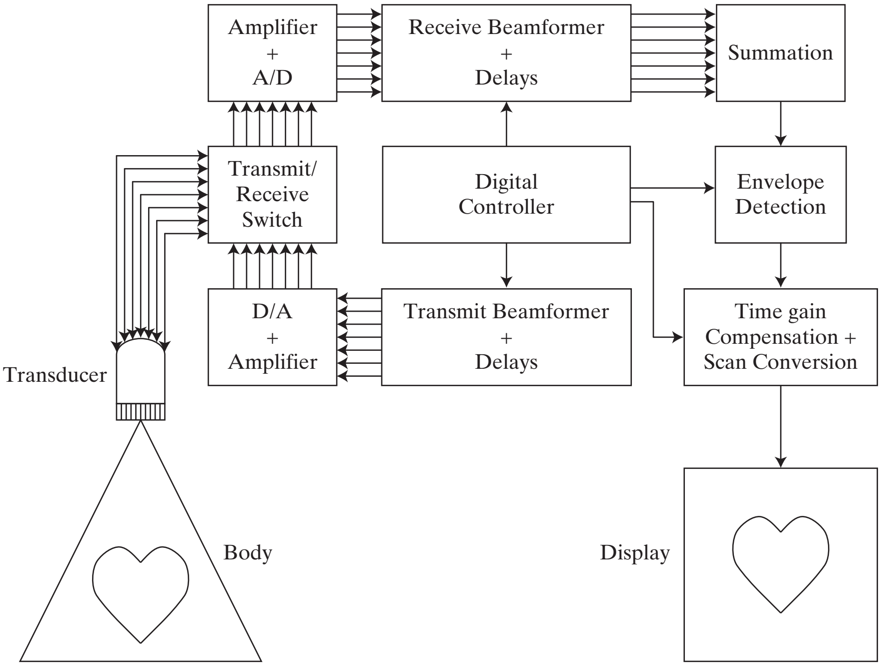


Figure 11.2
Block diagram of an ultrasound imaging system.

an acoustic wave. They also satisfy the reciprocal property that a mechanical displacement creates an electric potential, which means that they can also sense an acoustic wave. These concepts are illustrated in Figure 11.3.

Lead zirconate titanate, or PZT, is the piezoelectric material used in nearly all medical ultrasound transducers. It is a ceramic ferroelectric crystal exhibiting a strong piezoelectric effect. These crystals can be manufactured in nearly any

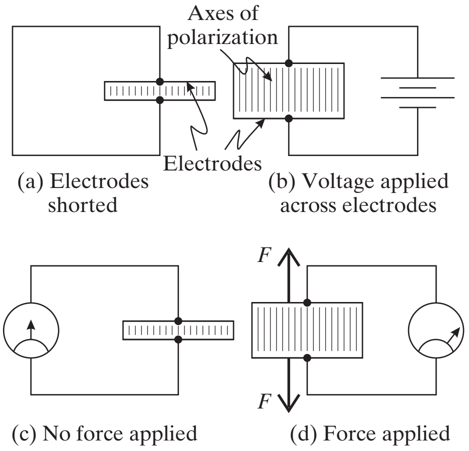


Figure 11.3 Piezoelectric behavior of an ultrasound transducer crystal.

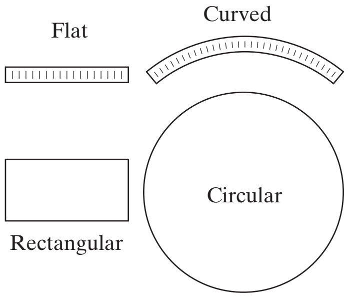


Figure 11.4 Shapes and axial orientations for PZT.

shape, and their axes of polarization (see Figure 11.3) can be oriented in nearly any direction, as shown in Figure 11.4. The most common transducer shapes are the circle, for single-crystal transducer assemblies found in Doppler monitoring systems, and the rectangle, for multiple-transducer assemblies such as those found in linear and phased arrays, as described later.

An ideal transducer material, when used in pulse-echo mode, should be both an efficient producer and sensitive receiver of ultrasound waves. The *transmitting constant d* of a transducer is the strain produced by a unit electric field and has units of meters per volt. The *receiving constant g* is the potential produced by a unit stress and has units of volt-meters per Newton. For PZT, $d = 300 \times 10^{-12}$ m/V and $g = 2.5 \times 10^{-2}$ V/(N/m). By comparison, for quartz (a natural piezoelectric material), $d = 2.3 \times 10^{-12}$ m/V and $g = 5.8 \times 10^{-2}$ V/(N/m). This implies that quartz is two orders of magnitude less efficient in production of ultrasound as PZT but comparable in its reception efficiency. Polyvinylidene fluoride (PVDF), a polymer film, also has piezoelectric properties and is commonly used as a probe when measuring the acoustic properties of other transducers. For PVDF, $d = 15 \times 10^{-12}$ m/V and $g = 14 \times 10^{-2}$ V/(N/m). Clearly, PVDF has poor efficiency as a transmitter of ultrasound but is much more efficient on reception than either PZT or quartz.

Resonance Transducers Transducer crystals exhibit *resonance*—that is, they tend to vibrate sinusoidally after electrical excitation has ended. The frequency of this vibration is called the *fundamental resonant frequency* f_T of a transducer.

In most systems, different transducers having resonant frequencies in the range 1–20 MHz can be "plugged" into the ultrasound scanner. The difference in resonant frequency between the different transducers has a profound influence on image quality, as we shall see.

The resonant frequency of a transducer is largely determined by the thickness of its piezoelectric crystal. When the front face of the transducer is moved forward (by the electrically induced strain, for example), an acoustic wave is initiated forward into the medium (perhaps a human body) and backward into the transducer crystal itself. The backward-traveling wave will hit the back face of the transducer and reflect toward the front face again. Resonance is set up when the returning wave strikes the front face at the time when the wave reaches a "crest," thus reinforcing the wave. This condition occurs when

$$f_T = \frac{c_T}{2d_T}, \tag{11.1}$$

where c_T and d_T are the speed of sound in the transducer and the thickness of the transducer, respectively. Since $\lambda_T = c_T/f_T$, the resonance condition is equivalently given by

$$\lambda_T = 2d_T .$$

EXAMPLE 11.1

The thickness of the piezoelectric crystal determines the resonant frequency of a transducer.

Question Consider a transducer made of a PZT crystal, which has a speed of sound of $c_T = 8,000$ m/s. If we want the transducer to work at a frequency of 10 MHz, what should the thickness of the crystal be?

Answer The relation between the thickness of the crystal and the resonant frequency is given by

$$f_T = \frac{c_T}{2d_T},$$

where $f_T = 10$ MHz and $c_T = 8,000$ m/s. So the thickness of the crystal should be

$$d_T = \frac{c_T}{2f_T} = 0.4\,\text{mm}.$$

■

Damping and Matching The piezoelectric crystal comprising a resonance transducer is plated with conductive electrodes on each side and mounted within an assembly called a *probe*, as illustrated in Figure 11.5. The transducer is usually *shock excited*, meaning the electrical signal resembles an impulse—that is, it has a very large peak voltage and a very short duration. Once excited, the transducer will resonate until the in-transducer wave loses energy, which causes a damping of the acoustic wave. In-transducer acoustic energy is lost in three ways: to the body, out the back of the transducer, and due to absorption within the crystal. The reflection coefficient between PZT and tissue is large,

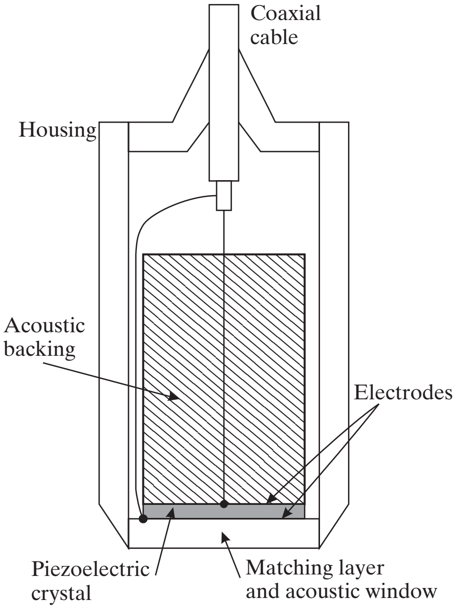


Figure 11.5
A simple ultrasound probe.

and the absorption coefficient of PZT is small. Therefore, to remove excessive vibrations, the transducer is backed with an epoxy material with an impedance that is nearly matched to PZT but with a high absorption coefficient. This permits acoustic energy to be transmitted out the back of the crystal, thus *damping* the remaining resonant acoustic wave inside the crystal.

Acoustic energy must also be permitted to exit through the front of the tissue and into the body. Since the acoustic impedance of tissue is about 20 times smaller than that of the piezoelectric crystal, there is a very large reflectivity at the crystal/tissue interface. Unless aided in some way, only a small fraction of the acoustic energy would escape into the tissue. It turns out that a thin layer of material, called a *matching layer*, between the crystal and the tissue will help transmit the pulse. The material should have a thickness that is one-quarter of the wavelength of the resonant pulse (as measured in the matching layer) and an acoustic impedance Z_m that is the geometric mean of that of the crystal Z_c and the tissue Z_t, that is, $Z_m = \sqrt{Z_c Z_t}$. The reader should be reminded that the matching layer must be coupled to the skin so that there is no air between the two, which is the reason that acoustic gel is used when applying the probe to the body.

In the case of a typical resonance transducer, all factors (i.e., backing and matching layer) are controlled so that the outgoing acoustic pulse has a well-defined center frequency but damps out after about three to five cycles. A typical transmit pulse is shown in Figure 11.6.

Broadband Transducers Some imaging scenarios require a shorter pulse, such as one that is only one half of a cycle of a sinusoidal pulse. In other cases, it might be necessary to use the same transducer to generate one pulse at 3 MHz and the next at 5 MHz. These demands cannot be met by the typical resonant transducer since it is designed to generate a narrowband pulse (meaning that is has a narrow range of frequencies) at a specific center frequency. In these cases, a *broadband* transducer is required.

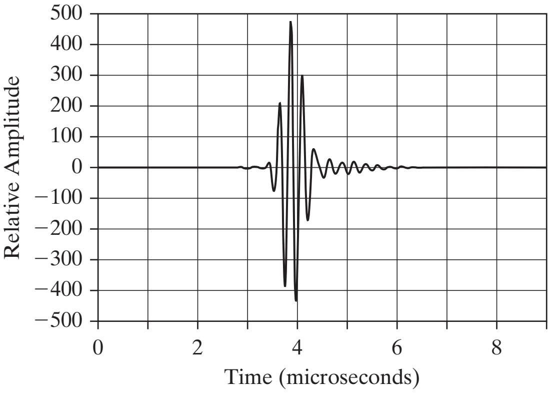


Figure 11.6 Typical transmit pulse from a resonance ultrasound transducer.

One way to build a broadband transducer is to simply increase the damping on the back of the crystal. In this case, a shock excitation would generate an initial pulse that would be quickly dissipated; what exits out the front and into the tissue is then a shorter pulse, which has a correspondingly higher bandwidth. A typical transducer in commercial use today will have somewhere between 50% and 100% bandwidth as measured from its center frequency. For example, a 3-MHz transducer with a 100% bandwidth would excite the tissue in the frequency range 1.5–4.5 MHz under shock excitation.

Any transducer with greater than 80% bandwidth is considered to be a broadband transducer. There are other more efficient designs for broadband transducers than the high damping strategy. For example, one design uses many small piezoelectric crystals embedded like "plugs" within a damping material. In this way, when all the crystals are excited simultaneously they will produce a coherent wave while being individually damped by the surrounding material. These transducers do not need a matching layer or a special additional acoustic backing, and are therefore more efficient.

Broadband transducers do not need to be shock excited but instead can be "driven" by an arbitrary waveform. In particular, if excited by any pulse that is within the bandwidth of the transducer, that pulse will be generated as an acoustic wave. For example, a 5-MHz transducer with a 100% bandwidth could in principle serve as a replacement for any narrowband transducer in the range 2.5–7.5 MHz by exciting the transducer with short sinusoidal pulses (like those produced by a resonant transducer). These transducers can also generate a sinusoidal pulse at one frequency and receive the harmonics that are produced by nonlinear propagation. These are therefore the type of transducers that are used for harmonic ultrasound imaging (see Section 11.8.1).

11.2.2 Ultrasound Probes

Single-Element Probes The simplest transducer assembly, also called an *ultrasound probe*, is shown in Figure 11.5. This single-element assembly has a field pattern that is well-modeled using the vibrating plate model developed in Section 10.5.2. This basic design accommodates either a lens or a curved crystal for focusing. For imaging, the ultrasound beam must be *steered* (also called

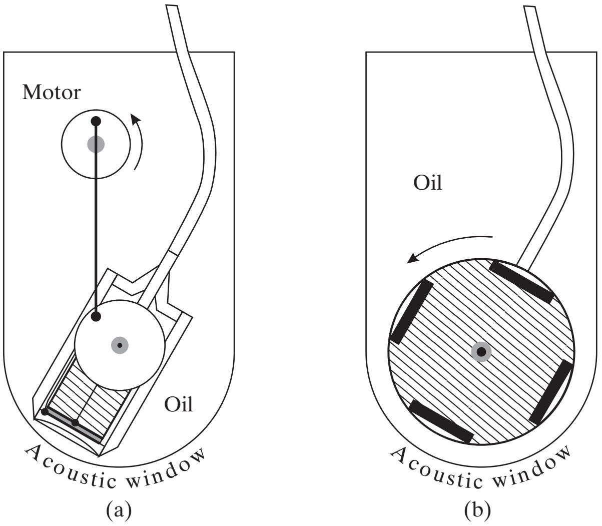


Figure 11.7
Mechanical scanner probe designs.

scanned or *swept*) within the body, typically within a plane. Early ultrasound systems used hand-held single-element probes that were manually scanned within a body cross section. Modern imaging systems, however, use either mechanical or electrical means to scan the beam. This is what permits real-time imaging.

Mechanical Scanners Mechanical scanning is accomplished by rocking or rotating a transducer crystal or set of crystals within the transducer assembly, as shown in Figure 11.7. In these designs, a transducer crystal is rapidly pulsed as it moves through the sector being imaged. Each acoustic pulse goes through an acoustic window and then propagates according to the speed of sound and the transducer's field pattern, as studied in Chapter 10. The echoes from any one scan line are received and processed before the transducer is pulsed again. In the *rocker* (or *wobbler*) design, shown in Figure 11.7(a), the transducer continues to travel through the same sector in a repeating fashion first clockwise, then counterclockwise, and so forth. In the *rotating* design shown in Figure 11.7(b), a new transducer is "switched in" as it enters the sector, so the scan sequence is always counterclockwise. Regardless of the specific design, the field of view for mechanical scanners is always shaped like a slice of pie, and called a *sector*.

Electronic Scanners Transducer assemblies having multiple elements can be electronically scanned in order to sweep the field of view. For two-dimensional (2-D) imaging, the arrangement of the elements in these assemblies is linear, as shown in Figure 11.8. Each element is rectangular and is focused in the longer dimension of the element using a lens (or curved elements). The width of each element and the attached electronics determine the manner in which the array is used to image the body. In particular, when the elements have widths on the order of a wavelength and are simply electronically grouped together making several elements appear as one, then the array is called a *linear array probe*. On the other hand, when the elements have widths on the order of a quarter wavelength and the timing of the firing of the elements are electronically controlled in order to steer and focus the beam, then the array is called a *phased array probe*.

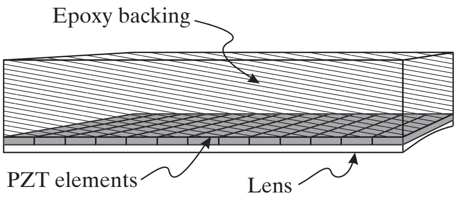


Figure 11.8 Basic arrangement of transducer elements in linear and phased arrays.

EXAMPLE 11.2

Consider a mechanical scanner with rotating design as shown in Figure 11.7(b), and suppose that the "pie-shaped" field of view (i.e., the sector) has an angle of 90° at the transducer.

Question Suppose N pulse-echo experiments are to be acquired for each transducer element over the 90° window. If no echo is received from further than 15 cm in range, and the speed of sound is $c = 1,540$ m/s, what is the maximum rate of revolution of the transducer?

Answer At a range of 15 cm, it takes

$$T = 2 \times \frac{15\,\text{cm}}{1,540\,\text{m/s}} = 195\,\mu\text{s}$$

for the transducer to receive the echo. During this period of time, the scanner probe cannot rotate more than 90°, otherwise the transducer leaves the window before it receives the echo. So the maximum rate of revolution is

$$r = \frac{1/4}{T} = 1,283\,\text{revolutions per second.}$$

■

11.3 Pulse-Echo Imaging

In the following sections, we develop a mathematical framework for basic ultrasound imaging. We rely on simplifying assumptions throughout, generally making bold assumptions at the outset and then refining them as our understanding grows. In subsequent sections, we will explore three very important augmentations to basic imaging: phased arrays (and electronic focusing), Doppler imaging, and 3-D imaging.

11.3.1 The Pulse-Echo Equation

Analysis of ultrasound imaging systems begins with the pulse-echo received signal equation developed in Chapter 10. We derived (10.73) using the Fresnel approximation and (10.79) using the Fraunhofer approximation. Either equation can be written more generally as

$$\mathtt{r}(t) = K\frac{e^{-\mu_a ct}}{(ct)^2}\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x,y,z)\mathtt{n}(t-2c^{-1}z)\tilde{q}^2(x,y,z)\,dx\,dy\,dz\,, \quad (11.2)$$

where

$$\tilde{q}(x,y,z) = zq(x,y,z)\,. \quad (11.3)$$

Here, $q(x, y, z)$ is the transducer field pattern. Because the input waveform in (11.2) is complex, the actual received signal is the real part of $\mathtt{r}(t)$. The envelope (which is usually desired instead of the high-frequency signal itself) is found by taking the complex modulus of $\mathtt{r}(t)$.

The time-dependent terms outside the integral in (11.2) cause very severe signal loss if not compensated. All systems are equipped with circuitry that performs *time-gain compensation* (TGC), which is a time-varying amplification, as shown in Figure 11.2. Nominally, the default gain is set to cancel the gain terms appearing in (11.2), which requires a gain of

$$g(t) = \frac{(ct)^2 e^{\mu_a ct}}{K} . \tag{11.4}$$

In practice, most systems have additional (range dependent) slide potentiometers, which allow the gain to be determined interactively by the operator. This permits the user to manually adapt the system to special circumstances requiring either more or less gain so that subtle features can be seen in the images. In our development here, we assume the nominal gain compensation, which leads to the gain-compensated (complex) signal

$$\mathtt{r}_c(t) = g(t)\mathtt{r}(t) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x, y, z)\mathrm{n}(t - 2c^{-1}z)\tilde{q}^2(x, y, z)\, dx\, dy\, dz\, . \tag{11.5}$$

The actual gain-compensated received signal is the real part of $\mathtt{r}_c(t)$ and its envelope is the complex modulus of $\mathtt{r}_c(t)$.

EXAMPLE 11.3

Suppose a system with nominal gain of 80 dB is working at a frequency of $f = 5$ MHz. The speed of sound is $c = 1,540$ m/s, and the coefficients for attenuation are $a = 1$ dB cm^{-1} MHz^{-1}, and $b = 1$.

Question What is the nominal TGC?

Answer The attenuation coefficient is

$$\alpha = af^b = 5 \text{ dB cm}^{-1}.$$

The amplitude attenuation factor μ_a is

$$\mu_a = \alpha/8.7 = 0.575 \text{ Np cm}^{-1}.$$

The nominal gain of the system is 80 dB. So the amplitude gain of the system is $K = 10,000$. Based on this information, the required TGC is

$$g(t) = \frac{(ct)^2 e^{+\mu_a ct}}{K} = 2.37 \times 10^6 t^2 e^{88,550t}.$$

The TGC plotted in log scale is shown in Figure 11.9. ■

Suppose our transducer has a resonant frequency of 5 MHz; then the transmitted signal has most of its spectral energy concentrated around 5 MHz.

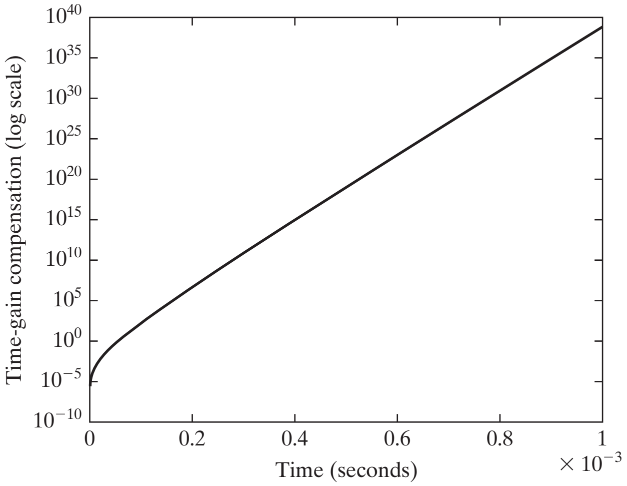


Figure 11.9
Time-gain compensation.

For example, a typical system might have around a 60 percent fractional bandwidth, meaning that a nominal 5 MHz system uses frequencies in the range 3.5–6.5 MHz. Because the system (transducer, body, amplification electronics) is modeled as linear, the received signal will have its spectral energy concentrated around 5 MHz. For imaging purposes, however, it is the signal strength of the returning echoes that is important, and the strength of a narrowband signal is captured by its envelope. Therefore, the first step in an ultrasound imaging system after amplification is envelope detection or demodulation.

Inexpensive ultrasound systems use a simple envelope detection procedure given by the two-step procedure depicted in Figure 11.10. The absolute value of the received signal is accomplished using a rectifier. The low-pass filter can be accomplished using a simple R-C circuit. Electrical engineers will recognize this as a simple AM demodulation scheme. Mathematically, it is simpler to model the envelope detection phase using the received complex signal. Since the received complex signal is a narrowband signal, it can be modeled as

$$\mathtt{r}(t) = r_e(t)e^{j\phi}e^{-j2\pi f_0 t} .$$

Mathematically, the envelope of this signal is extracted by taking its modulus (complex magnitude). Cascading gain compensation and envelope detection yields the (real) signal

$$e_c(t) = \left| \int_0^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R(x,y,z)\mathrm{n}(t - 2c^{-1}z)\tilde{q}^2(x,y,z)\, dx\, dy\, dz \right| . \tag{11.6}$$

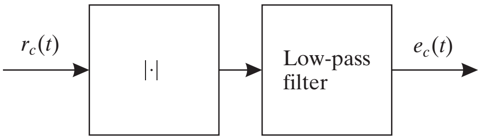


Figure 11.10
Envelope detection.

Equation (11.6) can be further simplified by recalling from Section 10.5.2 that n is a narrowband signal, which can be written as [see (10.51)]

$$\mathrm{n}(t) = n_e(t)e^{j\phi}e^{-j2\pi f_0 t}\,. \tag{11.7}$$

Substitution of (11.7) into (11.6) yields

$$\begin{aligned} e_c(t) = \bigg| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty & R(x,y,z) n_e(t - 2c^{-1}z) \\ & \times e^{j\phi} e^{-j2\pi f_0 (t-2c^{-1}z)} \tilde{q}^2(x,y,z)\, dx\, dy\, dz \bigg| \,. \end{aligned} \tag{11.8}$$

Phase terms that do not depend on x, y, or z disappear because of the modulus operator, leaving

$$e_c(t) = \left| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x,y,z) n_e(t-2c^{-1}z) e^{j2\pi f_0 2c^{-1}z} \tilde{q}^2(x,y,z)\, dx\, dy\, dz \right| \,. \tag{11.9}$$

We use the fact that $k = 2\pi f_0/c$ to get

$$e_c(t) = \left| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x,y,z) n_e(t-2c^{-1}z) e^{j2kz} \tilde{q}^2(x,y,z)\, dx\, dy\, dz \right| \,. \tag{11.10}$$

We will see in Section 11.5 that the signal $e_c(t)$, called the *A-mode signal*, is the fundamental signal in all of ultrasound imaging. It is used directly in A-mode, M-mode, and B-mode imaging.

11.4 Transducer Motion

So far in our analysis, our transducer has remained motionless, pointed down the z-axis. To acquire images, however, the transducer must move. In this section, we analyze the imaging equations that result when our transducer is allowed to move in the (x, y) plane.

Imagine moving the transducer in the $z = 0$ plane. The object does not move, but the transducer's field pattern does. This is like shining a flashlight on a different part of a visual scene, but since ultrasound imaging is a pulse-echo modality the transducer movement not only insonifies a different part of the body, but it also acquires echoes only from the part that insonified. Accordingly, with the transducer placed at $(x_0, y_0, 0)$ (still pointing parallel to the z-axis), the A-mode signal is

$$\begin{aligned} e_c(t; x_0, y_0) = \bigg| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty & R(x,y,z) n_e(t-2c^{-1}z) \\ & \times e^{j2kz} \tilde{q}^2(x-x_0, y-y_0, z)\, dx\, dy\, dz \bigg| \,. \end{aligned} \tag{11.11}$$

Since the envelope of the transmit pulse is fairly brief, it is nearly correct to say that a signal returning at time t must have originated at range

$$z_0 = ct/2\,, \tag{11.12}$$

where c is the speed of sound in the body. Notice that the "2" is needed to account for the round-trip delay (to get to range z and back). Equation (11.12) is known as the *range equation* and is fundamental in relating range to time in pulse-echo imaging.

Since the range equation relates z_0 and t, e_c can be thought of as a function of x_0, y_0, and z_0 instead of x_0, y_0, and t. Therefore, e_c can be thought of as an estimate of the reflectivity function as a function of spatial position. We formalize this notion by setting

$$\hat{R}(x_0, y_0, z_0) = e_c(2z_0/c; x_0, y_0)\,. \tag{11.13}$$

Combining (11.11) and (11.13) yields

$$\begin{aligned}&\hat{R}(x_0, y_0, z_0) \\ &= \left| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x,y,z)e^{j2kz} n_e(2(z_0-z)/c)\tilde{q}^2(x-x_0, y-y_0, z)\,dx\,dy\,dz \right| .\end{aligned} \tag{11.14}$$

The Geometric Assumption To get some insight into (11.14), let us make the drastic assumption that

$$\tilde{q}(x,y,z) = s(x,y)\,. \tag{11.15}$$

This is equivalent to assuming that the transducer energy travels down a cylinder having the same shape as its face. In this case, (11.14) reduces to

$$\hat{R}(x,y,z) = K \left| R(x,y,z)e^{j2kz} * \tilde{s}(x,y) n_e\left(\frac{z}{c/2}\right) \right| , \tag{11.16}$$

where $\tilde{s}(x,y) = s(-x,-y)$ and $*$ is 3-D convolution.

If we ignore the term e^{j2kz} in (11.16), then all terms are real, and the modulus bars are not needed. In this case, we see that the estimated reflectivity is the true reflectivity convolved with a blurring function, or impulse response, given by the product of the face shape and the envelope shape. This 3-D shape is known as the *resolution cell* of an ultrasound imaging system. It might seem that better estimates of R could be obtained if the pulse were very short and the face were very small—that is, if the resolution cell were small. Unfortunately, in reality, if the face gets small, the beam actually spreads (as discussed in Section 10.5.2); this is the problem with the geometric approximation. Also, if the pulse shortens, it is no longer narrowband, and problems with frequency-dependent attenuation arise. As always, there are compromises between these various factors that lead to well-balanced ultrasound systems.

Now we turn to the factor e^{j2kz}, which we ignored in the above discussion. It turns out that this term is the cause of *speckle*, the dominant artifact appearing in ultrasound images. To explain speckle, we note that the phase term e^{j2kz} multiplies the true reflectivity distribution $R(x,y,z)$. Since the resolution cell is far larger than π/k, e^{j2kz} can be expected to have many periodic cycles within a resolution cell. At any given time, a scatterer within the resolution cell will have a specific phase according to its z position. This phase can be modeled as a random variable, uniform over $[0, 2\pi)$. Integration over the entire resolution cell adds a large collection of complex numbers $R(x,y,z)e^{j2kz}$, where the phase of each is random. The modulus of this integral is a Rayleigh random variable, whose mean is dependent on the underlying reflectivity distribution.

Since resolution cells overlap in time [which is equivalent to range via (11.12)] and generally in azimuth (lateral position) in an image, the received signals are correlated; therefore, the Rayleigh random variables produce a spatially correlated random pattern, which multiplies the underlying desired image of reflectivity. This pattern, called a *speckle pattern*, appears as a splotchy granulation, oscillating from bright to dark across the image, with granule size that depends on the size of the resolution cell, the spatial distribution of scatterers, and the frequency of the pulse. Speckle is an undesirable, but generally dominant, feature in all medical ultrasound images.

Fresnel and Fraunhofer Approximations In order to further simplify (11.14) (but without using the drastic geometric field pattern approximation), we can replace $\tilde{q}$ with an expression for the actual transducer field pattern. Here, we consider the Fresnel and Fraunhofer approximations, two approximations that were developed in Section 10.5.2.

The Fresnel approximation for $\tilde{q}$ is found from (10.71) and (11.3). Substituting the resulting expression for $\tilde{q}$ into (11.14) yields

$$\hat{R}(x_0, y_0, z_0) = \left| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x,y,z) e^{j2kz} n_e\left(\frac{z_0 - z}{c/2}\right) \times [s'(x - x_0, y - y_0, z)]^2 \, dx\, dy\, dz \right|, \tag{11.17}$$

where

$$s'(x,y,z) = s(x,y) ** e^{jk(x^2+y^2)/2z}. \tag{11.18}$$

Unfortunately, (11.17) cannot be simplified further, and it does not yield an intuitively satisfying or practically useful expression describing the relationship between R and $\hat{R}$.

The Fraunhofer approximation for $\tilde{q}$ is found from (10.77) and (11.3). Substituting the resulting expression for $\tilde{q}$ into (11.14) yields

$$\hat{R}(x_0, y_0, z_0) = \left| \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty R(x,y,z) e^{j2kz} n_e\left(\frac{z_0 - z}{c/2}\right) \times e^{jk\left((x_0-x)^2+(y_0-y)^2\right)/z} \left[S\left(\frac{x_0 - x}{\lambda z}, \frac{y_0 - y}{\lambda z}\right)\right]^2 dx\, dy\, dz \right|. \tag{11.19}$$

We can now make two observations. First, the term $\exp\{jk\left((x_0 - x)^2 + (y_0 - y)^2\right)/z\}$ is approximately unity near the transducer axis (where $x \approx x_0$ and $y \approx y_0$) and far in range (where Fraunhofer applies). Therefore, we can replace it with the value 1. Second, the expression $[S(\cdot,\cdot)]^2$ changes slowly with respect to z, while n_e is only a brief pulse in z. Therefore, function $[S(\cdot,\cdot)]^2$ can be viewed as being *indexed* by z—that is, a different $[S(\cdot,\cdot)]^2$ is used for each range considered. Incorporating these two observations as approximations, we see that (11.19) can be approximately written as the following (three-dimensional) convolution:

$$\hat{R}(x,y,z) = \left| R(x,y,z)e^{j2kz} * \left[S\left(\frac{x}{\lambda z},\frac{y}{\lambda z}\right)\right]^2 n_e\left(\frac{z}{c/2}\right)\right| . \tag{11.20}$$

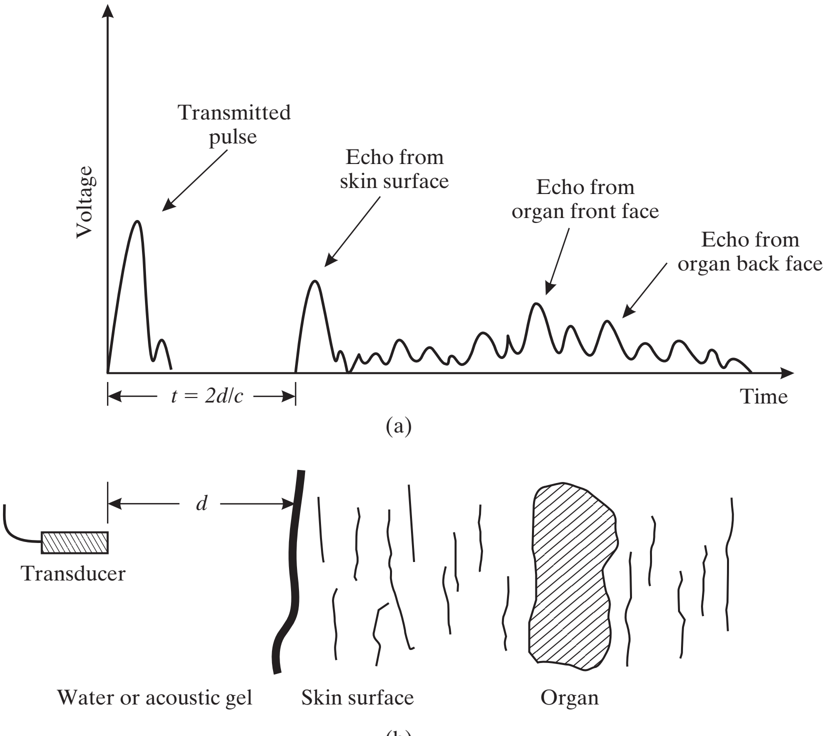
The resulting image is a local blurring of the object, where the blurring function (the resolution cell) gets wider with increasing depth. This system is not shift invariant but is approximately so within a narrow range of depth.

From (11.20), we get a feeling for the visual characteristics of ultrasound images. First, because of the term e^{j2kz}, speckle is still present in both the Fresnel and Fraunhofer approximations. Second, we see that the range resolution remains constant as a pulse propagates through the tissue. Third, we see that the lateral (azimuth) resolution degrades with range (except when focusing is used). Finally, we see that the lateral shape of the resolution cell usually has sidelobes, caused by the form of $S(\cdot,\cdot)$, which typically looks like a sinc function (see Section 2.2.4) in a direction parallel to the transducer's face. Sidelobes will cause artifacts in the resultant images, especially when there is a strong reflector in a sidelobe whose normal direction is aimed directly at the transducer.

11.5 Ultrasound Imaging Modes

11.5.1 A-Mode Scan

The starting point for ultrasound imaging systems is the envelope detected (gain-compensated) signal, $e_c(t)$ in (11.11); this signal is called the *A-mode signal* or



(a)

(b)

Figure 11.11
(a) A typical A-mode display and (b) its relationship to the anatomy.

amplitude-mode signal. By firing the transducer repeatedly a succession of these signals can be displayed on an oscilloscope. This display is called the *A-mode scan* (or *A-mode display*); an example is shown in Figure 11.11. The time between successive firings is called the *repetition time* (or *interval*), and is denoted by the symbol T_R. This interval should be long enough so that the returning echoes have died out but fast enough to capture any motion that might be present and of medical interest. For example, the A-mode display is useful when looking at heart valve motion.

11.5.2 M-Mode Scan

An *M-mode scan* is obtained by using each A-mode signal as a column in an image, with the value of the A-mode signal becoming the brightness of the M-mode image. Successive A-mode signals are displayed in successive columns, wrapping around when the last column is reached. Motion of objects along the transducer axis is revealed by bright traces moving up and down across the image, as shown in Figure 11.12. This mode is most often used to image the motion of the heart valves, and it is therefore usually displayed along with the ECG signal.

11.5.3 B-Mode Scan

A *B-mode scan* is created by scanning the transducer beam in a plane. One way to do this is to move the transducer in the x direction while its beam is aimed down the z-axis, as shown in Figure 11.13. Periodically firing pulses produces a succession of A-mode signals that are keyed to the x-position of the transducer. The B-mode image is created by brightness-modulating a computer display along a column using the corresponding A-mode signal. Given knowledge of the position of the transducer at the time of a pulse, it is a relatively simple matter to plot the correct line on a computer display and build up a full scan over time.

Scanning a single transducer was the dominant B-mode imaging method in early ultrasound imaging. Instead of relying upon a fixed method of translation, however, these early scanners had mechanical arms to hold the transducer and used potentiometric position sensors to determine the position and orientation

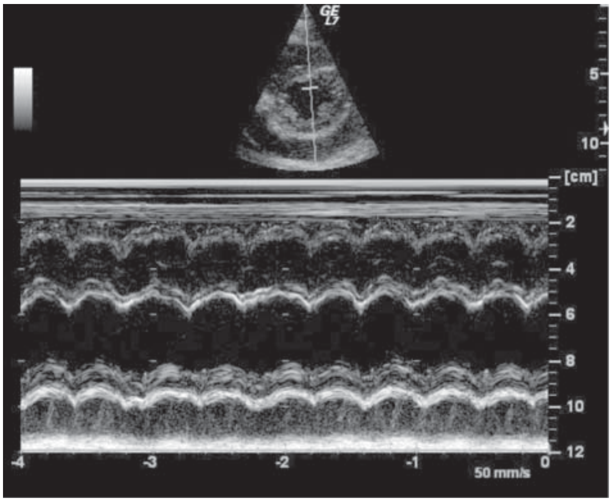


Figure 11.12
An M-mode scan is shown on the bottom part of this image. For positioning purposes, a line over which the M-mode data are acquired is shown on the B-mode image above. Courtesy of GE Healthcare.

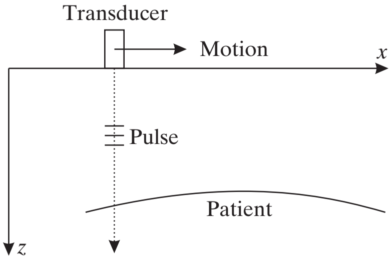


Figure 11.13
A simple B-mode scanner.

of the transducer at any time. The operator could often "fix" certain directions or orientations in order to produce a tomographic scan containing data from only one cross-sectional slice.

One important advantage of these manual-scan systems is that the transducer can be angled to hit the same point in the body from different directions. If a specular reflector exists at a point, a strong echo can be obtained from the direction normal to its surface. Therefore, images can be built up by taking the maximum echo from a given point within the body. When multiple views of the same tissue are included in a single B-mode image, it is referred to as *compound B-mode scanning*. Unfortunately, images made in this way often suffer from severe artifacts due to refraction. The sound beam refracts differently depending on the incidence angle to a surface; hence, what the electronic system computes as different measurements at the same point of tissue beyond the surface is often wrong. As a result, small intensity scatterers are often plotted at different points on the screen, producing object-dependent blurring. To reduce this blurring effect, these early systems often plotted only the largest amplitude returns—those corresponding to specular reflectors only.[1] These early images often looked like anatomical sketches, depicting only the outline of organs and tumors.

A further disadvantage of these early systems is that they were not real-time. Modern systems are generally real-time and images are acquired with the transducer probe held in one position. Three types of B-mode scanners now dominate: linear scanners, mechanical sector-scanners, and phased array sector-scanners, as shown in Figure 11.14. In practice, most B-mode images display the reflectivity value at a given point by acquiring data from a single orientation. However, phased arrays with electronic steering capability have made real-time

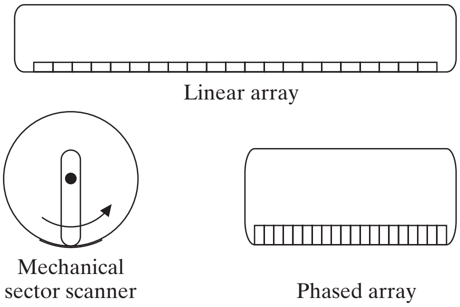


Figure 11.14
Three B-mode scanner types.

---

[1] Also, the earliest systems used storage oscilloscopes and therefore could plot only binary images (i.e., images with only two shades of gray).

compound imaging possible, and although the small angular separation of the two source positions (within the probe itself) does not add significantly to the specular echo diversity, it does have the benefit of reducing speckle. This mode of operation will be described in Section 11.8.2.

The *linear scanner* has a collection of transducers arranged in a line, thus mimicking the linear translation of a single transducer but without requiring motion. Generally, several transducers are tied together on transmit and receive to synthesize a larger aperture transducer which can be focused and can generate more power. The linear scanner requires a large flat area with which to maintain contact with the body; thus, abdominal imaging is its primary use, with obstetrics being a common use.

Sector scanners come in both mechanical and phased array varieties. The *mechanical sector scanner* simply pivots a transducer about an axis orthogonal to the transducer's axis (typically through the body of the transducer, as shown in Figure 11.14). Defining the pivot point as the origin, a collection of rays emanating from the origin can be imaged. Since the rays diverge with depth, the tissue is likely to be oversampled near the origin and undersampled far away. Since all medical ultrasound imaging systems use focused transducers, however, the lateral resolution will be best near the focal region and worse both nearer and farther away from the transducer.

The *phased array sector scanner* has a collection of very small transducer elements arranged in a line. The overall total linear dimension of a phased array sector scanner is much smaller than that of a linear scanner. The ultrasound beam is electronically steered over a sector by applying differing time-delays (or phases) to the individual transducers. One very important advantage of this type of system is that the receive focus can be varied over time with the expected position of returning pulses, thus providing a dynamic focus. A disadvantage of these systems is that spurious sidelobes (called "grating lobes") of acoustic energy are generated by the phasing action and can lead to image artifacts.

Both types of sector scanners require conversion of polar coordinate data (the A-mode signals returning from along rays) to rectangular coordinates (the image memory driving the computer display). The process of converting from polar coordinates to rectangular coordinates is called *scan conversion*. Proper scan conversion reduces sampling artifacts (moiré effects, see Section 3.6) without excessive low-pass filtering (and resultant loss of resolution).

EXAMPLE 11.4

A linear transducer array is operating at 2 MHz in water. The face of each transducer element is 1 mm wide in the x direction and 10 mm in the y direction.

Question If we want geometric approximation to be valid up to 10 cm in range, how many transducers have to be grouped together?

Answer The wavelength in water at 2 MHz is $\lambda = c/f = 1{,}484$ m/s$/2 \times 10^6$ s^{-1} = 0.742 mm. The range at which the beam transitions from geometric to far field is $z = D^2/\lambda = D^2/0.742$ mm = 10 cm. We have

$$D = \sqrt{100 \times 0.742} = 8.61 \text{ mm}\,.$$

So, at least nine transducer elements must be grouped together. ■

Depth of Penetration Ultimately, although there are many other factors involved, attenuation limits the depth of penetration of ultrasound systems. Suppose our transducer/preamplifier can handle at most an L dB loss (80 dB is typical). From Example 10.5,

$$20 \log \frac{A_z}{A_0} = -L\,, \tag{11.21}$$

where A_z is interpreted as the pressure amplitude after traversing a distance of z. From (10.33) and (10.34),

$$\alpha = -\frac{1}{z} 20 \log \frac{A_z}{A_0} \tag{11.22}$$

and from (10.35) with $b = 1$ we have

$$\alpha \approx af\,. \tag{11.23}$$

The total range that a wave can travel before being attenuated below the system threshold is

$$z = -\frac{L}{af}\,. \tag{11.24}$$

The depth of penetration d_p can be only half this distance since a round-trip is required, giving

$$d_p = \frac{L}{2af}\,. \tag{11.25}$$

A short table of d_p, assuming $a = 1$ dBcm^{-1}MHz^{-1} and $L = 80$ dB, is given in Table 11.1.

Pulse Repetition Rate A new pulse can be generated only after all echoes from the previous pulse have died out. After imaging tissues at the maximum depth of penetration, all potential remaining echoes are below the level of detection. Therefore, the pulse repetition interval T_R has a lower bound given by the round-trip time to the depth of penetration:

$$T_R \geq \frac{2d_p}{c}\,. \tag{11.26}$$

TABLE 11.1

Typical Depth of Penetration for Given Frequencies

Frequency (MHz)	Depth of Penetration (cm)
1	40
2	20
3	13
5	8
10	4
20	2

The *pulse repetition rate* is defined as

$$f_R = \frac{1}{T_R} \,. \tag{11.27}$$

Assuming $a = 1$ dB cm^{-1} MHz^{-1} and $L = 80$ dB, the minimum pulse repetition interval is 0.267 ms. In other words, in this case $f_R = 3,750$ pulses per second can be generated without danger of confusing the returning pulses with the wrong transmitted pulse.

B-Mode Image Frame Rate Suppose N pulses are required to generate an image (linear or sector). Then, the image frame rate F is

$$F = \frac{1}{T_R N} \,. \tag{11.28}$$

Continuing the example from above, if $N = 256$, then $F = 14.6$ frames/second. Typical frame rates in commercial ultrasound systems are between 10 and 100 frames/second.

A moving picture shown at 15 frames/second creates a great deal of flicker, and this would be unacceptable in an ultrasound system. Scan conversion solves this problem. The A-mode data are read into a dual-ported memory as fast as they come in. Simultaneously, the scan converter is reading out the data at a higher rate in order to generate around 60 frames/second for the computer display. Aside from flicker, however, it may be desirable to capture motion occurring faster than what can be viewed at 15 frames/second. A standard feature in most modern ultrasound systems is the capability to reduce the field of view—for example, by reducing the angle of the sector or the linear extent of a linear scanner—so that it may be scanned more rapidly (because N is reduced).

EXAMPLE 11.5

An ultrasound imaging system is operating in B-mode and requires 256 pulses to generate an image. Assume that the transducer is sensitive to at most 80 dB loss.

Question If the material being imaged has a speed of sound $c = 1,540$ m/s and $a = 1$ dB cm^{-1}MHz^{-1}, what should the working frequency be to achieve a frame rate of 15 frames/second?

Answer Given the frame rate and the number of pulses needed to generate an image, we can compute the pulse repetition rate as

$$T_R = \frac{1}{FN} = 0.26\,\text{ms} \,.$$

Since

$$T_R \geq \frac{2d_p}{c} = \frac{L}{afc} \,,$$

we have

$$f \geq \frac{L}{acT_R} = 1.99\,\text{MHz} \,.$$

■

11.6 Steering and Focusing

Phased arrays use electronic steering and focusing to achieve a B-scan without transducer movement. They do this using a small linear array of transducer elements (typically between 64 and 128), with total dimension of around 1 cm (height of each element) by 3 cm. The width of each element is about 1/4 wavelength, which is typically 0.2–0.75 mm wide. The extreme rectangular shape of each element allows mechanical focusing to keep the sound largely in the image plane but guarantees that each element will have a nearly circular field pattern within the image plane. On transmit, a directed beam is achieved by timing the firing of each element so that the sound each produces adds coherently in the desired direction and incoherently otherwise. On receive, the array is made direction-sensitive by delaying the received signals and summing them coherently. In the following two sections, we expand on these ideas and bring some elementary trigonometry to bear in the analysis of phased array systems.

11.6.1 Transmit Steering and Focusing

As shown in Figure 11.15, adding a separate delay element to each transducer of a phased array makes it possible to steer the transducer's acoustic beam. To analyze this situation further, assume that T_0 (transducer 0) generates a pulse at $t = 0$. We want to determine what firing times are necessary for the remaining transducers in order to generate a plane wave heading off in direction θ. At time t, the leading edge of the pulse generated by T_0 has traveled distance $r_0(t) = ct$. Therefore, the line $L(t, \theta)$ representing the wavefront in the direction θ at time t is [by analogy to (2.10)] given by

$$L(t,\theta) = \{(x,z) |\; z\cos\theta + x\sin\theta = r_0(t)\} \,. \tag{11.29}$$

By simple geometry, the distance between T_i (the ith transducer) and this line is

$$r_i(t) = r_0(t) - id\sin\theta \,. \tag{11.30}$$

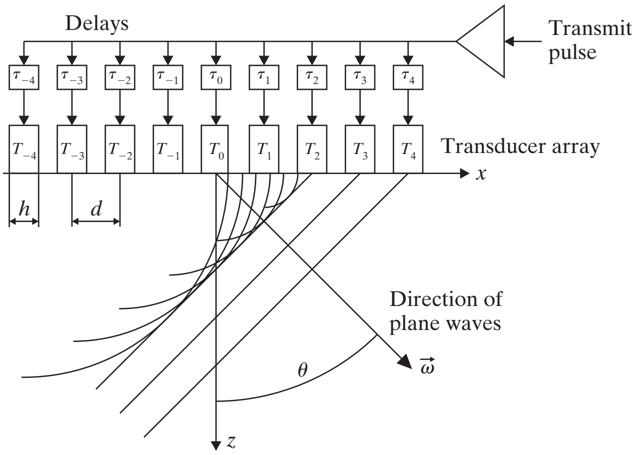


Figure 11.15
Geometry for steering a phased array.

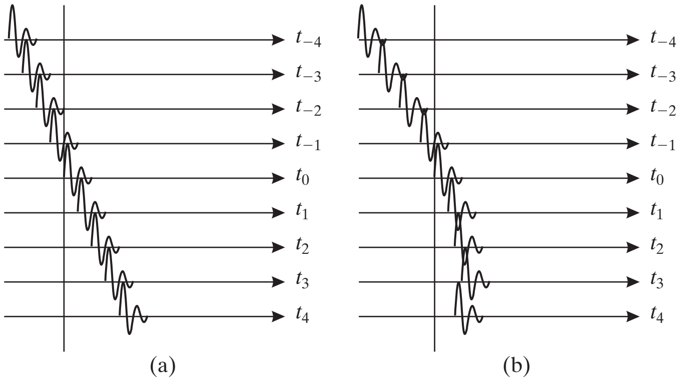


Figure 11.16
Timing diagram for (a) steering and (b) focusing.

To generate a pulse that will also have a leading edge on $L(t, \theta)$, the pulse must be generated at time

$$t_i = \frac{r_0(t) - r_i(t)}{c} = \frac{id \sin\theta}{c} \,. \tag{11.31}$$

The sequences of pulses generated at these times are depicted in Figure 11.16(a). In practical terms, we cannot have negative times, so the resulting times must be shifted to positive time. Assume that there are $2N+1$ transducer elements centered at the origin, and let $t_{\min} = \min\{t_i, i = -N, \ldots, N\}$. Then, the time delays are given by

$$\tau_i = t_i - t_{\min}, \;\; i = -N, \ldots, N \,. \tag{11.32}$$

Typically, the phased array will be focused, which is just a refinement of steering, as shown in Figure 11.17. To analyze steering and focusing together, let the focal point be at coordinate (x_f, z_f); T_i is at coordinate $(id, 0)$. Therefore, the range from T_i to the focal point is

$$r_i = \sqrt{(id - x_f)^2 + z_f^2} \,. \tag{11.33}$$

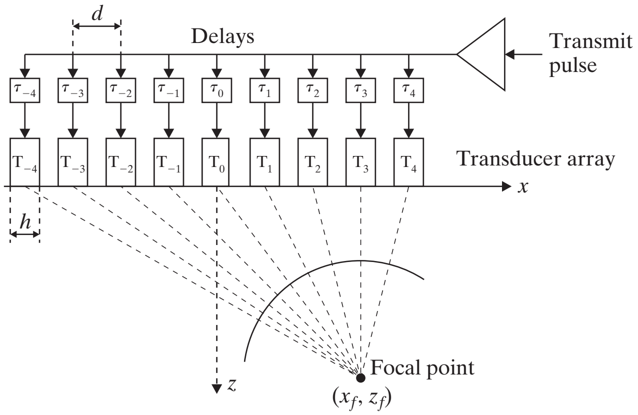


Figure 11.17
Geometry for steering and focusing a phased array.

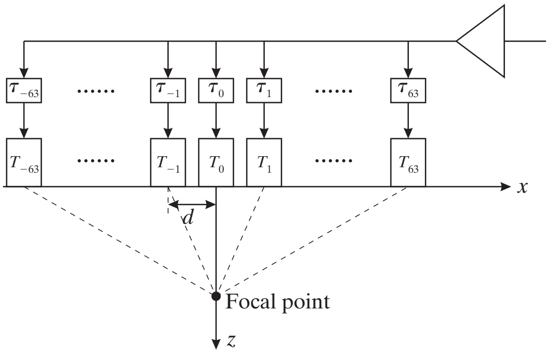


Figure 11.18
A linear transducer array with an on-axis focal point.

Assume as before that T_0 fires at $t = 0$. Then, T_i must fire at

$$t_i = \frac{r_0 - r_i}{c} = \frac{\sqrt{x_f^2 + z_f^2} - \sqrt{(id - x_f)^2 + z_f^2}}{c} \tag{11.34}$$

to reach the focal spot at the same instant, and the delays are also given by (11.32). Figure 11.16(b) depicts the pulse timings that would be used to generate a focused beam. The difference between these delays and those of steering alone is that the delays are not simply a multiple of a base delay and the transducers need not be fired in the same order as their geometric order. A typical depth for the focal spot would be $z_f = d_p/2$; in this way, the narrowest portion of the beam would be at one-half the depth of penetration.

EXAMPLE 11.6

A phased array transducer with 127 elements is operating in water. Adjacent transducer elements are separated by $d = 0.8$ mm (see Figure 11.18).

Question We desire a focal point at $z = 5$ cm on the z-axis. Suppose the outermost transducers fire at $t = 0$, when does the center element fire?

Answer The distance between the focal point and the centers of the outermost transducers is

$$r_{63} = \sqrt{(63d)^2 + (5\text{ cm})^2} = 7.1\text{ cm}.$$

The difference between r_{63} and r_0 is $r_{63} - r_0 = 2.1$ cm. In water, the speed of sound is 1,484 m/s, so the center element fires at

$$\tau_0 = \frac{2.1\text{ cm}}{1{,}484\text{ m/s}} = 14.15\,\mu\text{s}.$$

■

11.6.2 Beamforming and Dynamic Focusing

Although the acoustic energy from a phased array pulse is timed to provide coherent summation in a direction or at a point, in fact the energy propagates quite widely due to the circular wave pattern of each element of the array. In order to achieve better sensitivity to echoes occurring along the dominant

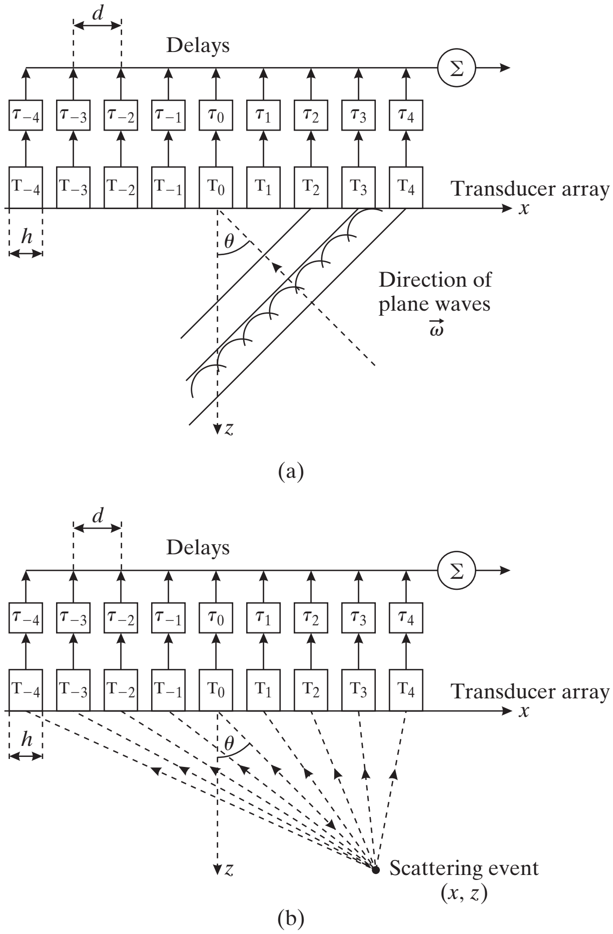

d
Delays
Σ
τ_{-4} τ_{-3} τ_{-2} τ_{-1} τ_0 τ_1 τ_2 τ_3 τ_4
T_{-4} T_{-3} T_{-2} T_{-1} T_0 T_1 T_2 T_3 T_4
Transducer array
x
h
θ
Direction of plane waves $\vec{\omega}$
z
(a)
d
Delays
Σ
τ_{-4} τ_{-3} τ_{-2} τ_{-1} τ_0 τ_1 τ_2 τ_3 τ_4
T_{-4} T_{-3} T_{-2} T_{-1} T_0 T_1 T_2 T_3 T_4
Transducer array
x
h
θ
z
Scattering event (x, z)
(b)

Figure 11.19 Geometry for (a) beamforming and (b) dynamic focusing of a phased array.

(steered) direction, two techniques are commonly employed: beamforming and dynamic focusing.

Beamforming is analogous to steering in that it increases the transducer's sensitivity to a particular direction. As shown in Figure 11.19(a), a plane wave incident upon the transducer array from a direction θ will hit one transducer at the end of the array first and then successively hit the remaining transducers. If the first transducer to be hit, and thereafter each successive transducer, delays its received waveform so that all the waveforms can be coherently summed, then the entire array is sensitized to the direction θ. The required receive delays are identical to the transmit delays specified in (11.31) and (11.32) for the transmission of a plane wave in direction θ. Likewise, a fixed focal point can be achieved by adopting the delays used for combined steering and focusing in (11.32) and (11.34).

Additional sophistication is achieved by using *dynamic focusing*, which increases the transducer's sensitivity to a particular point in space at a particular time. Suppose that T_0 is fired at $t = 0$ in the direction θ (steered or focused), as shown in Figure 11.19(b). Assuming that a scattered spherical wave originates at position (x, z), then the total distance the pulse travels from T_0 to arrive at transducer T_i is

$$r_i = \sqrt{x^2 + z^2} + \sqrt{(id - x)^2 + z^2}\,. \tag{11.35}$$

The time difference between the pulse's arrival at T_0 and T_i is

$$t_i = \frac{\sqrt{x^2 + z^2} - \sqrt{(id - x)^2 + z^2}}{c}\,. \tag{11.36}$$

In fact, as the transmit pulse propagates in the direction θ, the (x, z) position of the scattering center changes with the position of the pulse. Therefore, the time differences also change as a function of time. The position of the transmitted pulse along the θ axis is given by $(x, z) = (ct \sin\theta, ct \cos\theta)$. Therefore, the time differences are given by

$$t_i(t) = \frac{\sqrt{x^2(t) + z^2(t)} - \sqrt{(id - x(t))^2 + z^2(t)}}{c} \tag{11.37}$$

$$= t - \frac{\sqrt{(id)^2 + (ct)^2 - 2ctid \sin\theta}}{c}\,. \tag{11.38}$$

To avoid negative time delays, it is necessary to shift these times by subtracting $t_{\min} = -Nd/c$. Hence, dynamic focusing is achieved by dynamically altering the receive time delays according to

$$\tau_i(t) = t - \frac{\sqrt{(id)^2 + (ct)^2 - 2ctid \sin\theta}}{c} + \frac{Nd}{c}\,. \tag{11.39}$$

EXAMPLE 11.7

Consider a dynamic focusing phased array with 127 transducer elements that are separated by $d = 0.8$ mm. A point scatter on the z-axis at a range of 5 cm starts to move in the $+z$ direction with a speed of $v = 5$ cm/s at time $t = 0$.

Question Assuming that the speed of sound is $c = 1,500$ m/s, plot the difference between the delays $\tau_{63}(t)$ and $\tau_0(t)$.

Answer Let

$$\tau_d(t) = \tau_{63}(t) - \tau_0(t)\,.$$

The distance between the scatter and the center of the 63rd transducer at time t is

$$r_{63}(t) = \sqrt{[(63 \times 0.8)\ \text{mm}]^2 + [(5 + 5t)\ \text{cm}]^2}\,.$$

The distance between the scatter and the center of the middle transducer at time t is

$$r_0(t) = (5 + 5t)\ \text{cm}\,.$$

So, the difference in the time delay is

$$\tau_d(t) = \frac{r_{63}(t)}{c} - \frac{r_0(t)}{c} = \frac{1}{c}\left[r_{63}(t) - r_0(t)\right],$$

which is plotted in Figure 11.20. ■

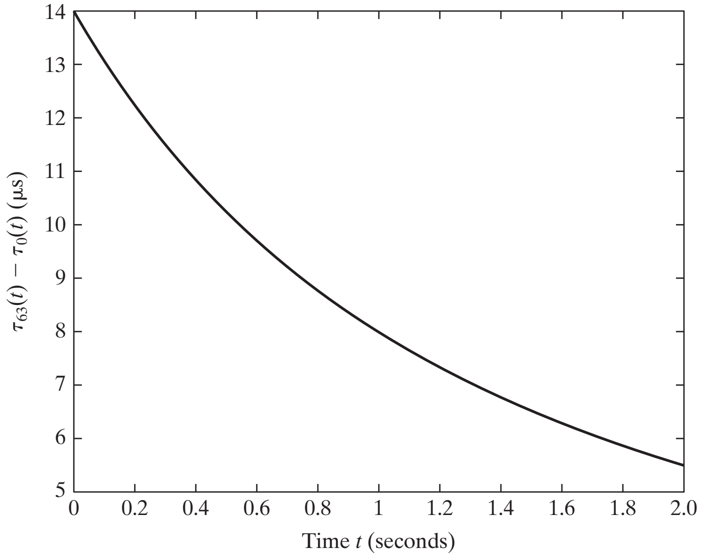


Figure 11.20
The difference in time delay between the outermost and the center transducer in a dynamic focusing phased array with a moving scatterer.

11.7 Three-Dimensional Ultrasound Imaging

Three-dimensional ultrasound is made possible by either scanning conventional linear or phased arrays or by using 2-D arrays and electronically scanning a volume beneath the stationary transducer. The geometries of volumes acquired in these ways are illustrated in Figure 11.21. The mechanical wobbler design that we described, is the most common design for 3-D imaging probes. For this purpose, a mechanical mechanism like that shown in Figure 11.7(a) is used to rotate an entire array of transducers, such as the one shown in Figure 11.8. The correct geometry is achieved by rotating the array in Figure 11.8 into the page by 90° in order to position it "end on" in Figure 11.7(a). Given this geometry, at each angular position of the wobbler, the array can acquire an entire 2-D image (each shaped like a sector) that is orthogonal to the plane of

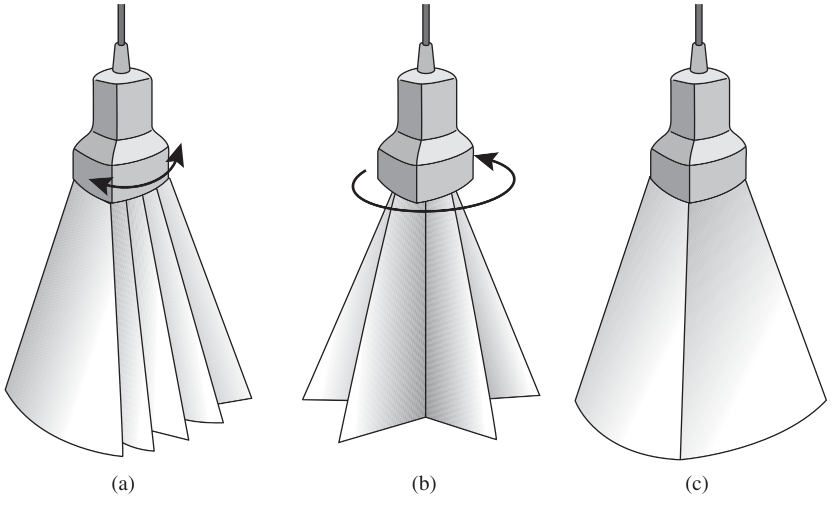


Figure 11.21
Geometry of 3-D ultrasound images. (a) Rocking a 1-D array using a mechanical scanner; (b) rotating a 1-D array; and (c) a stationary 2-D array that can acquire a volume without moving the probe.

our drawing [see Figure 11.21(a)]. The resulting volume of data is generated by reconstructing a series of 2-D images and then positioning them into a 3-D volume according to their rotational position at the time of acquisition.

By rotating the array around its axis, a series of rotated sector scans can be acquired, as shown in Figure 11.21(b). This type of system is less popular, in part because it requires more room on the patient's skin to carry out this rotation and may therefore not be physically capable of imaging some parts of the body. 2-D transducer arrays are also available commercially. The elements are typically square and there are the same number of elements in each dimension of the square transducer face. In this case, a volume can be acquired without motion by electronically scanning the beam beneath the transducer, as shown in Figure 11.21(c). A popular hybrid imaging mode in which only two orthogonal scans are acquired [like that depicted in Figure 11.21(b)] using a two-dimensional array (without transducer motion). In this way, orthogonal images can be collected and displayed in real time. Clearly, the 2-D array also requires more space on the patient's skin than a conventional array and may also not be suitable for some imaging scenarios.

A final 3-D mode not depicted here is a "free hand" mode, in which the transducer (a 1-D array acquiring a 2-D slice) is moved freely over the patient while acquiring data. The transducer is tracked using external markers so that the positions of all acquired slices are known in the laboratory frame, and a digital representation of the data is reconstructed into a volume by the scanner. Trackerless freehand systems are also under development. These systems require the operator to slowly scan the probe in the through-plane direction across the exposed tissue. A computer algorithm reconstructs a 3-D scene from the collection of 2-D slices that are observed by tracking the speckle patterns from frame to frame, which gives cues about the relative positions of the slices in 3-D.

11.8 Image Quality

An ideal ultrasound imaging system would reconstruct and display $R(x, y, z)$, the spatial distribution of reflectivity. This is not possible for two reasons. First, the transducer's impulse response function, defined in part by its field pattern, will blur out the true reflectivity. Second, the use of envelope detection will create a noiselike artifact called *speckle*; this accounts for the somewhat "blotchy" appearance of ultrasound images. There are other reasons, not the least of which is that some objects reflect, rather than scatter, ultrasound; thus, they do not even have the desired scattering property that we have called reflectivity.

11.8.1 Resolution

The resolution cell in (11.20) gives the most basic description of image resolution in ultrasound imaging. The range resolution is dominated by the pulse shape (envelope), as shown in Figure 10.6, while the lateral resolution is determined by the field pattern, as shown in Figure 10.5. Looking closely at virtually any ultrasound image, one can see in the pattern of echoes from small scatterers that the range resolution is constant, while the lateral resolution changes with range. This is particularly noticeable in Figure IV.3, for example.

There are a couple of ways to improve resolution that we have not considered to this point: through-plane focusing and harmonic imaging. We now give a brief description of these two approaches.

Through-Plane Focusing A key factor relating to the quality of ultrasound B-mode images is *slice thickness*. Whether the transducer is mechanically or electronically scanned, the resolution of the B-mode image is determined in part by the pulse duration—which determine resolution *range*—and the transducer aperture and in-plane focus properties—which determines resolution in the *azimuth* direction. Just like in computed tomography (CT), in which the slice thickness determined by the through-plane width of the x-ray beam affects the quality of the reconstructed image, the width of the ultrasound beam in the through-plane direction—also called the *elevation* direction—affects the quality of the reconstructed images.

Up to this point, we have assumed a fixed slice thickness, which is determined by the shape of the transducer in the elevation direction. In the case of a rectangular element without a manufactured curvature or an added lens, the standard nearfield and Fraunhofer approximations (see Figure 10.5) can be used to approximate the field pattern (and therefore slice thickness) in the elevation direction. We know that in this case, assuming the frequency of the transducer is fixed, the only way to change the field pattern is by changing the dimension *D* of the transducer. In order to enable this function "on the fly," we can fabricate additional arrays of transducer elements in the elevation direction, as shown in Figure 11.22. With dimensions selected appropriately, the slice thickness can be controlled as shown. In this case, three dimensions can be selected to yield thicker or thinner slices at specified ranges. Exciting only the center row ① yields the thinnest nearfield but the widest far field, while exciting the three central rows (①, ②, and ③) yields a wider nearfield but narrower farfield. Exciting all

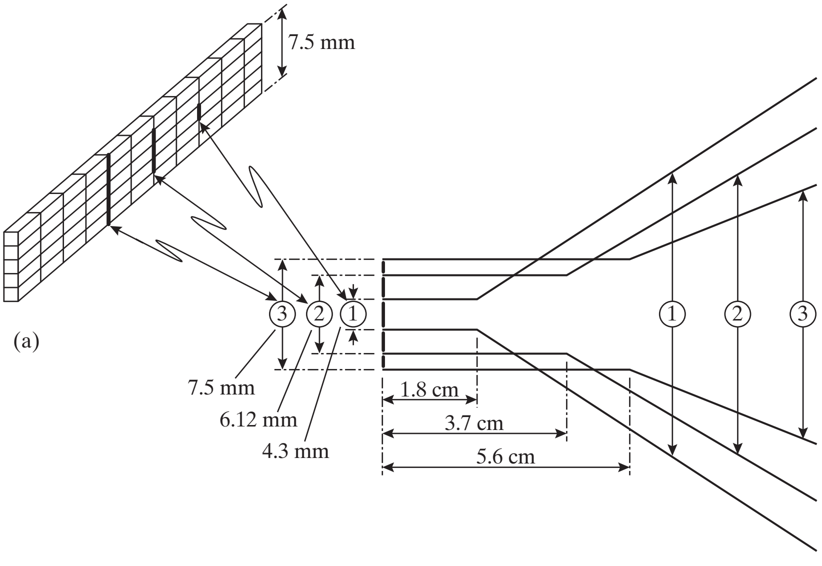


Figure 11.22
Geometry of a 1.5-D array. (a) There are multiple arrays in the elevation direction (vertical in this diagram). (b) Different slice thicknesses are achieved by exciting different rows together.

three rows yields the widest slice thickness in the nearfield but the narrowest in the far field.

Since we are aware of the possibility to introduce transmit time delays, it should be clear that with a phased array system, we can improve on the static scenario above. In particular, by firing all five rows of transducers (in the array illustrated in Figure 11.22), but starting first with the outer two, then the middle two and then the central row, we can achieve some degree of focusing of the beam in the elevation direction. We can choose an arbitrary focal length, and if there are even more rows available, the quality of the focus can be further improved. This flexible focal length has been exploited in some systems to provide a more uniform slice thickness by scanning the plane multiple times, changing the focal depth in both the elevation (slice thickness) and azimuth (lateral dimension) to improve resolution at different depths. In the scenario of Figure 11.22, for example, we might use ① to fill up pixels that represent the smallest ranges, ② to fill in intermediate ranges, and ③ to fill in the farthest ranges. With three scans required to produce a complete B-mode image, one sacrifices frame rate in order to gain resolution, which might not be feasible in some scenarios.

Harmonic Imaging We saw in Section 10.3.6 that sinusoidal waves can become more like sawtooth waves as they propagate due to the nonlinear effects of wave propagation. We also saw that sawtooth waves possess energy at harmonic frequencies—that is, multiples of the fundamental frequency. Through the use of broadband transducers, as described in Section 11.2.1, it is possible to generate waves at a fundamental frequency (also known as the first harmonic) and acquire echoes comprising the fundamental and second harmonic frequencies.

Now consider the signal that is created by bandpass filtering the echoes at the second harmonic frequency, that is, filtering out the fundamental frequency. The resulting signal contains only that part of the echo arising from nonlinear interactions between the wave and the tissue. This would, on the surface, seem to offer no gains and perhaps a few disadvantages. Let us review what happens and why this signal is interesting and sometimes advantageous.

We know that nonlinear effects occur primarily at larger acoustic pressures. In fact, nonlinear effects occur because the speed of sound increases with acoustic pressure. Therefore, the main beam, which has the largest acoustic pressure, will also have the largest nonlinear effect, and the center of the main beam will have a larger effect than its edges. The side lobes and the grating lobes (present in phased array systems, as noted in Section 11.5.3) will not contribute much to the returning harmonic signal. This has the effect of improving resolution and reducing artifacts and is the primary reason that harmonic imaging is used. Because the harmonic signal in the through-plane direction also diminishes more rapidly, the through-plane resolution is also improved.

We know that the effects of nonlinear wave propagation change the propagating acoustic wave from sinusoidal to sawtooth increasingly over range. Therefore, the harmonic part of a returning echo tends to increase in amplitude over range. But at the same time, the attenuation of the higher frequency signal that must return from the scattering site is larger than the fundamental wave due to frequency-dependent attenuation. On balance, the two effects trade off with each other quite well, and harmonic images at comparable depth to conventional B-mode images can be achieved.

In order to generate sufficient separation of the fundamental and higher harmonic signals, the initial pulse in harmonic imaging is typically a longer sinusoidal burst than in conventional imaging. This provides a more narrowband acoustic pulse at the outset, from which the higher harmonics can be more easily distinguished. This comes at the cost of slightly poorer range resolution. A clever way to retain range resolution is to use two successive (conventional, short) pulses, where the second is the inverse of the first—that is, where the first pulse goes positive the second pulse will be negative. Then the returning echoes from the two pulses are subtracted, the fundamental signal is nulled, and all that remains are the higher harmonics.

In the preceding paragraphs, we have discussed the first harmonic as a primary target to acquire but also used the more general higher harmonics when applicable. It is most common to image the second harmonic because it generally falls conveniently within the bandwidth of a broadband transducer, but it is possible to image higher harmonics if the transducer is sufficiently broadband. The primary benefit is improved resolution.

11.8.2 Noise and Speckle

Electronic Noise Although the appearance of speckle is certainly ‘‘noiselike,’’ we have learned that it does not arise from a random process and is therefore not noise in the sense we use the term in this book. In fact, what little noise there is in ultrasound systems arises from the very small electronic noise that is present in the wires and other components of the system. When there is very high gain—for example, at large ranges where TGC requires such gains in order to compensate for attenuation—this initially tiny noise can become amplified sufficiently to become visible in an image. Its appearance is similar to that of speckle, but because it is statistically independent between pixels, it has a very high spatial frequency, which is not dependent on the transducer’s frequency or field pattern. Regions where noise might become apparent in a B-mode image are generally at large ranges in homogeneous regions where scatterers are not present to obscure the noise. Thus, noise in ultrasound images is easily distinguished from speckle, and its main impact is to limit the depth of penetration.

Compound Scanning for Speckle Reduction Compound scanning is the acquisition of the same 2-D scene from one or more different angles within the plane followed by averaging of the acquired images.[2] Images acquired with manual motion of a single transducer element—as required in the earliest years of medical ultrasound imaging—were routinely compounded in order to visualize boundaries that were not favorably oriented for one angle of acquisition (i.e., not orthogonal to the transducer axis). As processing of the A-mode signals improved over the years, imaging nonorthogonal boundaries became less critical while speckle became more bothersome. Compound scanning has therefore re-emerged as an important speckle reduction technique.

We know that although speckle looks like noise in an image, it actually arises from the coherent and incoherent summation of echoes from scatterers

---

[2] Technically this approach is called *angular compounding* since other ways of acquiring separate images followed by averaging are also called compounding.

within a resolution cell. This creates a pseudorandom variation of intensity from pixel to pixel, which looks like noise to us and is desirable to reduce. This leads to the notion of *speckle reduction* in ultrasound imaging.

Images acquired from different angles will have resolution cells that are oriented differently at each point in a B-mode image, and therefore the geometric relationships of the scatterers will not be the same. As a result, although the overall appearance of the dominant echoes will be the same in the images, the speckle pattern will be slightly different. When the images are averaged, the dominant appearance will be reinforced and the speckle pattern will be reduced.

But how can these different angles be achieved without moving the probe? We have just studied how phased arrays can steer their beams in order to create sector-style B-mode images, but we developed the mathematics considering that the central transducer was always our "origin." Suppose instead that the central transducer was assigned to be at the 1/3 or 2/3 position along the array. Then the phased array transducer acts as if it has been shifted slightly to the left or right and the angle at which targets are viewed is slightly different than when the center of the transducer as the origin. This approach is called the *displaced aperture* approach to compound imaging using phased arrays. Three or more images can be acquired in quick succession and averaged, which achieves a small measure of speckle reduction.

11.9 Summary and Key Concepts

Ultrasound imaging is accomplished with a pulse-echo system based on a single transducer probe. Such systems have high performance and low cost. In this chapter, we presented the following key concepts that you should now understand:

1. The *transducer*—a primary component of an ultrasound imaging system—both generates and receives sound pulses in the *pulse-echo mode* of operation.
2. Most transducers are based on a *piezoelectric crystal* and resonate at a fundamental frequency characteristic of that transducer.
3. The ultrasound beam must be *steered*, *scanned*, or *swept* across the body to produce a useful field of view.
4. The *basic imaging equation* is the *pulse-echo equation*, which arises from generalization of either the Fresnel or Fraunhofer approximation.
5. The basic imaging equation can be transformed into an equation describing the estimated *reflectivity* of the tissues.
6. *Ultrasound imaging modes* include *A-mode scanning*, *M-mode scanning*, *B-mode scanning*, and *3-D ultrasound imaging*.
7. Practical B-mode scanning requires *steering* or *beamforming* and *focusing* of the ultrasound beam.
8. *Harmonic imaging* can be used with broadband transducers to improve resolution.
9. *Compound imaging* can be achieved using phased arrays by averaging images acquired by *aperture displacement*.

Further Reading

Cobbold, R.S.C. *Foundations of Biomedical Ultrasound*. New York, New York: Oxford University Press, USA, 2006.

Kremkau, F.W. *Diagnostic Ultrasound: Principles and Instruments*, 7th ed. Philadelphia, PA: W.B. Saunders, 2005.

Macovski, A. *Medical Imaging Systems*. Englewood Cliffs, NJ: Prentice Hall, 1983.

Pierce, A.D. *Acoustics: An Introduction to its Physical Principles and Applications*. New York, NY: McGraw-Hill, 1989.

Prager, R.W. *Three-dimensional ultrasound imaging*. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine. vol. 224, pp. 193–223, 2010.

Wells, P.N.T. *Biomedical Ultrasonics*. London, England: Academic Press, 1977.

Wells, P.N.T. *Ultrasound Imaging*. Physics in Medicine and Biology. vol. 51, pp. R83–R98, 2006.

Problems

Ultrasound Image Formation and Imaging Modes

11.1 An $L \times L$ square transducer, submersed in water, is centered at the origin and is pointed down the $+z$-axis. A large layer of homogeneous fat begins $z = z_0$ and extends to the depth of penetration, so that the only possible returns are those from the interface between water and fat. The transducer is capable of rotating toward the x-axis by θ degrees. Assume that there is no attenuation in the water. Assume that the transmitted pulse is a gated sinusoid of duration T and height A_0.

(**a**) Assume that z_0 is in the far field and derive an expression for the strength of return as a function of angle. Make approximations where appropriate and do not forget the sidelobes.

(**b**) Assume that the transducer and preamplifier are able to detect signals 80 dB down from the transmitted pulse height. Assume also that a B-mode binary image is generated so that all signals above the detectable threshold are plotted in black and all signals below the threshold are plotted in white. Sketch the B-mode image that would result from this experiment.

11.2 Two identical square transducers with width 1 cm face each other. The first one is located at the origin and points in the $+x$ direction; the second one is located at $(10\text{ cm}, 0, 0)$ and points toward the origin. The first one will be used to image the second one. Assume a homogeneous medium with $\rho_0 = 1,000\text{ kg/m}^3$, $c = 1,500\text{ m/s}$, and $\alpha = 1\text{ dB/cm}$. Assume that the first transducer fires a perfect geometric beam with peak transmit acoustic pressure measured at its face of 12.25 N/cm^2; assume that the second transducer is a perfect reflector.

(**a**) Sketch the A-mode signal. Label the axes carefully and identify the time-of-return and peak-height (as an acoustic pressure at the face) of the returning pulse.

(**b**) At time $t = 2$ s the second transducer begins to move back-and-forth along the x-axis with x position $x(t) = 10 + \sin 2\pi(t - 2)$ cm,

$t \geq 2$ seconds. Sketch the M-mode plot for $0 \leq t \leq 5$ seconds. Label the axes carefully; identify key points on your plot.

(c) Now suppose the second transducer stops at its original position and we allow the first transducer to move along the y axis. Sketch the resulting B-mode image. Label the axes carefully; identify key points on your plot. Make a sketch of the peak-height of the returning pulse as a function of the y position of the first transducer.

11.3 You are using a single transducer to examine a heart valve. Assume that in a given heart cycle the range of the valve is given by $z(t) = 16 + 0.5e^{-t/\tau}u(t)$ cm where $\tau = 10$ ms, $u(t)$ is the unit step function, and the speed of sound is 1,540 m/s.

(a) Sketch $z(t)$ over a couple of heart cycles (assume the heart rate is 1 Hz).

(b) Assume your transducer generates a pulse at $t = 0$ using a typical transmit waveform. Carefully sketch the A-mode signal (as a function of time) that you would observe on an oscilloscope.

(c) Now assume that you repeatedly fire your transducer every 1 ms. Sketch the M-mode image that would be generated, being careful to label the axes.

(d) Suppose you wanted to image the motion of this valve by making a B-mode image of it. Suppose it could be covered by 10 scan lines (given the beam size at a 16 cm range). Describe what steps you would take to make this image in real-time. Do you think it is possible?

11.4 An ultrasound transducer is pointing down the z-axis, on which two point scatterers are located at z_1 and z_2 in the Fraunhofer field (see Figure P11.1). The transducer fires a narrowband burst with a rect-shaped envelope in the $+z$ direction. The envelope is $n_e(t) = \text{rect}((t + T/2)/T)$ where $T = \lambda/2c$, λ is the wavelength, and c is the speed of sound in the material. Sketch the estimated reflectivity for

(a) $z_2 - z_1 = \lambda/2$.

(b) $z_2 - z_1 = \lambda/8$.

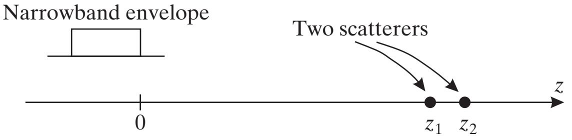


Figure P11.1 An ultrasound transducer and two scatterers on axis. See Problem 11.4.

Ultrasound Transducer Array

11.5 A linear ultrasound transducer array has 101 transducers, each having width $d = 0.1$ mm, that are separated by $h = 0.1$ mm. Assume the simple field pattern geometry of Figure 10.5 and $c = 1{,}540$ m/s.

(a) What is the maximum frequency at which the transducers can operate and still generate a plane wave at angle $\theta = 30°$?

(b) How long will it take to sequence through the array in order to generate the plane wave in part (a)?

11.6 Consider the phased array in Figure 11.15. Suppose we want to image a cross section of a patient as shown in Figure P11.2 by acquiring a scan line every $\Delta\theta = 1^\circ$. Assume there are 101 transducers operating at a frequency of 2 MHz and $c = 1,540$ m/s and $d = 0.6$ mm.

(**a**) Find an expression for the time delay of the ith transducer as a function of angle of the transmitted plane wave.

(**b**) How long does it take to scan the entire FOV assuming $R = 20$ cm?

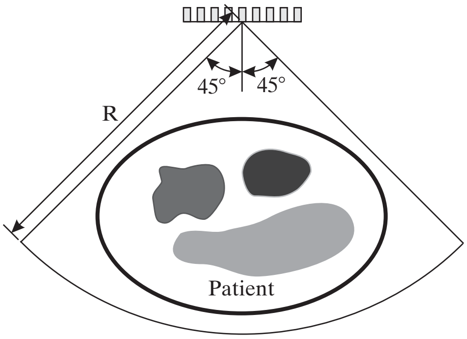


Figure P11.2
See Problem 11.6.

11.7 Consider a linear array of 100 flat, rectangular, 3.0 Mhz transducers, each 1.5 mm wide and 2.1 cm deep. Assume, for simplicity, that they are packed tightly together so that the array is 15 cm long by 2.1 cm deep. The transducer array is tested in a medium having a speed of sound of 1,500 m/s.

(**a**) If 14 consecutive transducers on the transducer array are electronically grouped on each transmit and receive, what depths (or ranges) are considered to be in the far field?

(**b**) At what time interval after the transmit pulse will an echo from an object at range 7 cm arrive back at the transducer array?

(**c**) A B-mode image is created by successively grouping 14 transducers on the array, sliding one transducer at a time down the length of the array. Suppose the depth of penetration is 20 cm. What is the frame rate for this linear array image?

(**d**) Describe two ways that the frame rate can be increased.

11.8 A phased array ultrasound transducer consists of three elements (L,C,R) each of dimensions $h = 3$ mm and $d = 8$ mm, as shown in Figure P11.3. Each element has a square face, $s(x,y) = \text{rect}(\frac{x}{h}, \frac{y}{h})$. In the first scenario [Figure P11.3(a)], the transducer has a silicone sheet of thickness $s = 2$ mm on its front face, which is placed on the skin. The acoustic impedances Z and the speed of sound c in silicone and in skin are given in the table below. For parts (a) to (e), assume that only one element, C, is generating and receiving ultrasound waves. The envelope of the wave is $n_e(t) = \text{rect}(\frac{t}{10\ \mu s})$ and its frequency is 2 Mhz.

	Silicone	Skin
Acoustic impedance (Z) (kg/m^2 second)	1.4×10^6	1.5×10^6
Speed of sound (c) (m/s)	1,500	1,550

(**a**) Sketch the A-mode signal from the time of the transmit pulse through the time of the echo from the silicone-skin interface. Assume a geometric field approximation and ignore attenuation and speckle. Clearly label key times and amplitudes.

In order to reduce the size of the echo from the skin, an acoustic gel of width $g = 3$ mm is introduced between the silicone and the skin, as shown in the figure [Figure P11.3(b)]. The gel is selected such that $Z_{\text{gel}} = \sqrt{Z_{\text{silicone}} Z_{\text{skin}}}$.

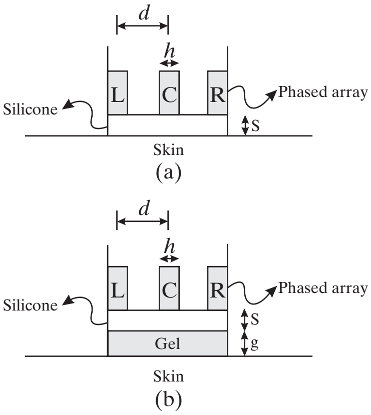


Figure P11.3
See Problem 11.8.

(**b**) Find the speed of ultrasound in the gel, c_{gel}, so that there can be no refraction at the gel-skin interface regardless of the angle of incidence.

(**c**) Sketch the A-mode signal from the time of the transmit pulse through the time of the echo from the gel-skin interface. Assume a geometric field approximation and ignore attenuation and speckle. Clearly label key times and amplitudes.

(**d**) Suppose you want your imaging system to be insensitive to the echo from the skin. What should the sensitivity L (in dB) be?

In parts (e) and (f) ignore the silicone and gel.

(**e**) Is point F in Figure P11.4 in the near field or far field of the central transducer?

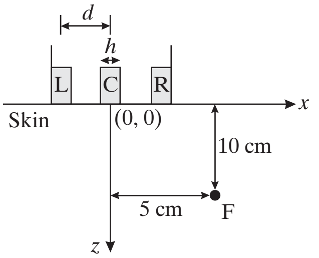


Figure P11.4
See Problem 11.8.

(f) Find the time delays τ_L, τ_C, and τ_R so that, on transmit, the phased array will focus on the point F.

Ultrasound Imaging System Design and Image Quality

11.9 A 1 cm × 1 cm ultrasound transducer is placed on the $z = 0$ plane and pointed in the $+z$ direction. Its sides are parallel to the x and y axes, and its center is at the origin. The transducer is submerged in a homogeneous medium with speed of sound $c = 1,540$ m/s. The working frequency is 2.5 MHz. A line scatterer parallel to y-axis passing through the point $(0, 0, 5)$ cm is being imaged.

(a) Write down an expression for the scatterer in mathematical form.

(b) What is the estimated reflectivity?

(c) If a second line scatterer parallel to the first one is placed at the same range, what is the minimal distance between them for the system to distinguish them?

(d) Repeat parts (b) and (c) when the scatterer is at range $z = 20$ cm.

11.10 You are asked to select the best ultrasound imaging parameters for imaging a deep-lying small structure. Assume that this object is located at 20 cm depth and is embedded within a homogeneous tissue with sound speed $c = 1,500$ m/s and absorption coefficient $\alpha = af$, where $a = 1.0$ dB/(cm MHz). Also assume that the system is sensitive to at most a 100 dB signal loss.

You can choose one of three operating frequencies: $f = 1$ MHz, $f = 2$ MHz, or $f = 5$ MHz. Given any frequency, you can choose one of two transducer diameters: $D = 1$ cm or $D = 2$ cm. Assume that all transducers have flat faces and use the split-range field pattern approximation (i.e., geometric and far-field approximations in their appropriate ranges). Which combination of f and D will give the best image of this structure? Justify your answer.

11.11 An $L \times L$ ultrasound transducer is centered at the origin and pointing in the $+z$ direction. The narrowband signal generated by the transducer has an envelope

$$n_e(t) = \operatorname{sinc}\left(\frac{\pi t}{\Delta T}\right).$$

Two point scatterers with reflectivity of R are located at $(0, 0, z_0)$ and $(0, 0, z_0 + \Delta z)$. Ignore attenuation and multiple reflections.

(a) Assume the scatterers are within the region where the geometric assumption holds and find the estimated reflectivity $\hat{R}(x, y, z)$.

(b) How far should the scatterers be separated in order for the imaging system to distinguish them?

(c) If the scatterers are in the far field where the Fraunhofer approximation holds, what would be the estimated reflectivity $\hat{R}(x, y, z)$?

(d) Derive an expression for the depth resolution in the far field as a function of z_0.

11.12 An ultrasound imaging system is equipped with two square transducers. One operates at 5 MHz and the other one operates at 12 MHz. The 5 MHz transducer is 2.0 cm by 2.0 cm and the 12 MHz one is 0.4 cm by 0.4 cm. The imaging system is tested in a medium having a speed of sound of 1,560 m/s at both frequencies. The amplitude attenuation factor of the medium satisfies

$$\mu_a[\text{cm}^{-1}] = 0.04\ \text{cm}^{-1} \cdot \text{MHz}^{-1} \times f[\text{MHz}]\,.$$

We want to image objects at ranges up to 20 cm.

(a) Find the system sensitivities $L_{5\text{MHz}}$ and $L_{12\text{MHz}}$?

(b) What are the widths of the main lobes at the range of 20 cm? (Using the simple 3 dB model of a field pattern.)

(c) What is the maximum pulse repetition rate for the system?

(d) Suppose the transducers were scanned back and forth to obtain a B-mode image. What is the maximum frame rate if 128 lines are used to form an image?

(e) What will happen to the B-mode image if the speed of sound in the medium is nonuniform but we still use a uniform speed assumption?

11.13 A 1 cm × 1 cm square ultrasound transducer, submerged in oil, is centered at the origin and pointed down the $+z$-axis, as shown in Figure P11.5. A large layer of fat begins at $z_0 = 20$ cm, and extends to infinity, so that the only possible echoes are those from the interface between oil and fat. Suppose the transducer operates at 1 MHz and the density, speed of sound, and absorption coefficient of oil and fat are:

$$\rho_{\text{oil}} = 950\,\text{kg/m}^3,\ c_{\text{oil}} = 1,500\,\text{m/s},\ \alpha_{\text{oil}} = 0.95\ \text{dB/cm}\,,$$
$$\rho_{\text{fat}} = 920\,\text{kg/m}^3,\ c_{\text{fat}} = 1,450\,\text{m/s},\ \alpha_{\text{fat}} = 0.63\ \text{dB/cm}\,.$$

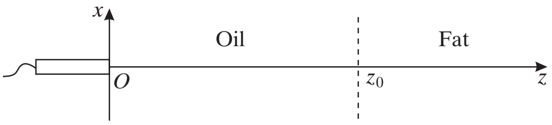


Figure P11.5
See Problem 11.13.

Assume that the transducer is sensitive to at most $L = 65$ dB pressure amplitude loss.

(a) What is the depth of penetration inside the oil?

(b) What is the approximate beamwidth at the given oil/fat interface?

(c) Suppose the peak acoustic pressure of the generated pulse is 20 N/cm^2. Compute the peak pressure of the returning pulse. Is it detectable by the transducer system? Does this result conflict with the answer in part (a)? Explain why or why not.

Now, assume that the interface is not fixed; instead, it is given by $z_0(t) = 20 - 5\cos(2\pi f_0 t)$ cm, where $f_0 = 100$ Hz.

(**d**) Assume that the transducer fires a pulse at $t = 0$. Sketch the A-mode signal. Label the axes carefully, and identify the time-of-return of the returning pulse.

(**e**) Assume that the transducer fires every 10 ms starting from $t = 0$. Sketch the M-mode image that would be generated. Label the axes carefully.

11.14 Consider a single, flat, square, ultrasound transducer whose sides have lengths of 0.5 cm and whose resonant frequency is 2.5 MHz. Suppose it is embedded in a homogeneous medium whose density ρ is 920 kg/m^3 and whose acoustic impedance Z is 1.35×10^6 kg/m^{2}s.

(**a**) What range z_f characterizes the range at which the far field begins for this transducer in this medium?

(**b**) For a typical system (i.e., make appropriate assumption(s) about the system), what does the absorption coefficient α of this medium have to be if the depth of penetration is 20 cm.

(**c**) What is the lateral resolution (FWHM) of this transducer at a range $z = 10$ cm?

(**d**) Assume that this transducer could be scanned (slid) rapidly in the x direction and repetitively pulsed in order to create a rectangular-shaped B-mode image. It is desired to obtain A-mode data separated by 1 mm in the x direction and to scan a 12 cm wide region. Ignoring any potential mechanical problems with scanning the transducer, what is the maximum frame rate that can be achieved for the depth of penetration given in (b)?

(**e**) Suppose there is an interface at range 10 cm, normal to the transducer's axis, and assume that the second medium (beyond 10 cm) has density 1,070 kg/m^3 and acoustic impedance 1.7×10^6 kg/m^{2}s. If the acoustic pressure of the initial pulse has amplitude A_0 and the absorption coefficient is $\alpha = 1.5$ dB/cm in the first medium, what is the amplitude of the reflected acoustic pressure pulse when it arrives at the transducer?

11.15 Consider a phased array ultrasound system operating at 2 MHz and capable of B-mode and M-mode imaging that is located on the starship Prometheus. We are examining a fetus *in utero* (in the uterus of a living woman), which is suspected to be an alien.

(**a**) Calculate the depth of penetration that can be acheived with an 80 dB dynamic range. Assume an absorption coefficient $\alpha = af$ dB cm^{-1}, for $a = 1.0$ dB cm^{-1} MHz^{-1} and f in MHz.

(**b**) What is the maximum frame rate acheivable, for a 256×256 B-mode image? Assume a speed of sound $c = 1,480$ m/s.

(**c**) We notice a quickly moving fetal heart, located at a depth of $z =$ 10 cm and lateral position $x = 5$ cm. We wish to focus the pulse at this location. Assume we have 129 transducer elements, where the ith (i from -64 to 64) element is located at position $x_i = \frac{i}{128}$ cm and range $z = 0$ cm. Calculate an expression for the delay times, τ_i, and find the value of τ_{64}. Let $\tau_0 = 0$. Delays may be negative.

(**d**) An M-mode image is produced from the signal focused on the heart. Calculate maximum the pulse repetition rate for this image. Alien hearts oscillate like a sine wave with frequency 500 beats per minute. Is our M-mode frame rate high enough to sample this periodic signal without introducing aliasing? How about our B-mode frame rate? Explain.

11.16 Consider the transducer array shown in Figure P11.6, which has elements that alternate in frequency between 3 MHz (white) and 6 MHz (grey). Assume that each transducer is only able to transmit and receive at its own frequency. Assume $L = 90$ dB and $\alpha = af$ where $a = 1$ dB/(cm MHz). In the figure below, let $d = h = 0.4$ mm. Assume the speed of sound is 1,540 m/s.

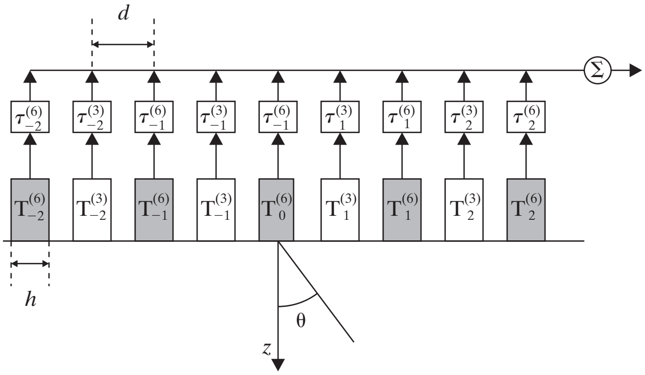


Figure P11.6
See Problem 11.16.

(**a**) What is the depth of penetration for the two transducer types?

Suppose both arrays will fire simultaneously, sending out sound in the direction $\theta = 20°$. The 6-MHz system will focus at a depth of $z = 5$ cm, while the 3-MHz system will focus at a depth of $z = 10$ cm.

(**b**) Assuming that T_0 fires at $t = 0$, write an expression for the transmit delays $\tau_i^{(3)}$ of the 3-MHz system and compute the actual delay $\tau_2^{(3)}$ in μs.

(**c**) Assuming T_0 fires at $t = 0$, write an expression for the transmit delays $\tau_i^{(6)}$ of the 6-MHz system and compute the actual delay $\tau_1^{(6)}$ in μs.

(**d**) Let $A_z^{(6)}$ and $A_z^{(3)}$ be the amplitudes for the 6-MHz and 3-MHz echoes received from range z, assuming a perfect reflection. The displayed A-mode signals will be switched from the 6-MHz echoes to the 3-MHz echoes when $A_z^{(6)}$ reach a signal loss of 30 dB. At what range, z_{switch}, does this "switchover" occur?

(**e**) Find the ratio, $A_0^{(3)}/A_0^{(6)}$, between the transmit amplitudes of the two systems such that $A_z^{(6)} = A_z^{(3)}$ at $z = z_{\text{switch}}$, assuming a perfect reflection.

Harmonic Imaging

11.17 We examine how to use nonlinearities to improve depth of penetration in ultrasound. The idea is to emit a pulse at 2.5 MHz and measure the returning echoes at the second harmonic, 5.0 MHz. Our imaging system has a dynamic range of $L = 80$ dB, and the tissue has an amplitude attenuation factor of $\mu_a = af$ dB cm^{-1} (for $a = \frac{1}{8.7}$ dB MHz^{-1} cm^{-1}).

(a) Assume a tumor is located at depth d. We emit a $f_0 = 2.5$ MHz pulse with amplitude A_0. What is its amplitude A_1 at depth d? Answer in terms of A_0 and d.

(b) Through nonlinear interactions in the tumor, the scattered waveform is transformed into a sawtooth wave with the same amplitude A_1. It can be expressed as

$$g(t) = A_1 \frac{2}{\pi} \sum_{n=1}^{\infty} (-1)^n \frac{1}{n} \sin(2\pi n f_0 t).$$

What is the Fourier transform of this waveform? *Your answer will involve an infinite sum.*

(c) What is the amplitude A_2 of the sinusoidal component at $f_1 = 5$ MHz at the site of the tumor? What is its amplitude A_3 once it has traveled back to the transducer (assume a plane wave)? Answer in terms of A_0 and d.

(d) Find the maximum value of d such that the tumor could still be detected by receiving the 5-MHz signal.

(e) Consider a traditional ultrasound system operating at $f_1 = 5$ MHz. Calculate its depth of penetration.

(f) Design a convolution filter, $h(t)$, which will remove all but the 5-MHz component of the signal $g(t)$. It should eliminate frequencies outside the range of 4–6 MHz, and have unit gain for frequencies within that interval.

This page intentionally left blank

Magnetic Resonance Imaging

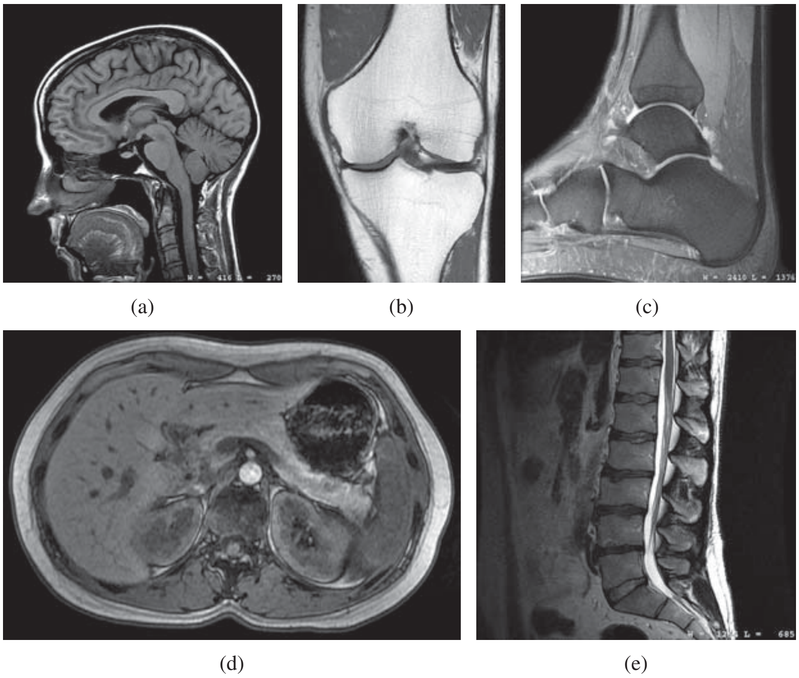
PART V

Overview

In this part of the book, we consider magnetic resonance imaging (MRI). Like computed tomography (CT), MRI produces high-resolution, high-contrast cross-sectional (tomographic) images through the body. Like ultrasound imaging, MRI does not involve ionizing radiation. The combination of high image quality and risk-free imaging has made MRI one of the most widely used medical imaging modalities.

In MRI, the signals arise from the nuclear magnetic resonance properties of tissues; these properties are ‘‘stimulated’’ by the application of fixed magnetic and variable radio frequency fields. Since the *pulse sequences* that govern the time- varying application of these fields can be changed by the operator, MRI—like nuclear medicine—has the potential to create a number of different images representing different underlying signals. Clever combinations of pulse sequences can be used to differentiate fat from water, measure water diffusion, image bone, measure motion, and estimate blood flow. Functional magnetic resonance imaging, or *fMRI*, has become a powerful research tool, especially in understanding brain function and complements its clinical use for anatomic imaging. In addition, paramagnetic ‘‘contrast agents’’ or ‘‘tracers’’ akin to those used in radiography and nuclear medicine are being developed to improve image contrast and measure additional functions. Ultimately, MRI may have the ideal combination of structural and functional imaging capabilities in a risk-free modality.

A typical MRI system looks like a CT system, with a large gantry with a ‘‘tunnel’’ through which the patient, lying on a bed, is inserted. This gantry houses the apparatus for applying the magnetic fields and RF fields, as well as the *receiver coils* from which the raw signal data arise. MRI, like CT, is a cross-sectional or tomographic imaging modality. Image reconstruction, while conceptually equivalent to that in CT, is derived directly from the nature of the raw signals in frequency space.



(a) (b) (c)

(d) (e)

Figure V.1
MR images of (a) the head showing the brain, spinal column, tongue, and vocal tract; (b) the knee; (c) the ankle; (d) the liver; and (e) the lumbar spine. Images (a), (c), (d), and (e) courtesy of GE Healthcare. Image (b) courtesy of Osirix.

In clinical practice, MRI is mainly used to depict anatomy rather than function. It is capable of producing high-contrast cross-sectional images throughout the entire body, as shown in Figure V.1 (see also Figures I.1(d) and I.4(b)). But MRI goes even further by providing different ways to obtain images—through modification of the underlying pulse sequence—in order to obtain different tissue contrasts to better reveal pathology.

Figure V.2 shows several image pairs demonstrating that one pulse sequence might be better at revealing the pathology of the subject than the other. Figure V.2(a) shows a PD-weighted image of the wrist, revealing a small surface tear of the triangular fibrocartilage disc. Figure V.2(b) shows a fat-saturated PD-weighted image showing the same tear and also revealing bone marrow edema and cystic changes of the scaphoid, lunate, and triquetrum. A stroke is hard to see on the T2-weighted image in (c), while the diffusion weighted image in (d) clearly shows it as the bright white region on the left side of the image. This difference is due to the subtle changes in water concentration in the tissues that occur in an early stroke. In a multiple sclerosis patient, only the most apparent white matter lesions can be seen on the T1-weighted MR image shown in (e), while several lesions are visible as bright spots and streaks in the white matter in the fluid-attenuated inversion recovery MR image shown in (f).

Because of the capability to provide anatomic images with high contrast in soft tissues without the use of ionizing radiation or contrast agents, MRI has displaced CT in many clinical applications.

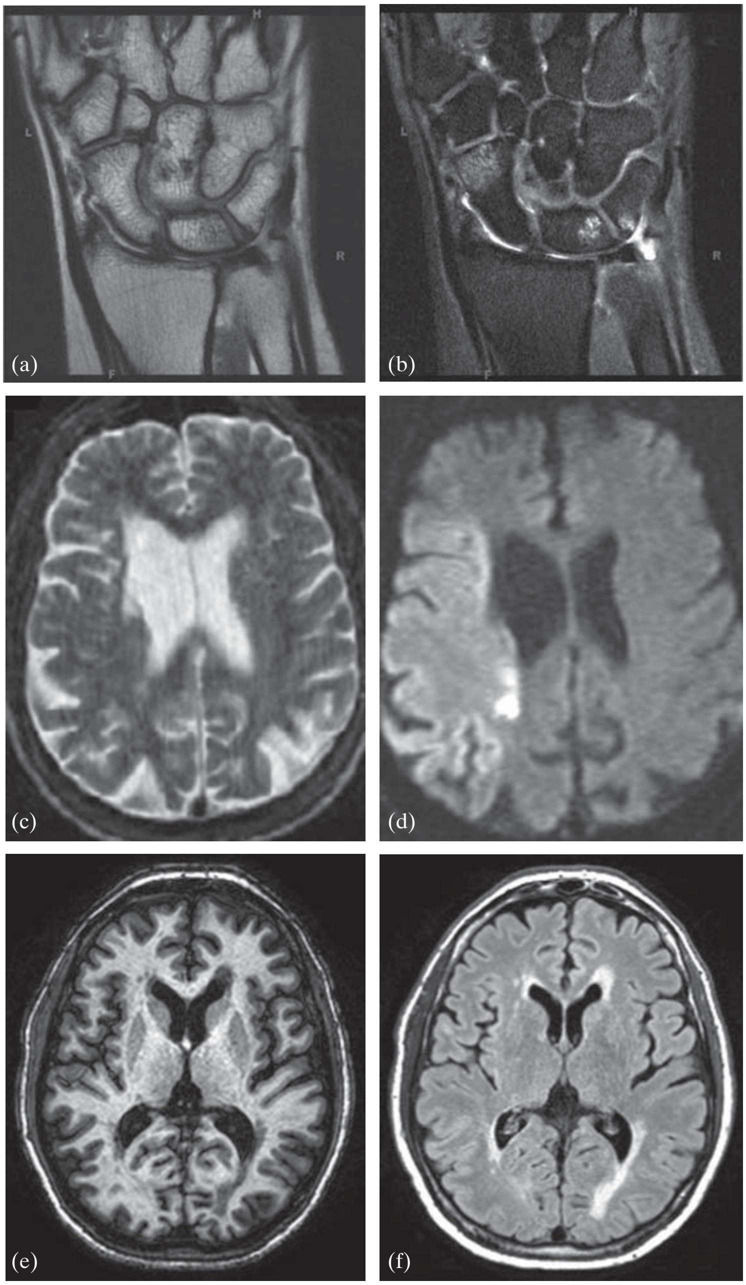


Figure V.2
(a) and (b) show two images of a wrist fracture. (c) and (d) show two images of a stroke. (e) and (f) show two MR images of a multiple sclerosis patient. Images (a) and (b) courtesy of Dr. Avneesh Chhabra, Department of Radiology, Johns Hopkins Hospital. Images (c) and (d) courtesy of Dr. David Yousem, Department of Radiology, Johns Hopkins Hospital.

CHAPTER

12

Physics of Magnetic Resonance

12.1 Introduction

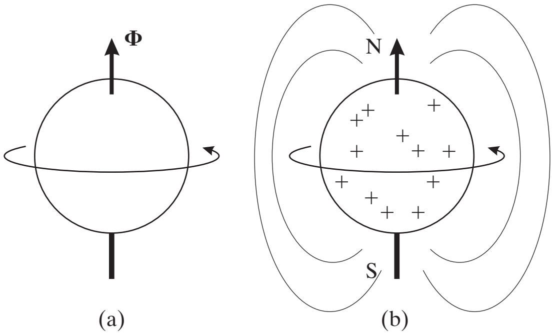
Magnetic resonance imaging (MRI) is made possible by the physical phenomenon called *nuclear magnetic resonance* (NMR). In this chapter, we provide an overview of NMR in order to be able to understand MRI. It turns out that it is not necessary to understand the quantum physics aspects of NMR, although it is informative to appreciate these origins, if for no other reason than to know where to look for a deeper understanding of the physical principles of MRI. We can then rapidly progress from the quantum principles to the macroscopic principles, which can be understood from the principles of classical physics. It is macroscopic *spin systems* and their interaction with electric and magnetic fields that give rise to the signals from which MR images are made. The behavior of these spin systems is described by the so-called *Bloch equations*, which are phenomenological.[1] The Bloch equations tie three important physical properties of these spin systems—proton density, longitudinal relaxation, and transverse relaxation—to the generation of signals that can be measured by an MR scanner and turned into an image, as described in Chapter 13.

12.2 Microscopic Magnetization

Nuclear magnetic resonance is concerned with the nuclei of atoms, but not with radioactivity as in nuclear medicine. Instead, it is the *charge* and *angular momentum* possessed by certain nuclei that are of interest in NMR. We know that all nuclei have positive charges because they are composed of only protons and neutrons. It is also true that a nucleus with either an odd atomic number or an odd mass number has an *angular momentum* $\boldsymbol{\Phi}$. (We note that a bold symbol is used here, and throughout this chapter and the next, to denote a vector quantity.) Said to have *spin*, these nuclei are NMR active. We can visualize such a nucleus as a small ball that is rotating about an axis, as shown in Figure 12.1(a).

---

[1] The term *phenomenological* refers to a process or equation that is based on an observed phenomenon rather than one that is derived from fundamental or "first" principles.



(a) (b)

Figure 12.1
Visualizations of (a) the angular momentum of a nucleus and (b) the microscopic magnetization of a nucleus.

Collections of identical nuclei—regardless of the molecular environment in which they are found—are called *nuclear spin systems*. The nuclei of ^{1}H, ^{13}C, ^{19}F, and ^{31}P are important in NMR because they are prevalent in biological systems and thus each can provide a large enough NMR signal to be detected above the background noise. In whole-body MR imaging of anatomy, however, we are concerned only with the nuclei of ^{1}H. ^{1}H atoms are present in very high density within the body (mainly because of the water content in our bodies), and an ^{1}H nucleus gives a very strong NMR signal. Because the nucleus of the hydrogen atom is a single proton, it is sometimes said that MRI images protons, but this is not strictly true because protons are present in other atoms but are not imaged in conventional MRI.

Each nucleus that has a spin also possesses a microscopic magnetic field. Although the physical principles for this fact can only be rigorously derived using quantum physics, we can use a classical approach to argue for this as well. We envision each nucleus having a positive charge that is spinning around an axis, as shown in Figure 12.1(b). Since a magnetic field is present whenever there are circulating charges (as in a loop of wire), the nucleus must possess a magnetic field. The microscopic magnetic field has a magnetic moment vector $\boldsymbol{\mu}$, given by

$$\boldsymbol{\mu} = \gamma \boldsymbol{\Phi}\,, \tag{12.1}$$

where γ is the *gyromagnetic ratio* and has units of radians per second per tesla. It is useful to define

$$\bar{\gamma} = \frac{\gamma}{2\pi}\,, \tag{12.2}$$

which has units of Hz per tesla. Some gyromagnetic ratios for common nuclei used in NMR and MRI are given in Table 12.1.

In general, there is no preferred orientation for nuclei in a given sample of material. The randomly oriented individual nuclear spins cancel each other out macroscopically, and the resulting spin system has no macroscopic magnetic field. However, spin systems do become macroscopically magnetized when they are placed in a magnetic field, because in this case the microscopic spins tend to align with the applied external magnetic field. This property of nuclear spin systems is termed *nuclear magnetism*.

TABLE 12.1

Common Gyromagnetic Ratios

Nucleus	$\gamma\!\!\!-$ [MHz/T]
^{1}H	42.58
^{13}C	10.71
^{19}F	40.05
^{31}P	11.26

The magnitude and direction of an external magnetic field can be represented as a vector $\mathbf{B}_0$, given by

$$\mathbf{B}_0 = B_0\hat{\mathbf{z}}\,. \tag{12.3}$$

Here, B_0 is the magnitude of the magnetic field in tesla and $\hat{\mathbf{z}}$ is a unit vector pointing in the $+z$ direction within a laboratory (fixed) frame. Intuition suggests that the microscopic spins should all point in the same direction as the applied magnetic field. Quantum physics predicts otherwise. Each nuclear species possesses a *spin quantum number*, I, which takes on nonnegative integer multiples of 1/2. All of the nuclei in Table 12.1, in fact, have $I = 1/2$ and their spins systems are called *spin 1/2 systems*. For spin 1/2 systems in equilibrium, the microscopic magnetization can point in only one of two possible orientations relative to the direction $\hat{\mathbf{z}}$. In particular, they can be 54° off $\hat{\mathbf{z}}$ or $180^\circ - 54^\circ = 126^\circ$ off $\hat{\mathbf{z}}$, with a very slight preference for the 54-degree (the so-called "up") orientation, which is the low energy state. Furthermore, the phase of $\boldsymbol{\mu}$—that is, its orientation around the z-axis—is random. Putting all these facts together, the spin system overall becomes slightly magnetized in the $\hat{\mathbf{z}}$ direction. This gives rise to the notion of "bulk" or macroscopic magnetization, which we now study in detail in order to understand how MR imaging works.

12.3 Macroscopic Magnetization

Consider a specific spin system (e.g., the nuclei of ^{1}H) within a volume of material. When an external static magnetic field $\mathbf{B}_0$ is applied, the spin system becomes magnetized. This can be modeled using a bulk *magnetization vector* $\mathbf{M}$, which is really just the sum of a large number N_s of individual nuclear magnetic moments

$$\mathbf{M} = \sum_{n=1}^{N_s} \mu_n\,.$$

This situation is shown in Figure 12.2. If the sample is left undisturbed in the static field $\mathbf{B}_0$ for a long period of time,[2] $\mathbf{M}$ will reach an equilibrium value

---

[2] The definition of "long" and "short" times will become clear after we look at relaxation processes.

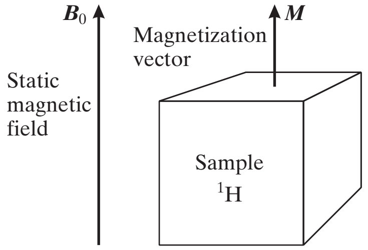


Figure 12.2
The magnetization vector **M**.

$\mathbf{M}_0$ that is parallel with the direction of $\mathbf{B}_0$ and has a magnitude M_0 [which is dependent only on its spatial position $\mathbf{r} = (x, y, z)$]. The magnitude of $\mathbf{M}_0$ is given by

$$M_0 = \frac{B_0 \gamma^2 h^2}{4kT} P_D\,, \tag{12.4}$$

where k is Boltzmann's constant, h is Planck's constant, T is temperature (from absolute zero), and P_D is the so-called *proton density*—that is, the number of targeted nuclei per unit volume. In whole-body imaging, the hydrogen nucleus is just a proton, which accounts for the name *proton density*. We are going to see that if M_0 is larger, the NMR signal is larger. Therefore, both larger magnets and higher proton densities produce larger NMR signals. When we learn about spatial encoding for MRI in Chapter 13, we will see that spatial differences in P_D give rise to signal differences, which in part accounts for image contrast.

It is of fundamental importance in NMR and MRI that **M** is a function of time. In addition to the changes in this vector as it approaches equilibrium within a magnet, it turns out that **M** can be manipulated in a spatially dependent fashion by using external radio frequency (RF) excitations and magnetic fields. Thus, it should be understood that $\mathbf{M} = \mathbf{M}(\mathbf{r}, t)$, where $\mathbf{r}$ is a three-dimensional laboratory coordinate and t is time. In analogy to the microscopic angular momentum studied at the nuclear scale, there is also a bulk *angular momentum* **J** corresponding to the sample, which is related to the sample's magnetization by

$$\mathbf{M} = \gamma \mathbf{J}\,. \tag{12.5}$$

In MR imaging, the "sample" is a small volume of tissue—that is, a voxel. The value of an MR image at a given voxel is determined by two dominant factors: the tissue properties and the scanner imaging protocol. There are several tissue properties that affect the appearance of an MR image, including two relaxation parameters T_1 and T_2, which we will define below, and the proton density. The second dominant factor is the way in which the vector **M** is manipulated by the scanner hardware and software using a so-called *pulse sequence*. As we shall see, the pulse sequence manipulates the magnetic field within the field of view in order to manipulate **M** over space and time. For now, we focus on a particular sample (voxel) and study how $\mathbf{M} = \mathbf{M}(t)$ changes with time. The so-called *equations of motion* for $\mathbf{M}(t)$ are based on the *Bloch equations*, which we now begin to develop.

12.4 Precession and Larmor Frequency

Because $\mathbf{M}(t)$ is a magnetic moment, it experiences a torque when an external, time-varying magnetic field $\mathbf{B}(t)$ is applied. The equation describing this relationship is

$$\frac{d\mathbf{J}(t)}{dt} = \mathbf{M}(t) \times \mathbf{B}(t)\,, \tag{12.6}$$

where $\mathbf{J}$ is the angular momentum vector associated with $\mathbf{M}$. Note that the symbol × denotes the cross product of two vectors. Using (12.5) to eliminate $\mathbf{J}$ in (12.6) yields

$$\frac{d\mathbf{M}(t)}{dt} = \gamma \mathbf{M}(t) \times \mathbf{B}(t)\,, \tag{12.7}$$

which is valid over a short period of time, where "short" will soon be defined.

Now suppose that $\mathbf{B}(t)$ is a static magnetic field oriented in the z direction —that is, $\mathbf{B}(t) = \mathbf{B}_0$. If the initial magnetization vector $\mathbf{M}(0)$ was oriented at an angle α relative to the z-axis, then the solution of (12.7) is

$$M_x(t) = M_0 \sin\alpha \cos(-\gamma B_0 t + \phi)\,, \tag{12.8a}$$

$$M_y(t) = M_0 \sin\alpha \sin(-\gamma B_0 t + \phi)\,, \tag{12.8b}$$

$$M_z(t) = M_0 \cos\alpha\,, \tag{12.8c}$$

where

$$M_0 = |\mathbf{M}(0)|\,, \tag{12.9}$$

$\mathbf{M}(t) = (M_x(t), M_y(t), M_z(t))$, and ϕ is an arbitrary angle. These equations describe a *precession* of $\mathbf{M}(t)$ around $\mathbf{B}_0$ with a frequency

$$\omega_0 = \gamma B_0\,, \tag{12.10}$$

which is known as the *Larmor frequency*.

As expressed in (12.10), the Larmor frequency has units of radians per second. In this book, we have expressed frequencies in cyclic rather than radial units. Therefore, we define the equivalent Larmor frequency

$$\nu_0 = \bar{\gamma} B_0\,, \tag{12.11}$$

which has units of cycles per second or hertz. Using (12.11) and (12.2), the equations of precession in (12.8a) can be written as

$$M_x(t) = M_0 \sin\alpha \cos(-2\pi\nu_0 t + \phi)\,, \tag{12.12a}$$

$$M_y(t) = M_0 \sin\alpha \sin(-2\pi\nu_0 t + \phi)\,, \tag{12.12b}$$

$$M_z(t) = M_0 \cos\alpha\,. \tag{12.12c}$$

Precession is an important concept in NMR and MRI. Most people have seen precession in a child's top or dreidel. A spinning top typically begins with the axis of the top, which is also its axis of rotation, pointing vertically.

As the top slows, its spin axis begins to rotate about the vertical axis—this is precession. In NMR, $\mathbf{M}(t)$ represents the axis of the top; the direction of the static magnetic field—$\hat{\mathbf{z}}$, in this case—represents the vertical axis. This situation is shown schematically in Figure 12.3. Notice that precession occurs in the *clockwise* direction when viewed opposite the direction of the static magnetic field.

It might be assumed that the Larmor frequency is constant for a given spin system (e.g., ^{1}H) within a sample. After all, the gyromagnetic ratio is constant for a given spin system, and B_0 is supposed to be a constant. In reality, B_0 cannot be exactly constant. There are three sources of B_0 fluctuations of concern in MRI: magnetic field inhomogeneities, magnetic susceptibility, and chemical shift. Magnet design and calibration is a very important issue in MRI. We shall see in the following chapter that *shimming* the main magnet to produce a very homogeneous field is critical to image quality. Generally, the main field inhomogeneity is kept to a few parts per million (ppm) and is slowly varying over the field of view, so it can be either ignored or compensated by postprocessing.

Magnetic susceptibility refers to the material property that decreases or increases the magnetic field within the material relative to the surrounding field. *Diamagnetic* materials slightly decrease the field, *paramagnetic* materials slightly increase the field, and *ferromagnetic* materials strongly increase the field. The magnetic field in the presence of susceptibility can be modeled as

$$\hat{B}_0 = B_0(1 + \chi)\,, \tag{12.13}$$

where χ is a spatially variable quantity called the *diamagnetic susceptibility*. Both carbon and molecular oxygen (O_2) are diamagnetic, and as a result the body is mostly diamagnetic. This means that the magnetic field within the body is slightly lower than the field outside, and there are sudden changes in the field when moving from inside to outside the body. These sudden changes can have a deleterious effect on images because our assumptions leading to image equations assume a constant magnetic field. But differences in susceptibility due to natural biological events, such as iron deposits in the brain or removal of oxygen from blood hemoglobin due to neuronal activity, can also be imaged. Thus, like many intrinsic tissue properties, the presence of differences in tissue susceptibility can be either suppressed or emphasized in NMR or MRI acquisitions.

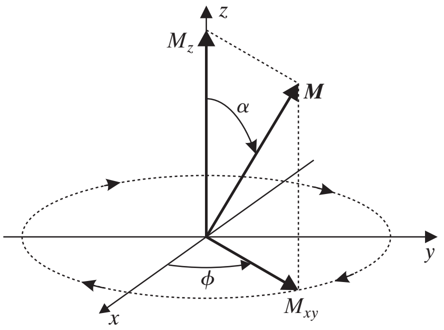


Figure 12.3
The magnetization vector $\mathbf{M}$ precesses about the z-axis.

Chemical shift is a measure of the change in Larmor frequency due to the chemical environment of the imaged nucleus—for example, the molecules to which the ^{1}H atom is attached. In the chemical shift phenomenon, the local electron cloud can shield the nucleus from the full effects of the main magnetic field. Unlike susceptibility, chemical shift is a very local property since it is tied to the molecular environment of a nucleus; the spatial impact of chemical shift is therefore tied to the locations and concentrations of the particular molecule causing the effect. Chemical shift can be modeled as a change in the magnitude of the magnetic field, by

$$\hat{B}_0 = B_0(1 - \varsigma)\,, \tag{12.14}$$

where ς is the *shielding constant* and is a measure of the degree of chemical shift. The corresponding shifted Larmor frequency is given by

$$\hat{\nu}_0 = \nu_0(1 - \varsigma)\,. \tag{12.15}$$

Chemical shift is usually expressed in ppm:

$$\varsigma(\text{ppm}) = \varsigma \times 10^6\,. \tag{12.16}$$

For example, the hydrogen nuclei in fat (CH_2) are shifted down by 3.35 ppm from those in water (H_2O). This corresponds to a shift of the Larmor frequency by −214 Hz at 1.5 T.

Groups of nuclei within a given spin system that have the same Larmor frequency (including chemical shift effects) are called *isochromats*. For example, the hydrogen nuclei in water form an isochromat, while those in fat form another isochromat. The importance of isochromats for a given spin system is that they have only very slightly different Larmor frequencies. The impact of this fact on imaging will become clear as we proceed.

12.5 Transverse and Longitudinal Magnetization

In order to obtain a deeper understanding of MRI, it is necessary to conceptualize the magnetization vector $\mathbf{M}(t)$ as having *two* components. The first is the *longitudinal component*, which is oriented along the axis defined by the static magnetic field. The second is the *transverse component*, which is oriented in a plane orthogonal to the direction of the static magnetic field. The *longitudinal magnetization* is given by $M_z(t)$—simply the z component of $\mathbf{M}(t)$. The *transverse magnetization* is defined by

$$M_{xy}(t) = M_x(t) + jM_y(t)\,, \tag{12.17}$$

which captures the two orthogonal components $M_x(t)$ and $M_y(t)$ as one complex quantity. The angle of the complex number M_{xy}, given by

$$\phi = \tan^{-1}\frac{M_y}{M_x}\,, \tag{12.18}$$

is called the *phase* or *phase angle* of the transverse magnetization. Figure 12.3 illustrates this concept as a projection of the magnetization vector $\mathbf{M}(t)$ against the z-axis and the x-y plane, respectively. The three Cartesian components of $\mathbf{M}(t)$ are completely represented by just these two components, one of which is complex. Thus, understanding the evolution of $M_{xy}(t)$ and $M_z(t)$ yields, equivalently, an understanding of the temporal evolution of the three spatial components of $\mathbf{M}(t)$.

12.5.1 NMR Signals

Using (12.12) and (12.17), the transverse magnetization can be written as

$$M_{xy}(t) = M_0 \sin \alpha e^{-j(2\pi \nu_0 t - \phi)} \,. \tag{12.19}$$

The origin of the *observed signal* in MRI can now be understood. The rapidly rotating transverse magnetization creates an RF excitation within the sample. This RF excitation will in turn induce a voltage—a measurable signal—in a coil of wire located outside the sample. This signal is recorded for use in MRI. It is a common misconception that MRI uses *radio waves* to image the body. This is not true. In fact, it is *Faraday induction* that is used both to manipulate nuclear spin systems and to generate signals from active NMR samples. Thus, while it is true that the signal frequencies in whole-body imaging are in the RF frequency band, radio waves themselves play very little role in MR imaging. Radio waves are generated by the coils used to induce NMR and they will penetrate the body. Their frequencies are well below that of ionizing radiation, however, and their only effect on the body is to cause slight heating.

Although it is common to refer to the transverse magnetization as the "signal," in order to actually receive a voltage signal from these rotating spins, there must be a coil of wire—a so-called *RF coil*—close to the sample. In Chapter 13, we will study various RF coil designs that are used in imaging; for now, we simply need to understand the abstract relationships.

Faraday's law of induction states that a time-varying magnetic field cutting across a coil of wire will induce a voltage in the wire. We can determine the amplitude of the induced voltage using the *principle of reciprocity*—that is, we first look at the magnetic field produced at a point in space as a result of a current flowing in the coil of wire. Specifically, suppose that the magnetic field at $\mathbf{r}$ produced by a unit direct current in the coil is given by $\mathbf{B}^r(\mathbf{r})$. Now reverse the scenario: suppose that there exists a time-varying magnetic field $\mathbf{M}(\mathbf{r}, t)$ (our magnetic moment) throughout the object. Then, the voltage induced in the coil is given by

$$V(t) = -\frac{\partial}{\partial t} \int_{\text{object}} \mathbf{M}(\mathbf{r}, t) \cdot \mathbf{B}^r(\mathbf{r}) \, d\mathbf{r} \,, \tag{12.20}$$

where $\cdot$ represents the dot (or inner) product.

To find a simpler expression for the induced voltage in an NMR experiment, we first assume that the object is homogeneous—that is, $\mathbf{M}(\mathbf{r}, t) = \mathbf{M}(t)$. We also assume that the coil produces a uniform field when excited, so $\mathbf{B}^r(\mathbf{r}) = \mathbf{B}^r$. We also note that the z component of magnetization $M_z(t)$ is a slowly changing quantity (we will see why later); hence, its temporal derivative is small and can be ignored. Using these simplifications in (12.20) and breaking each field into its component representation yields

$$V(t) = -\frac{\partial}{\partial t}\int_{\text{object}} M_x(t)B_x^r + M_y(t)B_y^r \, d\mathbf{r} \tag{12.21a}$$

$$= -V_s \frac{\partial}{\partial t}\left[M_x(t)B_x^r + M_y(t)B_y^r\right], \tag{12.21b}$$

where V_s is the volume of the sample.

The transverse magnetization components are found as the real and imaginary parts of $M_{xy}(t)$ in (12.19),

$$M_x(t) = M_0 \sin\alpha \, \cos(-2\pi\nu_0 t + \phi), \tag{12.22a}$$

$$M_y(t) = M_0 \sin\alpha \, \sin(-2\pi\nu_0 t + \phi). \tag{12.22b}$$

Using these expressions in (12.21) yields

$$V(t) = -2\pi\nu_0 V_s M_0 \sin\alpha \left[B_x^r \sin(-2\pi\nu_0 t + \phi) - B_y^r \cos(-2\pi\nu_0 t + \phi)\right]. \tag{12.23}$$

Writing the components of the reference field as

$$B_x^r = B^r \cos\theta_r, \tag{12.24a}$$

$$B_y^r = B^r \sin\theta_r, \tag{12.24b}$$

and further simplifying (using a trigonometric identity) yields

$$V(t) = -2\pi\nu_0 V_s M_0 \sin\alpha \, B^r \sin(-2\pi\nu_0 t + \phi - \theta_r). \tag{12.25}$$

This shows that the basic NMR signal is a sinusoid at the Larmor frequency.

The magnitude of the NMR signal in (12.25),

$$|V| = 2\pi\nu_0 V_s M_0 \sin\alpha \, B^r, \tag{12.26}$$

is of central interest, since we would ordinarily want to maximize it. Recall from (12.4) that M_0 is proportional to B_0, and ν_0 is also proportional to B_0. Therefore, (12.26) reveals that signal strength is proportional to B_0^2, which means that higher field magnets will yield larger signals. Current whole-body scanners use field strengths of 1.5 and 3.0 tesla, but scanners with fields up to 7 tesla are available and approved for human use by the FDA, and these scanners yield much higher signal strengths. We see also that the maximum signal is produced when $\alpha = \pi/2$. The angle α is called the *tip angle* (or sometimes the flip angle) and is something we can control when acquiring MR data. We will find that it is sometimes beneficial to use smaller tip angles in order to image faster, as will be discussed in the next chapter, but this practice comes at the expense of signal strength. The last factor of interest is the sample volume V_s, a quantity that is, to some degree, under our control. In NMR, we can derive a larger signal by putting more of the substance under study into the scanner. We will discover in

MRI that V_s represents *voxel size*, and that larger signal strength is achieved by having larger voxels.

EXAMPLE 12.1

From (12.26), we can see that in order to improve the resolution while keeping the signal strength at the same level, we need to increase B_0. Now consider a scanner with $B_0 = 1.5$ tesla.

Question If we want to double the resolution in all three dimensions, what B_0 should we use to keep the signal strength unchanged?

Answer By doubling the resolution in all three dimensions, we halve the voxel size in all directions. This change reduces the volume of a voxel by a factor of $1/2^3 = 1/8$. From (12.26), we know that in order to keep the signal strength at the same level, we must have B_0 increased by a factor of $\sqrt{8}$. So, in order to double the resolution, we need to use

$$B_0 = 1.5 \times \sqrt{8} = 4.24 \text{ tesla.}$$

■

12.5.2 Rotating Frame

It is sometimes convenient to express and visualize the evolution of the magnetization vector in a frame of reference, called the *rotating frame*, that is rotating at the Larmor frequency ν_0. The coordinates in the rotating frame are related to those in the stationary frame by

$$x' = x\cos(2\pi\nu_0 t) - y\sin(2\pi\nu_0 t)\,, \tag{12.27a}$$

$$y' = x\sin(2\pi\nu_0 t) + y\cos(2\pi\nu_0 t)\,, \tag{12.27b}$$

$$z' = z\,. \tag{12.27c}$$

In this frame of reference, (12.19) becomes

$$M_{x'y'}(t) = M_0 \sin\alpha e^{j\phi}\,. \tag{12.28}$$

In other words, in the rotating frame $M_{x'y'}$ is a *stationary vector* in the complex plane with magnitude $M_0 \sin\alpha$ and phase angle ϕ.

We now look at the method used to tip $\mathbf{M}(t)$ away from $\mathbf{B}_0$ in order to elicit an NMR signal.

12.6 RF Excitation

We have seen that the magnetization vector $\mathbf{M}$ will precess if it is initially oriented away from $\mathbf{B}_0$, and the transverse component of this precession will induce a current in an antenna surrounding the sample. By putting RF current through an antenna surrounding the sample, the spin system can be deliberately excited and thereby control the behavior of $\mathbf{M}$. For example, this is how we can produce magnetization vectors that are not parallel to $\mathbf{B}_0$. In particular, we can excite the spin system using RF signals so that the stimulated system will in turn induce RF signals as output.

Equation (12.7) can be used to understand how RF excitation will allow us to control the magnetization vector. Consider a system in equilibrium (so that $\mathbf{M}(t)$ lines up with $\mathbf{B}_0$). If a small magnetic field $\mathbf{B}_1 = B_1\hat{\mathbf{x}}$ oriented in the x direction is turned on (adding to the main field), then (12.7) predicts a small motion of $\mathbf{M}(t)$ in the $+y$ direction. One understands this as an incremental precession around the x-axis due to the presence of the additional field $\mathbf{B}_1$. This is the beginning of a process that can lead, ultimately, to the magnetization vector $\mathbf{M}(t)$ resting on the transverse plane and to precess about the z-axis once $\mathbf{B}_1$ is turned off.

However, there is a difficulty with the above approach to spin system excitation. As soon as the incremental precession around the x-axis begins, the spin system is no longer in equilibrium, and the magnetization vector $\mathbf{M}(t)$ begins to precess (also) around the z-axis, which has a much larger field strength and therefore a much larger rate of precession. Once that precession begins, an analysis of the location of $\mathbf{M}(t)$ (see Problem 12.4) reveals that the desired effect of $\mathbf{B}_1$ is lost. In fact, if we are to continually push $\mathbf{M}(t)$ toward the transverse plane (the desired goal in most instances), we will need to track the position of $\mathbf{M}(t)$ as it precesses around the z-axis, and apply a $\mathbf{B}_1$ field whose orientation will produce the correct motion, in accordance with (12.7).

Since $\mathbf{M}(t)$ is precessing at the Larmor frequency, the first step in tracking the position of $\mathbf{M}(t)$ is to apply the $\mathbf{B}_1$ field at the Larmor frequency rather than keeping it constant. In this way, the excitation becomes an *RF excitation*. This has a desirable effect, since the precessing vector $\mathbf{M}(t)$ is pushed down toward the transverse plane whenever it coincides with the $\pm y$-axis (see Problem 12.4). Such an RF excitation is said to be *linearly polarized* because its $\mathbf{B}_1$ field is oriented along only one linear axis. It is possible to improve this approach by adding another RF field oriented in the y direction (still orthogonal to the main field). By applying a quadrature excitation (sine instead of cosine) to this y-oriented B_1 field, the magnetization vector is continuously pushed toward the transverse plane. Such an RF excitation is said to be *circularly polarized* since the direction of the $\mathbf{B}_1$ field traces out a circle in the transverse plane.

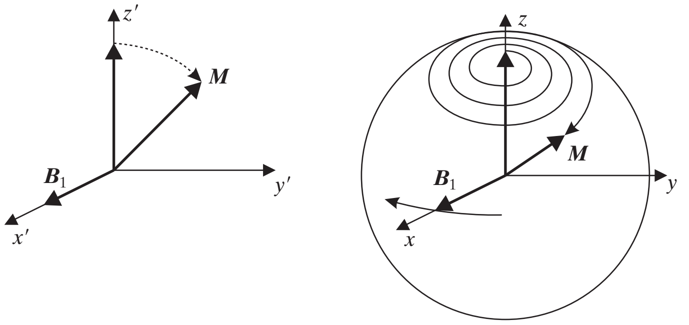
Circularly polarized RF excitations are produced using quadrature RF coils (see Chapter 13), which are in common use today. The circularly polarized RF field can be modeled as a complex magnetic field in the transverse plane

$$B_1(t) = B_1^e(t)e^{-j(2\pi\nu_0 t-\varphi)}\,, \tag{12.29}$$

where $B_1^e(t)$ is the envelope of $B_1(t)$ and φ is its initial phase. The simplest envelope is a rectangular pulse with amplitude B_1 and duration τ_p, yielding a simple RF burst. In the rotating frame, the RF field is given by

$$B_1(t) = B_1^e(t)e^{j\varphi}\,. \tag{12.30}$$

For simplicity, we will assume in the following that $\varphi = 0$, so that B_1 is oriented in the x' direction of the rotating frame. This excitation causes $\mathbf{M}$ to precess in the clockwise direction in the y'-z' plane (in analogy to the precession we observed earlier for the static $\mathbf{B}_0$ field), as shown in Figure 12.4(a). This type of motion of $\mathbf{M}$ is called *forced precession* because it occurs in response to a deposition of RF energy that comes from outside the sample. The frequency of



(a) (b)

Figure 12.4
(a) In the rotating frame, the magnetization vector **M** precesses around the x'-axis in response to an RF excitation. (b) In the laboratory frame, the actual motion of **M** is a spiral.

this precession is given by $\nu_1 = \bar{\gamma} B_1$, where $B_1 = |B_1^e(t)|$. Of course, the actual precession must include the rotation in the x-y plane. The actual evolution of $\mathbf{M}(t)$ is a spiral from the z-axis toward the x-y plane in a clockwise orientation when viewed from the $+z$-axis, as shown in Figure 12.4(b).

The final tip angle and phase of **M** depend on both the amplitude and duration of $B_1^e(t)$. If we turn off the RF after **M** has precessed down into the transverse plane, then the pulse is called a $\pi/2$ (pronounced "pi over two") pulse. A $\pi/2$ pulse is commonly used as an *excitation pulse* because it elicits the maximum signal from a sample that is in equilibrium. An RF pulse twice as long as this is called a π pulse; it will place **M** along the $-z$-axis. This is often called an *inversion pulse* because it inverts the orientation of **M**. In general, the final tip angle after an RF excitation pulse of duration τ_p is given by

$$\alpha = \gamma \int_0^{\tau_p} B_1^e(t)\, dt\,. \tag{12.31}$$

For the special case of a rectangular pulse, **M** is tipped through an angle

$$\alpha = \gamma B_1 \tau_p\,. \tag{12.32}$$

An RF excitation pulse that changes the tip angle by α is referred to as an *α pulse*.

EXAMPLE 12.2

We apply an RF pulse to a sample of protons. The sample is in equilibrium with the B_0 field in the $+z$ direction.

Question We need to tip the magnetization vector **M** into the x-y plane in 3 ms. What should the strength of RF excitation be?

Answer We need a $\pi/2$ pulse to tip the magnetization vector **M** aligned in z direction into the x-y plane. The gyromagnetic ratio for protons is $\bar{\gamma} = 42.58$ MHz/T. The tip angle is

$$\alpha = \pi/2 = 2\pi\bar{\gamma} B_1 \tau_p = 2\pi \times 42.58\,\text{MHz/T} \times B_1 \times 3\,\text{ms}\,.$$

So the strength of the RF excitation should be

$$B_1 = 1.96 \times 10^{-6}\,\text{tesla} = 0.0196\,\text{gauss}.$$

■

12.7 Relaxation

After application of an α pulse, and assuming that $\alpha \neq \pi$, $\mathbf{M}$ will precess in response to the presence of the main magnetic field $\mathbf{B}_0$, as described earlier. According to (12.7), this precession will never end. If this were true, then RF waves would emanate from the sample and could be detected forever as a sinusoidal voltage in an external antenna. Obviously, this situation cannot be true; there must be some mechanism to dampen this (otherwise perpetual) oscillation. In fact, there are two independent relaxation processes that together cause the received signal to vanish: longitudinal relaxation and transverse relaxation. We now describe each of these in detail.

Transverse relaxation acts first to cause the received signal to decay. Also known as *spin-spin relaxation*, transverse relaxation is caused by perturbations in the magnetic field due to other spins that are nearby. This interaction, heavily influenced by random microscopic motion, causes spins to momentarily speed up or slow down, changing their phases relative to other nearby spins, as illustrated in Figure 12.5. This *dephasing* causes a loss of coherence of the RF wave produced by the spin system and a concomitant loss of signal in the receiver antenna. The resultant signal is known as a *free induction decay* (FID), as illustrated in Figure 12.6.

The signal decay in an FID is well modeled as an exponential decay. The time constant of this decay—called the *transverse relaxation time*—is given the symbol T_2 and, like a conventional time constant, has units of time. Accordingly, rather than modeling the transverse magnetization as a never-ending complex exponential as in (12.19), a more accurate representation is given by

$$M_{xy}(t) = M_0 \sin\alpha e^{-j(2\pi\nu_0 t - \phi)} e^{-t/T_2} . \tag{12.33}$$

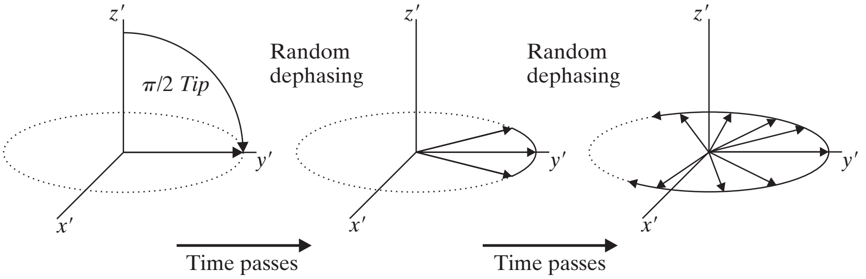


Figure 12.5
Dephasing caused by transverse relaxation.

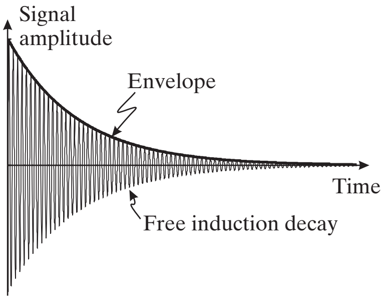


Figure 12.6
A free induction decay. The signal decay is caused by transverse relaxation.

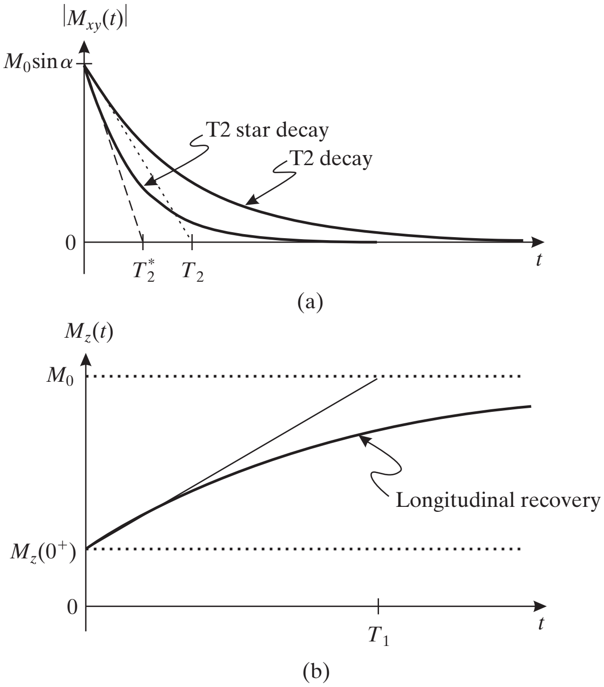


Figure 12.7
(a) Transverse and (b) longitudinal relaxation.

T_2 is generally different for various types of tissues and is one of the important physical properties of the body that gives rise to contrast in MR images, as we shall see. The decay of signal strength $|M_{xy}(t)|$ due to T_2 relaxation is shown graphically in Figure 12.7(a).

After this careful presentation of T_2, we are now going to "pull a fast one on you," and reveal that the received signal actually decays more rapidly than T_2. In fact, local perturbations in the static field B_0 cause the received signal to decay exponentially with a time constant T_2^* (pronounced "tee two star"), which satisfies $T_2^* < T_2$. This situation is illustrated in Figure 12.7(a). It is useful to model the decay associated with these "external" field effects using a time constant T_2'. The relationship between the three transverse relaxation constants is

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \,. \tag{12.34}$$

It is natural to ask why we need the concept of T_2 if it does not fit reality. We will find that although the initial signal decays with T_2^*, there remains an underlying magnetization coherence that lasts longer—in fact, it decays with time constant T_2. This is consistent with the fact that the relaxation due to T_2' effects is reversible. In making images, we will crucially depend on *refocusing* this latent coherence through the concept of *spin echoes*. So, although the initial signal disappears after a few T_2^*'s, there is an underlying signal potential lasting a few T_2's. We describe how to exploit this potential in Section 12.9.

The second relaxation mechanism that causes a loss of signal is called *longitudinal relaxation* or *spin-lattice relaxation*. This process concerns the longitudinal magnetization $M_z(t)$, which recovers back to its equilibrium value M_0 as a rising exponential, as depicted in Figure 12.7(b). Although transverse relaxation is a decreasing exponential and longitudinal relaxation is an increasing exponential, both processes lead to a loss of NMR signal.

Suppose that an α pulse ($\alpha \neq \pi$) is applied at $t = 0$. Then, the longitudinal magnetization obeys

$$M_z(t) = M_0(1 - e^{-t/T_1}) + M_z(0^+)e^{-t/T_1} , \tag{12.35}$$

where T_1 is a material property called the *longitudinal relaxation time*. The notation $M_z(0^+)$ refers to the longitudinal magnetization immediately after the α pulse and is given by

$$M_z(0^+) = M_0 \cos\alpha . \tag{12.36}$$

It is implied by this notation that the duration of the α pulse is negligible with respect to the longitudinal relaxation time. This is a good assumption since, typically, α pulses are a few milliseconds long, while longitudinal recovery lasts several hundred milliseconds.

This recovery of longitudinal magnetization is illustrated in Figure 12.7(b). Like T_2, T_1 is generally different for various types of tissue and is also responsible for generating contrast in MR images, as we shall see. For tissues in the body, the relaxation times are in the ranges $250\text{ ms} < T_1 < 2{,}500\text{ ms}$ and $25\text{ ms} < T_2 < 250\text{ ms}$. Usually, $5T_2 \leq T_1 \leq 10T_2$ (although there are some exceptions), and for all materials $T_2 \leq T_1$.

Although we have been using the term "equilibrium" throughout the chapter, our understanding has been limited to some notion of an "unchanging" or "quiescent" system. We can now understand what equilibrium means mathematically. A sample is said to be in *equilibrium* if its longitudinal magnetization $M_z(t)$ is essentially equal to its final value M_0 throughout the sample. In practice, a sample is considered to be in equilibrium if there have been no external excitations for at least $3T_1^{\max}$, where $T_1^{\max}$ is the largest T_1 in the sample. In the next chapter, we will encounter the term *steady-state*; it is worth noting now that this is a different notion than equilibrium. Equilibrium implies that there has been no excitation for some time into the past; steady-state describes a spin system that has been *periodically excited* so that the spin system is undergoing a periodic longitudinal relaxation process.

EXAMPLE 12.3

Suppose a sample is in equilibrium, and a $\pi/2$ pulse is applied.

Question What happens to the longitudinal magnetization of the sample?

Answer The longitudinal magnetization immediately after the α pulse is $M_z(0^+) = M_0 \cos(\pi/2) = 0$. Therefore, (12.35) yields

$$M_z(t) = M_0(1 - e^{-t/T_1}) ,$$

where M_0 is the equilibrium value of longitudinal magnetization. ■

EXAMPLE 12.4

Suppose a sample is in equilibrium, and an (unspecified) α pulse is applied.

Question What are the transverse and longitudinal magnetizations of the sample, expressed in the rotating frame?

Answer Since the sample is initially in equilibrium, the longitudinal magnetization immediately after the α pulse is given by (12.36). Substituting (12.36) into (12.35) yields the longitudinal magnetization

$$M_z(t) = M_0(1 - e^{-t/T_1}) + M_0 \cos\alpha e^{-t/T_1} .$$

This expression is the same whether expressed in the laboratory coordinate frame or the rotating frame. The transverse magnetization, expressed in the rotating frame, is just $M_{xy}(t)$ in (12.33) multiplied by $e^{+j2\pi\nu_0 t}$, which yields

$$M_{x'y'}(t) = M_0 \sin\alpha e^{j\phi} e^{-t/T_2} .$$

■

12.8 The Bloch Equations

Putting together both the *forced* and *relaxation* behavior of a magnetic spin system yields the *Bloch equations*

$$\frac{d\mathbf{M}(t)}{dt} = \gamma \mathbf{M}(t) \times \mathbf{B}(t) - \mathrm{R}\{\mathbf{M}(t) - \mathbf{M}_0\} , \qquad (12.37)$$

which describe the behavior of **M** in the laboratory frame.[3] Here, $\mathbf{B}(t)$ is composed of the static field and the RF field,

$$\mathbf{B}(t) = \mathbf{B}_0 + \mathbf{B}_1(t) ,$$

and R is the *relaxation matrix*, given by

$$\mathrm{R} = \begin{pmatrix} 1/T_2 & 0 & 0 \\ 0 & 1/T_2 & 0 \\ 0 & 0 & 1/T_1 \end{pmatrix} . \qquad (12.38)$$

The Bloch equations are used to construct models of the behavior of magnetization vectors during excitation. From this behavior, the NMR signal can be inferred by computing the transverse magnetization. For most applications, these equations are transformed into the rotating frame of reference (see Problem 12.9).

EXAMPLE 12.5

The Bloch equations describe the behavior of **M** in the laboratory frame.

Question Find the equations for the components of **M** in the x-y plane, and verify that the transverse relaxation after a $\pi/2$ pulse (in the x direction) satisfies the equations.

---

[3] This equation is referred to in the plural because it is a vector equation comprising three scalar equations.

Answer The Bloch equations are given above in (12.37). By expanding the cross product, we have

$$\frac{d}{dt}\begin{pmatrix} M_x(t) \\ M_y(t) \\ M_z(t) \end{pmatrix} = \gamma \begin{pmatrix} M_y(t)B_z(t) - M_z(t)B_y(t) \\ -M_x(t)B_z(t) + M_z(t)B_x(t) \\ M_x(t)B_y(t) - M_y(t)B_x(t) \end{pmatrix} - \begin{pmatrix} \frac{1}{T_2}M_x(t) \\ \frac{1}{T_2}M_y(t) \\ \frac{1}{T_1}(M_z(t) - M_{0z}) \end{pmatrix}.$$

After a $\pi/2$ pulse, the RF field B_1 is shut down, and only B_0 is nonzero. Therefore, $B_x(t) = B_y(t) = 0$, and the equations for $M_x(t)$ and $M_y(t)$ are simpler:

$$\frac{d}{dt}\begin{pmatrix} M_x(t) \\ M_y(t) \end{pmatrix} = \gamma \begin{pmatrix} M_y(t)B_z(t) \\ -M_x(t)B_z(t) \end{pmatrix} - \begin{pmatrix} \frac{1}{T_2}M_x(t) \\ \frac{1}{T_2}M_y(t) \end{pmatrix}.$$

The transverse relaxation after a $\pi/2$ pulse is given by

$$\begin{aligned} M_x(t) &= M_0 \cos[-(2\pi \nu_0 t - \pi/2)]e^{-t/T_2} = -M_0 \sin(2\pi \nu_0 t)e^{-t/T_2}\,, \\ M_y(t) &= M_0 \sin[-(2\pi \nu_0 t - \pi/2)]e^{-t/T_2} = -M_0 \cos(2\pi \nu_0 t)e^{-t/T_2}\,. \end{aligned}$$

The initial phase is $\pi/2$ because the RF pulse is applied in x direction, the magnetization vector M is tipped into y direction. By substituting the relaxation equations into the Bloch equations for $M_x(t)$, we get

$$\begin{aligned} \frac{dM_x(t)}{dt} &= \frac{d}{dt}\left[-M_0 \sin(2\pi \nu_0 t)e^{-t/T_2}\right] \\ &= -2\pi \nu_0 M_0 \cos(2\pi \nu_0 t)e^{-t/T_2} + \frac{1}{T_2}M_0 \sin(2\pi \nu_0 t)e^{-t/T_2} \\ &= \gamma B_0 M_y(t) - \frac{1}{T_2}M_x(t)\,. \end{aligned}$$

Similarly, we have

$$\frac{dM_y(t)}{dt} = -\gamma B_0 M_x(t) - \frac{1}{T_2}M_y(t)\,.$$

■

12.9 Spin Echoes

Pure transverse relaxation, characterized by the time constant T_2, is a random phenomenon. The fact that the FID decays faster, with time constant T_2^*, is due to fixed perturbations in the magnetic field. These fixed perturbations cause the precession of some spins to speed up and others to slow down (relative to the nominal rate of rotation predicted by the Larmor frequency). As described in Section 12.7, in a very brief period nearby spins are largely dephased—that is, they begin to point in different directions in the transverse plane—as illustrated by the top row of Figure 12.8.

The existence of *spin echoes* is due to the fact that the faster (slower) spins, which now lead (lag) the spin system, can be made to lag (lead) the spin system using a short-duration 180° pulse, as shown in Figure 12.8(d). From this new phase position, the fast spins "catch up" and the slow spins "fall

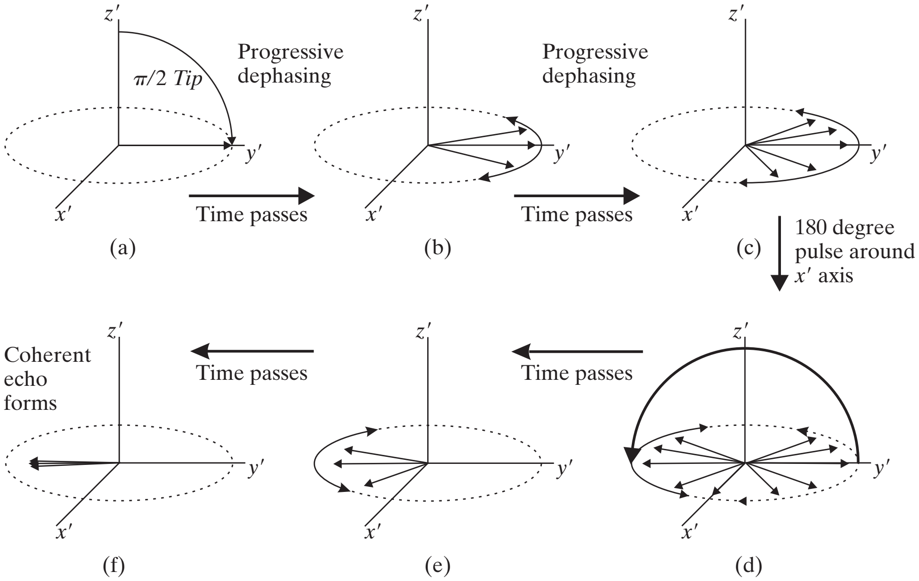


Figure 12.8
Formation of a spin echo.

back," forming a *spin echo*, as shown in the bottom row of Figure 12.8. Thus, a spin echo is the signal that is generated by the transverse spins recovering their coherence due to a deliberate 180° RF pulse, after a loss of coherence due to transverse relaxation processes. The time interval from the initial $\pi/2$ pulse to the formation of the spin echo is known as the *echo time* and is given the symbol T_E. This time is under our control because we specify the application time of the 180° pulse, which is at $T_E/2$.

Until this section, we have not had to consider the timing of NMR signals; we have simply generated an α pulse and watched what happens. In the generation of spin echoes, however, we see that it is the succession of two pulses that matters, and the timing of the second pulse relative to the first determines when the signal (echo) occurs. As we will discover in Section 12.10 and in even greater detail in Chapter 13, it is the type and ordering of excitations and their relative timings that gives rise to different tissue contrasts, as well as a host of other image characteristics such as resolution, noise, and artifacts. In the case of spin echoes, Figure 12.9 shows a simple timing diagram, which depicts a *pulse sequence* that can be followed to generate the echo signal.

There are two mechanisms that combine to make spin echoes decrease in amplitude over time. First, because time elapses during the spin echo process,

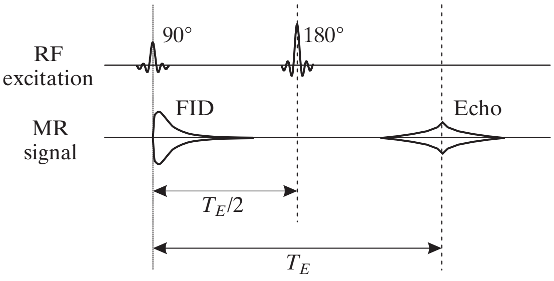


Figure 12.9
Pulse sequence diagram for generating spin echoes.

longitudinal relaxation causes the magnitude of the transverse magnetization to decrease. This idea is evident in Figure 12.8, where the transverse component of the magnetization vector is shown to be getting smaller with time. Second, because of the random effects of transverse relaxation, the phase of the coherent echo is never perfectly aligned, as illustrated in Figure 12.8(f). Since T_2 is often much smaller than T_1, the first effect can often be ignored; therefore, the amplitude of an echo is approximately given by the ideal T_2 exponential decay of the tissue [see (12.33)] at $t = T_E$. This means that we can elicit *multiple echoes* using multiple 180° pulses, and the signal strength of these echoes will decay exponentially with time constant T_2, until they disappear at approximately $3T_2$.

EXAMPLE 12.6

Suppose two ^{1}H isochromats are in different locations in a 1.5 T magnet, and the fractional difference in field strength is 20 ppm.

Question How long will it take before these isochromats are 180° out of phase?

Answer Suppose one sample has field strength B_0 and the other B_0'. The magnitude of the field strength difference can be found as follows:

$$\begin{aligned} B_0' &= B_0(1 - 20 \times 10^{-6}), \\ |B_0' - B_0| &= (20 \times 10^{-6})B_0 . \end{aligned}$$

The difference in Larmor frequency is then

$$\begin{aligned} \Delta\nu &= \gamma\!\!\!^{-}|B_0' - B_0| \\ &= \gamma\!\!\!^{-}(20 \times 10^{-6})B_0 \\ &= 42.58 \text{ MHz/T} \times (20 \times 10^{-6}) \times 1.5 \text{ T} \\ &= 1,277.4 \text{ Hz} . \end{aligned}$$

Since 180° is half a cycle, the time it will take before these isochromats are 180° out of phase is

$$\Delta t = \frac{1/2 \text{ cycle}}{1,277.4 \text{ cycle/s}} = 391\ \mu\text{s} .$$

Question What will be their phase difference at $T_E/2$ if the echo time is 4 ms?

Answer T_E is defined to be the echo time; therefore, $T_E/2 = 2$ ms. The number of cycles occurring in 2 ms is

$$\begin{aligned} \text{Number of cycles} &= 1,277.4 \text{ cycles/s} \times 2 \text{ ms} \\ &= 2.555 \text{ cycles} . \end{aligned}$$

The phase difference is therefore

$$\begin{aligned} \text{Phase difference} &= 2.55 \text{ cycles} \times 2\pi \text{ radians/cycle} \\ &= 16 \text{ radians} . \end{aligned}$$

Thus, after this relatively short amount of time, the two signals have completely lost coherence. ■

12.10 Basic Contrast Mechanisms

We now understand that it is the transverse magnetization $M_{xy}(t)$ that produces the measurable MR signal; the larger its magnitude, the larger the measured signal. If we are to see a contrast between tissues—for example, different image intensities in the gray matter (GM) and white matter (WM) of the brain—the measured signal must be different in those tissues. In MRI, our ability to generate tissue contrast depends on both the intrinsic NMR properties of the tissues—that is, P_D, T_2, and T_1—and the characteristics of the externally applied excitations (i.e., the pulse sequence). So far, we have seen that it is possible to control the *tip angle* α and the *echo time* T_E of the RF excitation. It is also possible to control the interval between successive α pulses; this is the so-called *pulse repetition interval* which is given the symbol T_R. In this section, we consider the response of small volumes containing different tissues to the same external application of RF excitations. We will see how tissue contrast can be manipulated by externally controllable parameters.

Figure 12.10 shows three images of the same slice through a human skull. These clearly show quite different tissue contrasts in the three images. The contrast in the images are classified as (a) P_D-weighted, (b) T_2-weighted, and (c) T_1-weighted. This does not mean that the brightness of these images are proportional to P_D, T_2, and T_1, but merely that the differences in intensity seen between different tissues are largely determined by the differences in P_D, T_2, and T_1, respectively, of the tissues. In the brain, there are two dominant tissue types, GM and WM matter, which are surrounded by cerebrospinal fluid (CSF). The NMR properties of these three tissues at 1.5 T are given in Table 12.2.

P_D-Weighted Contrast In P_D-weighted images, the image intensity should be proportional to the number of hydrogen nuclei in the sample. This weighting can be obtained with the kind of NMR experiment we have already considered. We merely need to start with the sample in equilibrium, apply an excitation RF pulse,

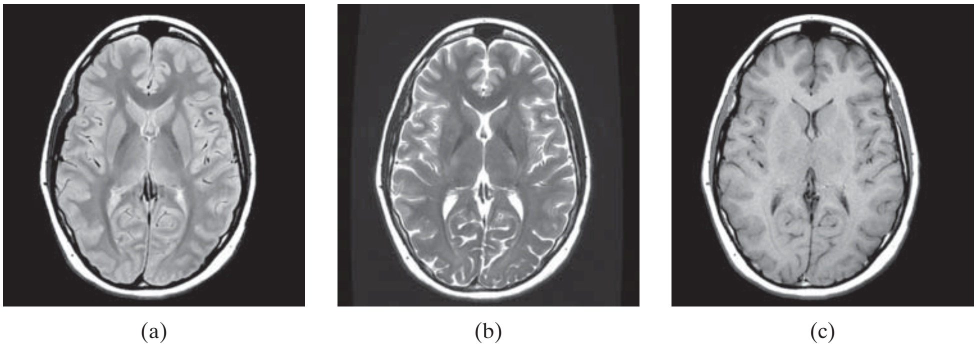

(a) (b) (c)

Figure 12.10
Three images of the same slice through the skull. Contrast between the tissue types are classified as (a) P_D-weighted, (b) T_2-weighted, and (c) T_1-weighted. Courtesy of GE Healthcare.

TABLE 12.2

Brain Tissue Parameters Measured at 1.5T.

Tissue Type	Relative P_D	T_2 (ms)	T_1 (ms)
White matter	0.61	67	510
Gray matter	0.69	77	760
Cerebrospinal fluid	1.00	280	2,650

and image quickly, before the signal has a chance to decay from T_2 effects. Thus, a P_D-weighted contrast can be obtained by using a long T_R (which allows the tissues to be in equilibrium) and either no echo or a short T_E (in order to minimize T_2 decay). In order to obtain the maximum signal, the tip angle should be $\pi/2$. The image shown in Figure 12.10(a) was obtained using $T_R = 6,000$ ms, $T_E = 17$ ms, and $\alpha = \pi/2$. It directly reflects the contrast expected from Table 12.2. In practice, 6,000 ms is an unusually long repetition time. It is generally impractical in MRI to use more than about $T_R = 3,500$ ms because the images take too long to acquire.

T_2-Weighted Contrast To reveal T_2 contrast, differences in the transverse relaxation times of different tissues must be apparent. We know that due to T_2^* effects, the FID decays more rapidly than pure transverse relaxation would predict, and this is not sufficient time to observe T_2 differences between tissues. To obtain T_2-weighted images, therefore, spin echoes must be used. But what should T_E be? We have already seen that if T_E is small, we get P_D-weighting. On the other hand, if T_E is large, the signal strength would be too small to detect above the noise level. In practice, for T_2 contrast, T_E should be selected to be approximately equal to the T_2 values of the tissues being imaged. Figure 12.10(b) is the result of using $T_R = 6,000$ ms, $T_E = 102$ ms, and $\alpha = \pi/2$. The use of a large T_R here is consistent with obtaining maximum signal strength, but its use also reduces intermingling of T_1 contrast (which will become clear in the next section). The appearance of Figure 12.10(b) is consistent with the data in Table 12.2, especially when one realizes that the image intensities are related to T_2 through a decaying exponential. In particular, we realize that GM and WM have completely decayed (greater than three time constants) by the time the CSF has decayed by only about a third of its starting intensity. This means that the GM and WM should have small contrast with respect to each other and large contrast with respect to CSF. WM is slightly darker than GM because its NMR signal has decayed slightly faster.

T_1-Weighted Contrast To obtain T_1-weighted contrast, differences in the longitudinal component of magnetization must be emphasized. This is done by exciting the tissue repeatedly before it has had a chance to fully recover its longitudinal magnetization. We know that if the sample is in equilibrium and excited using an α pulse, then the transverse magnetization is given by

$$M_{xy}(t) = M_0 \sin \alpha e^{-j(2\pi \nu_0 t - \phi)} e^{-t/T_2} \,. \tag{12.39}$$

This (potential) signal dies away with a time constant T_2, and since $T_1 \gg T_2$, it is already negligible when $t \approx T_1$. On the other hand, the longitudinal component, given by

$$M_z(t) = M_0(1 - e^{-t/T_1}) + M_z(0^+)e^{-t/T_1}\,, \tag{12.40}$$

is not negligible when $t \approx T_1$. So, if we set $T_R \approx T_1$ (for some tissue in the field of view), we have created a situation where the transverse component has vanished (and therefore no signal is available, even from an echo), and the longitudinal component has not returned to equilibrium.

Now suppose an α pulse is applied at $t = T_R$, where $T_R \approx T_1$ for some tissue. In this case, (12.39) no longer applies, because the sample was not initially in equilibrium. Instead, the longitudinal magnetization follows (12.40) and has longitudinal magnetization $M_z(T_R)$ just prior to excitation. For generality, we denote the longitudinal magnetization just prior to an RF excitation by $M_z(0^-)$. Then, the transverse magnetization after excitation follows

$$M_{xy}(t) = M_z(0^-)\sin\alpha\, e^{-j(2\pi\nu_0 t-\phi)}e^{-t/T_2} \tag{12.41a}$$

$$= M_{xy}(0^+)e^{-j(2\pi\nu_0 t-\phi)}e^{-t/T_2}\,, \tag{12.41b}$$

where $M_{xy}(0^+)$ denotes the magnitude of the transverse magnetization immediately after excitation.

In this scenario, $M_{xy}(0^+) < M_0 \sin\alpha$ for most tissues, and the exact value of $M_{xy}(0^+)$ depends directly on the longitudinal relaxation time of the particular tissue. This process is illustrated in Figure 12.11 for two tissues, one with a short T_1 (top row) and one with a long T_1 (bottom row). Starting in the left column at an early time when $t > 3T_2$, there is no transverse magnetization, and the longitudinal magnetization is starting to recover. In the middle column, the tissue with the short T_1 has undergone much more recovery than that with the

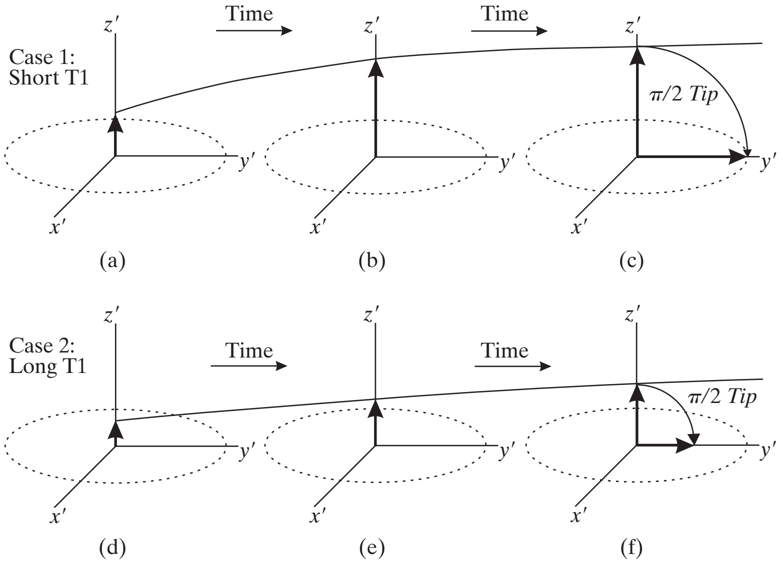


Figure 12.11
The principle behind T_1-weighted contrast.

long T_1. In the right column, a $\pi/2$ pulse tips the longitudinal magnetization into the transverse plane. The signal strength is proportional to the length of the magnetization vector in the transverse plane, and this length is partially determined by the T_1 value of the tissue.

The image shown in Figure 12.10(c) was obtained using $T_R = 600$ ms, $T_E = 17$ ms, and $\alpha = \pi/2$. From the data in Table 12.2, we see that this choice of T_R falls in between the T_1 values for GM and WM but is much smaller than that of CSF. Therefore, both GM and WM will have recovered approximately two-thirds of their longitudinal magnetization, whereas CSF will have recovered relatively little. As a result, the CSF signal should be very small, while that of GM and WM should be larger. This situation is reflected well in Figure 12.10(c), where the GM and WM are relatively bright, while the CSF is dark.

Inversion Recovery We found above that tissues with different T_1 values can be made to yield different signal intensities by controlling the pulse repetition time T_R of a pulse sequence. There is a conceptually simpler way to establish T_1 contrast, that is by applying a 180° RF pulse—which is called an *inversion pulse*—at the beginning of a pulse sequence, as shown in Figure 12.12(a). Previously, we used a 180° pulse to yield a spin echo, but that was for the manipulation of the transverse magnetization, whereas here it is for the manipulation of the longitudinal magnetization.

Consider a system in equilibrium. Application of a $\alpha = 180^\circ$ RF pulse simply inverts the longitudinal magnetization; that is

$$M(0^+) = -M_0\,. \tag{12.42}$$

This is illustrated in Figure 12.12(b). Notice that the transverse magnetization is zero after this pulse is applied (because $\sin\alpha = 0$). Thus, no signal is generated. However, the longitudinal magnetization must make its way back to equilibrium according to the "usual" exponential form given by (12.40). Substituting (12.42)

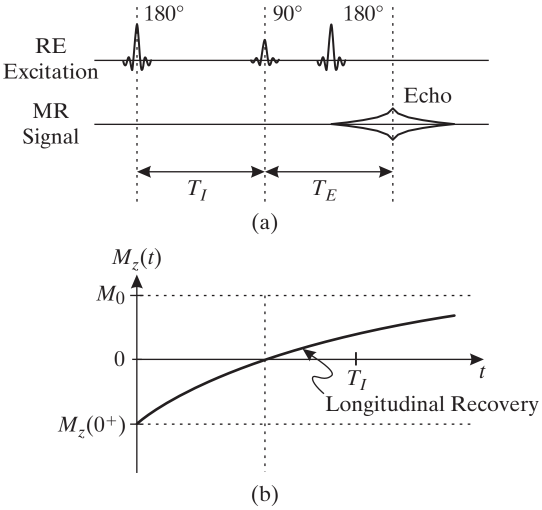


Figure 12.12
(a) Inversion recovery pulse sequence (spin echo version) and (b) time course of longitudinal relaxation of the nulled tissue.

into (12.40) and simplifying yields

$$M_z(t) = M_0(1 - 2e^{t/T_1}), \qquad (12.43)$$

which describes the evolution of longitudinal magnetization from the moment of the inversion pulse. Tissues with different T_1 values will relax at different rates, and this difference can be captured by applying an imaging pulse during the relaxation after the inversion pulse.

Consider a tissue having a longitudinal relaxation time of T_1. Its longitudinal magnetization will be precisely zero at

$$t_{\text{null}} = T_1 \ln 2, \qquad (12.44)$$

which is straightforward to prove by setting $M_z(t) = 0$ in (12.43) and solving for t. If an $\alpha = 90^\circ$ pulse is applied at $T_I = t_{\text{null}}$, where T_I is called the *inversion time*, then this particular tissue will have no transverse relaxation because

$$M_{xy}(t_{\text{null}}^+) = M_z(t_{\text{null}}^-) \sin\alpha = 0\,.$$

This tissue has been "nulled" and will not generate an NMR signal. However, since other tissues with different T_1 values will have nonzero transverse magnetization, they will contribute to the FID and their contrast relative to the nulled tissue is (mathematically) infinite. Figure 12.12(a) shows a spin echo inversion recovery pulse sequence. A spin echo is most commonly used in inversion recovery experiments since T_2^* effects will typically dephase the transverse magnetization by the inversion time T_I, leaving no signal without the spin echo.

The principle of inversion recovery can be applied to measure the T_1 of a given tissue by repeating the experiment with different inversion times T_I. When a particular inversion time, say $t_{I,\text{null}}$, gives a zero signal, then according to (12.44)

$$T_1 = T_{I,\text{null}} / \ln 2 \qquad (12.45)$$

for that tissue. An even more important use of an inversion pulse is to null out certain tissues such as fat (these are called STIR sequences) or fluid (these are called FLAIR sequences). When a given class of tissues, such as fat or fluid, is nulled, then all other tissues have very high contrast with respect to the nulled class, and this can be used to improve tissue discrimination in certain diseases.

12.11 Summary and Key Concepts

Nuclear magnetic resonance is the phenomenon behind magnetic resonance imaging. This phenomenon is well-known in chemistry and is now exploited for medical imaging. In this chapter, we presented the following key concepts that you should now understand:

1. *Magnetic resonance imaging* is based on the phenomenon of *nuclear magnetic resonance*.

2. A nucleus with an odd Z or A has *angular momentum* or *spin*; collections of such nuclei are called *nuclear spin systems*.
3. Such systems become magnetized when an external magnetic field is applied, producing a *bulk magnetization vector* that is a function of time.
4. The so-called *equations of motion* for this magnetization vector are based on the *Bloch equations*.
5. These equations describe a *precession* of the magnetization vector around the external magnetic field direction, with a frequency known as the *Larmor frequency*.
6. An *RF pulse* will cause the magnetization vector to precess around the external magnetic field.
7. The magnetization vector has two components: *transverse magnetization* and *longitudinal magnetization*.
8. The *observed signal* in MRI is an RF pulse produced by the rapidly rotating transverse magnetization.
9. *Relaxation* describes the gradual dampening of the precession (and associated signal); it has two components: *transverse* or *spin-spin relaxation* and *longitudinal* or *spin-lattice relaxation*.
10. *Contrast* in MR images is produced by manipulating the RF excitation pulse sequence to produce MR signals that are influenced by different weighted contributions of P_D, T_1, and T_2.

Further Reading

Bushong, S.C. *Magnetic Resonance Imaging*, 3rd ed. St. Louis, MO: Mosby, 2003.

Callaghan, P.T. *Principles of Nuclear Magnetic Resonance Microscopy*, New York, NY: Oxford University Press, 1993.

Haacke, E.M., Brown, R.W., Thompson, M.R., and Venkatesan, R. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, New York, NY: John Wiley and Sons, 1999.

Liang, Z.P. and Lauterbur, P.C. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*, Piscataway, NJ: IEEE Press, 2000.

Stark, D., Bradley, Walter G., and Bradley, William G. (Eds.) *Magnetic Resonance Imaging*. St. Louis, MO: Mosby, 1999.

Problems

Magnetization

12.1 A nonuniform magnetic field B pointing in the z direction is applied to a sample of protons. The field B (in tesla) varies as a function of z (in cm):

$$B(z) = 1 + 0.5z\,.$$

The magnetization vector M precesses around the z-axis. Suppose at time $t = 0$, all magnetization vectors have the same phase. At what time will the magnetization vector M at $z = 1$ cm and that at $z = 0$ have the same phase again?

12.2 Prove that the equations in (12.12) are solutions to (12.7).

12.3 Protons that are in different chemical species, such as the CH_2 and CH_3 groups in fat, will resonate at slightly different frequencies due to local shielding of the static field by the electronic environment. This change in the resonance frequency is called *chemical shift*; it has made NMR an indispensable tool for chemists and physicists. If N species are present in the sample, and we neglect all interactions between nuclei, the FID is composed of a set of N decaying oscillators. Write down an equation that gives an expression for the transverse magnetization under these circumstances.

RF Excitation and Relaxation

12.4 A sample of ^{1}H in the equilibrium condition in a static magnetic field $\mathbf{B}_0$ is excited by a short circularly polarized RF pulse. The RF pulse is a magnetic field in the transverse plane

$$B_1(t) = B_1^e(t)e^{-j2\pi\nu_0 t}\ \mathrm{G},$$

where ν_0 is the Larmor frequency of the sample. The envelope of the RF pulse is a triangle function with parameter T:

$$B_1^e(t) = \begin{cases} \frac{1}{10}\left(1 - \frac{|t-T|}{T}\right), & 0 \le t \le 2T \\ 0, & \text{otherwise} \end{cases}.$$

(**a**) Find the tip angle of the magnetization vector as a function of t for $0 \le t \le 2T$.

(**b**) What is the value of T to make $B_1(t)$ a $\pi/2$ pulse?

12.5 The longitudinal and transverse relaxation of a magnetization vector after an RF excitation is governed by the following differential equations (in the rotating frame):

$$\frac{dM_z}{dt} = -\frac{M_z - M_0}{T_1},$$

$$\frac{dM_{xy}}{dt} = -\frac{M_{xy}}{T_2}.$$

Solve the above equations for $M_z(t)$ and $M_{xy}(t)$, assuming that $M_z(0)$ and $M_{xy}(0)$ are known.

12.6 The following equations from Example 12.4 give the components of $\mathbf{M}$ after an α pulse (assuming the system is in equilibrium just before the α pulse):

$$M_z(t) = M_0(1 - e^{-t/T_1}) + M_0 \cos\alpha e^{-t/T_1},$$

$$M_{xy}(t) = M_0 \sin\alpha e^{j\phi} e^{-t/T_2}.$$

Suppose that we now excite the sample with a train of α pulses, separated by a time T_R. The equilibrium condition is true when T_R is long compared with T_1 and we can assume that M_z just before the pulse is equal to

M_o. Derive a more general formula for $M_z(t)$. You can assume that the transverse magnetization has completely dephased before each RF pulse, that is, $M_{xy}(T_R) = 0$. (Hint: In this more general formula, M_o will be replaced with the *steady-state* value of the longitudinal magnetization. Define M_z after the $(n+1)^{th}$ pulse to be M_z^{n+1}, and M_z after the n^{th} pulse to be M_z^n. Relate these two quantities with an equation. Derive another (very simple) equation from the steady-state condition. You now have enough information to solve the problem.)

12.7 In this problem, we answer the question: What is the tip angle that maximizes MR signal intensity for a pulse sequence with given T_R and T_E values? Solve this problem for given T_1, T_2, and T_2^* values. Assume that the magnetization is in the steady-state condition, that is, the magnetization is a periodic function of time with a period of T_R. Also assume that the transverse magnetization completely dephases before each RF pulse, that is, $M_{xy}(0^-) = 0$.

(**a**) Calculate $M_z(0^-)$, the steady-state magnetization value just before the RF pulses.

(**b**) Calculate $M_{xy}(T_E)$, the signal level in the steady-state condition.

(**c**) Calculate the optimum tip angle.

12.8 (**a**) For the pulse sequence given in Figure P12.1, write down the expression for the FID signal.

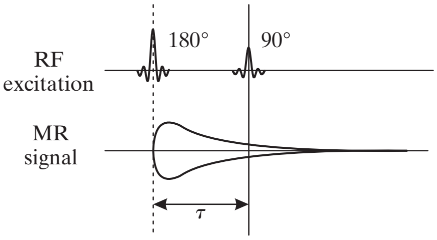


Figure P12.1 Pulse sequence for Problem 12.8.

(**b**) Explain how we can use this pulse sequence to measure T_1.

Bloch Equations and Spin Echoes

12.9 Transform the Bloch equations into the rotating frame.

12.10 (**a**) Explain why a π pulse is applied at $T_E/2$ in order to get an echo at $t = T_E$.

(**b**) Suppose a $\pi/2$ pulse is applied to a sample at $t = 0$, and a sequence of π pulses is applied to the sample at $t = \frac{2k+1}{2}T_E$, $k = 0, 1, \ldots$. If T_E is small compared with T_2, find an expression for the magnitude of $M_{xy}(kT_E)$ for $k = 1, 2, \ldots$.

12.11 A rectangular RF pulse given by

$$\mathbf{B}_1(t) = A\,\mathrm{rect}(t/10^{-5})[\cos(2\pi\nu_0 t)\hat{\mathbf{x}} + \sin(2\pi\nu_0 t)\hat{\mathbf{y}}]$$

is used to tip the magnetization vector away from its equilibrium position along the z-axis by 15°. What is the value of A required to achieve this tip angle?

12.12 A $\pi/2$ pulse is applied at $t = 0$, sending the longitudinal magnetization of a sample into the transverse plane in the direction of the x-axis. At that instant a spatially-varying (but temporally constant) magnetic field $\Delta B(\mathbf{r})$ oriented in the z direction is added to the static field B_0. The spatially-varying Larmor frequency is then given by

$$\nu_0(\mathbf{r}) = \bar{\gamma}(B_0 + \Delta B(\mathbf{r}))\,,$$

and the phase of the precessing sample is given by

$$\phi(\mathbf{r}, t) = -\gamma(B_0 + \Delta B(\mathbf{r}))t\,.$$

A π pulse is then applied at time τ.

(**a**) What is the phase $\phi(\mathbf{r}, t)$ immediately before and immediately after the π pulse?

(**b**) What is the phase $\phi(\mathbf{r}, t)$ at time $T_E = 2\tau$?

(**c**) Make a conclusive statement about the use of spin echoes when there are spatially-varying gradients in use.

Contrast Mechanism

12.13 Explain what is P_D-weighted contrast. Describe how to select imaging parameters to obtain P_D-weighted images. Explain why a large T_E cannot be used.

12.14 Suppose we are imaging a human brain (see Table 12.2 for the NMR properties of different brain tissues). We use the strategy described above in Problem 12.6 and acquire signals after the system is in the steady state. For simplicity, we assume the intensity of the reconstructed MR image is directly proportional to the magnitude of the transverse magnetic field right after the RF excitations. In order to get maximal signal strength, we use $\pi/2$ pulses.

(**a**) In order to get the best contrast between GM and CSF, what is the best value for T_R?

(**b**) Using the T_R value computed in part (a), what is the contrast between GM and CSF? What is the contrast between GM and WM?

12.15 We try using the pulse sequence shown in Figure P12.2 to image the human brain. A $\pi/2$ pulse is applied to the brain while in equilibrium.

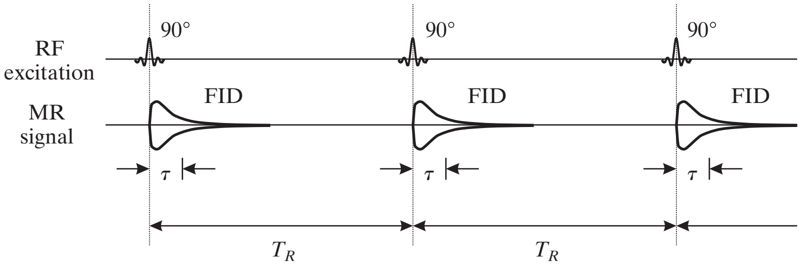


Figure P12.2
Pulse sequence for Problem 12.15.

Suppose the duration of the RF pulse is negligible. The FID signal is sampled at time τ and the pulse repetition interval is T_R.

(**a**) We use $T_R = 6,000$ ms, $\tau = 20$ ms. Is this a T_2-weighted contrast? If not, what modification is needed to obtain T_2-weighted images?

(**b**) Sketch the modified pulse sequence.

(**c**) Provide some reasonable values of τ, T_R, tip angle, and other parameters that should be used in the new pulse sequence.

CHAPTER

13

Magnetic Resonance Imaging

In the previous chapter, we studied nuclear magnetic resonance (NMR), the physics behind magnetic resonance imaging (MRI). We saw that it is possible to manipulate nuclear spin systems in such a way that a radio frequency signal is generated by the object in the form of a free induction decay (FID). We also saw that spin echoes can be used to recover an otherwise very brief or transient FID. We even saw that a variety of different types of tissue contrasts can be created by manipulation of the timing of the various excitations. What we did not see, however, is how the spatial dependency of the underlying object can be encoded so that images can be created.

In fact, looking back at the previous chapter, you will notice that we assumed there was a single object called the sample, which produced a single coherent signal called the FID or echo. A single object and its corresponding single signal was the basic viewpoint of the first 40 years or so of NMR. Then, in the early 1970s, Paul Lauterbur had the idea to spatially encode the NMR signal in order to create images. The first MR scanners were built in the late 1970s, and there has been a continuous growth in technical development and clinical use since that time.

In this chapter, we first explore the instrumentation necessary to create MR images. We then present the image formation process, starting from imaging equations and ending with computer algorithms that use these equations to generate MR images. Finally, we discuss the factors affecting image quality in order to get a sense of the limitations of MR imaging and of the possibilities for the future.

13.1 Instrumentation

13.1.1 System Components

As depicted in Figure 13.1, there are five principal components constituting an MRI scanner: the main magnet; a set of coils to provide a switchable spatial gradient in the main magnetic field; resonators or "coils" for the transmission

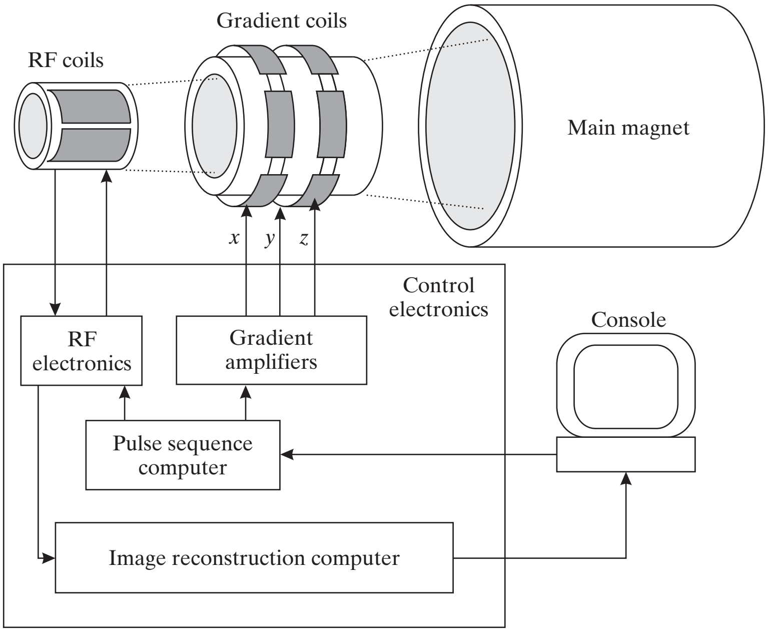
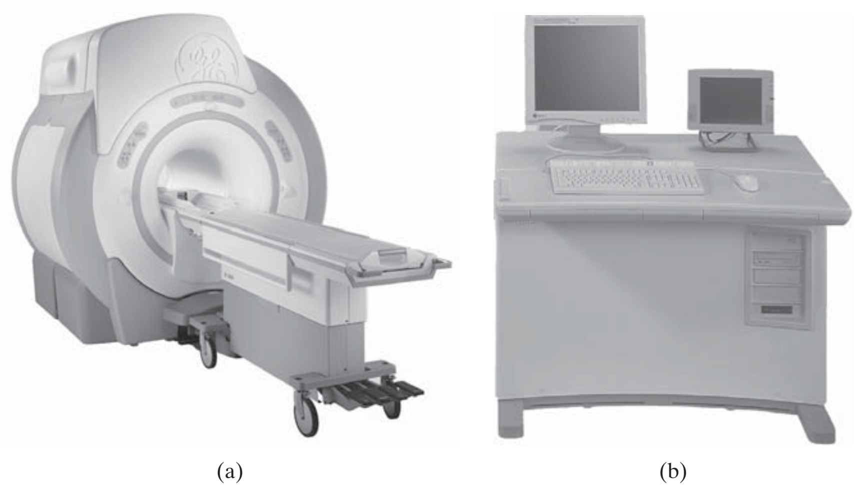


Figure 13.1
Block diagram of MR scanner components.

and reception of radio frequency pulses; electronics for programming the timing of transmission and reception of signals; and a console for viewing, manipulating, and storing images. Figure 13.2(a) shows a photograph of an MRI magnet with patient table and Figure 13.2(b) shows the operator console.

The magnet, gradient coils, and RF coils must be isolated from the electronic noise of the outside world—for example, broadcast radio signals—in order to prevent interfering signals. Accordingly, these elements are placed in a copper-lined room, which acts as a *Faraday cage*. All electronic signals that enter the scan room, such as gradient currents and power outlets, must first go through bandpass filters to ensure that no electrical noise in the range of the receiver electronics is present inside the room. One or more racks of electronics, including


(a)

(b)

Figure 13.2
(a) The main magnet with the patient table and (b) the console for operating the scanner. Courtesy of GE Healthcare.

the gradient amplifiers and transmit/receive electronics, are typically housed in a room adjacent to the scanner room.

Like computed tomography (CT) and positron emission tomography (PET) scanners, the majority of MRI systems on the market today have the patient lying on a sliding table inside a cylinder. This geometry requires large (>1 meter diameter) solenoidal coils of superconducting wire to be used to provide a strong (>1 tesla) main magnetic field. Other geometries, such as those with magnets having two pole pieces arranged either vertically or horizontally, have also been produced. These so-called *open MRI* systems give easier access to the patient but have lower field strengths (<1 tesla) and hence lower signal amplitude. We focus our attention on 1.5 T MR scanners, which represent approximately 70% of the installed scanners in the United States. Our presentation is general enough, however, to adequately describe the operation and imaging characteristics of both lower and higher strength MR scanners.

13.1.2 Magnet

The most common type of magnet used in MRI systems is the cylindrical superconducting magnet (typically with a 1 meter bore size). In a superconducting magnet, as depicted in Figure 13.3, coils of niobium-titanium wire are immersed in liquid helium held at about 4° K. Niobium-titanium is a superconductor at temperatures less than 9.5° K, and helium boils at 4.2° K. A dewar, also known as a *cryostat*, is used to keep the helium from boiling. The cryostat is essentially a "sandwich" that contains liquid helium between layers of liquid nitrogen and a vacuum, which help to insulate the helium.

Field strengths in MR scanners vary from 0.5 T (21 MHz) to 3.0 T (128 MHz), with a few systems operating at 9 T (383 MHz). The most common magnet used for whole body imaging operates at 1.5 T. This field strength has no special or optimal characteristics; it was simply the highest field achievable at the time the MRI market was being established (in the late 1970s). Because these magnets use superconducting wire, the current is on all of the time. For a 1.5 T magnet, the stored energy in this constant current is approximately 2.8 million joules; put another way, if the solenoid carrying the current was a single wire, that wire would carry 740 amperes of current.

There are two major challenges in the design and maintenance of superconducting magnets. First, the homogeneity of the magnetic field within the

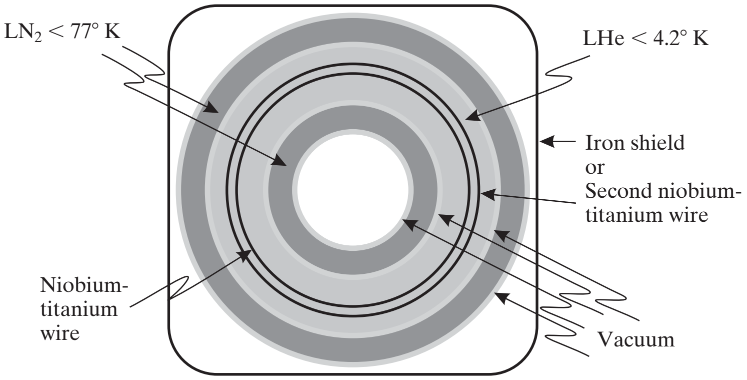


Figure 13.3 Arrangement of cold fluids for superconducting coils.

bore (field of view) must be maintained at better than ±5 ppm. The process of *shimming* is used to "tailor" the magnetic field in order to improve homogeneity. *Passive shimming* is the placement of small pieces of metal just inside the bore. The metal pieces disturb the field created by the superconducting wire and, when properly placed, the result is a more homogeneous field. *Active shimming* is the adjustment of currents in as many as 30 small coils of wire placed just within the bore. This process is more automated than passive shimming, but it is more expensive to manufacture.

The second challenge in the design of superconducting magnets is the minimization of the so-called *fringe field*—the magnetic field that is outside the bore of the magnet. Because these fields are very large, the fringe field can cause significant problems in devices that depend on either magnetic storage (e.g., credit cards) or moving currents (e.g., implanted medical devices). Because of these kinds of detrimental effects caused by the fringe fields, MR scanners must be sited carefully within a hospital or clinic building.

Two mechanisms can be used to reduce fringe fields, again one passive and one active. The passive mechanism is to simply put a large iron shield around the entire superconducting magnet. This tends to reduce the fringe field beyond the iron but also detrimentally affects the field within the magnet's bore. The active mechanism is to add an additional set of superconducting coils *outside* the primary coils (see Figure 13.3). The additional set has current moving in the opposite direction, which significantly lowers the field outside. Since the second set also reduces the field within the magnet, the primary coils have to operate at a higher current than would otherwise be the case, in order to maintain the desired magnetic field strength. Active shielding has the advantage that other large metal objects outside the scanner (e.g., steel beams within the walls of the building) will have less of an effect on the field within the bore.

13.1.3 Gradient Coils

As depicted in Figure 13.1, the *gradient coils* fit just inside the bore of the magnet (after any active shimming coils, if present). The function of the gradient coils is to provide a temporary change in the magnitude B_0 of the main magnetic field as a function of position in the magnet bore. There is no process described in Chapter 12 that requires such a function; in fact, we have gone to great lengths to ensure that the main field is constant. So, why would we want to deliberately perturb this "perfection" by creating an additional field that deliberately causes the main field to vary spatially? It turns out that this is the key to spatially encoding the NMR signal—the key to creating images.

Gradient coils provide the means to choose slices of the body for selective imaging. In this way, MRI can be *tomographic*—that is, it can image slices. Gradient coils also provide the means to spatially encode the pixels within a given image slice so that the individual FIDs and echoes coming from each one of thousands of pixels can be unraveled and turned into an image. We will see how all this works in Section 13.2. For now, let us study the gradient coils themselves.

There are usually three orthogonal gradient coils, one for each of the physical x, y, and z directions, as shown in Figure 13.4. For a cylindrical magnet, these gradient coils are wound on a cylindrical former and stabilized

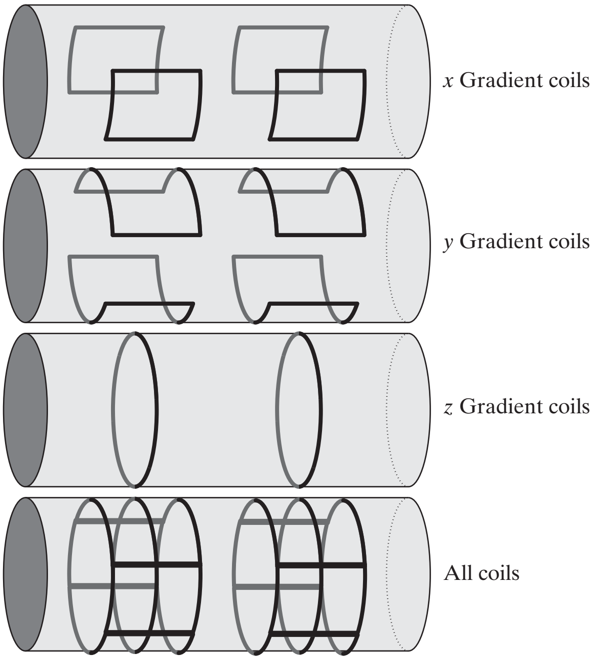


Figure 13.4 Arrangement of the gradient coils.

by "potting" the windings in epoxy. Circulating water or air is used to cool the current-carrying elements. Gradient coils are subject to large forces—sometimes called *Lorentz forces*—since they carry currents in a strong magnetic field. Furthermore, in order to acquire images, the gradient currents are repeatedly turned on and off, leading to significant vibration. In addition to the potential to break down the epoxy and thereby weaken electrical connections, this vibration is the source of the rather loud "knocking" sounds associated with MR scans.

The purpose of each of the gradient coils is to add or subtract a spatially dependent magnetic field to the main field. In particular, if all three coils are turned on at the same time with strengths G_x, G_y, and G_z, respectively, then the main field is (ideally) given by

$$\mathbf{B} = (B_0 + G_x x + G_y y + G_z z)\,\hat{\mathbf{z}}\,. \tag{13.1}$$

Notice that the gradient coils do not change the direction of the magnetic field; instead, they add to (or subtract from) the magnitude of the main field. This is a common point of confusion when first learning MRI and should be clearly understood before moving on. Ideally, the change in field strength resulting from currents in the gradient coils is a linear function of spatial position. This helps considerably in the development of algorithms to reconstruct images, as we shall see. The *gradient amplitude* is defined by the constants G_x, G_y, and G_z, which have units of gauss per centimeter. It is often written in vector form as

$$\mathbf{G} = (G_x, G_y, G_z)\,. \tag{13.2}$$

Using the notation $\mathbf{r} = (x, y, z)$ to denote a vector spatial position, we find that (13.1) can be written using a dot product notation as

$$\mathbf{B} = (B_0 + \mathbf{G} \cdot \mathbf{r})\,\hat{\mathbf{z}}\,. \tag{13.3}$$

The gradient coils are designed to produce a linear spatial perturbation of the magnetic field, as described above. As shown in Figure 13.4, the x- and y-gradients can be produced by pairs of *saddle coils* and the z-gradient can be produced using two *opposing coils*, each wound around the circumference of the cylindrical bore. The maximum gradient amplitude is determined by the maximum current that the coil can carry. This, in turn, is limited by the gradient amplifier, the heating of the coil, and the forces on the coil. Typically, 100–200 amperes of current are available for the x, y, and z coil each. The maximum gradient amplitude in clinical scanners is usually on the order of 1–6 gauss/cm (or 10–60 mT/m).

We will see that the faster a gradient can turn on the faster we will be able to acquire an image. The switching time from zero to the maximum gradient amplitude is also dependent on the coil and amplifier design. Typical switching times are on the order of 0.1–1.0 ms. The so-called *slew rate* is more commonly used to characterize the overall performance of a gradient coil and amplifier pair. The *slew rate* is the maximum achievable rate of change of the gradient value and is given in units of mT/m/ms. Typical values range from 5 to 250 mT/m/ms.

The gradient coils must also have an auxiliary *shielding coil* on the outside of the gradient coil cylinder in order to minimize the change in the magnetic field outside the gradient housing. Without this shield, the changing flux through surrounding metal components such as the magnet housing would cause significant eddy currents, which exist on time scales from milliseconds to seconds. These induced eddy currents completely change the temporal profile of the magnetic field gradients. Some of the most innovative engineering that has emerged from the development of MRI has been the design of gradient coils that can be switched rapidly without inducing large eddy currents in the metallic components of the magnet housing.

In MR imaging methods, the gradients are switched on and off very rapidly. Ideally, this change in field strength would occur instantaneously; however, there are physical limitations that prevent this. The first limitation is the self-inductance of the gradient coil, which can be overcome with high-voltage amplifiers, or by redesigning the coil to be smaller (usually this means shorter in the z direction). The tradeoff when making the gradients smaller is that the region of linearity in the change of the magnetic field becomes smaller, and hence the maximum achievable useful field of view (FOV) becomes smaller. The second limitation is the induction of eddy currents in the patient causing peripheral nerve stimulation (muscle twitching). The Food and Drug Administration (FDA) has set a limit of 40 T/s on the exposure to magnetic field switching. Above this limit, the probability of peripheral nerve stimulation is not negligible.

EXAMPLE 13.1

Although the bore of a magnet is one meter, the human body normally takes up considerably less room. For the sake of concreteness, let us say the diameter of a human body is d and $d \approx 0.5$ m—this is the field of view (FOV). In the center of the magnet, the additional field caused by a gradient is zero, rising to its maximum values on the edges of the FOV, which is at a radius of $d/2$.

Question How rapidly can an MR scanner legally establish a gradient of $G = 40$ mT/m in the human body, and what is the slew rate required to achieve this?

Answer The maximum contribution of the magnetic field in the human body due to the gradient coils is

$$G_{\text{max}} = 40 \text{ mT/m} \times \frac{d}{2} = 10 \ \text{ mT}\,.$$

We assume that the gradient waveform is linear until it reaches the desired final value. Therefore, the minimum time to reach the maximum value is

$$t_{\text{min}} = \frac{G_{\text{max}}}{40 \text{ Tesla/s}} = \frac{10 \text{ mT}}{40 \text{ Tesla/s}} = 0.25 \text{ ms}\,.$$

The slew rate required to achieve this is

$$\text{SR} = \frac{40 \text{ mT/m}}{0.25 \text{ ms}} = 160 \text{ mT/m/ms}\,.$$

■

13.1.4 Radio Frequency Coils

In the previous chapter, we saw the importance of radio frequency (RF) induction in NMR. Current that is oscillating at the Larmor frequency applied around the sample causes the nuclear spins to precess, tipping them toward the transverse plane. Once a spin system is excited, coherently rotating spins can induce RF currents (at the Larmor frequency) in nearby antennas, yielding measurable signals associated with the FID and echoes. Thus, RF *coils* (sometimes called *resonators*) in MRI systems, as shown in Figure 13.1, serve to both induce spin precession and to have currents induced in them by the spin system.

There are two basic types of RF coils: volume coils and surface coils. *Volume coils* are designed to (mostly) surround the object being imaged, while *surface coils* are designed to be placed on the surface in very close proximity to the object being imaged. Different types of RF coils are depicted in Figure 13.5. Volume coils are preferable to surface coils in most instances because their sensitivity (field) patterns are very uniform within the body. This means that transmitted energy is uniformly distributed throughout the sample so, for example, tip angles intended to be $\pi/2$ are fairly close to being $\pi/2$ radians everywhere. In addition, FIDs generated throughout the sample in response to an excitation are received in a relatively uniform manner using volume coils. Surface coils, in contrast, are very sensitive to sources close to the coil, but their sensitivity degrades rapidly away from the coil. Arrays of surface coils can be used to improve the coverage (e.g., "belts" containing four or more coils can be strapped around the torso), but the uniformity is still quite inferior to the carefully designed geometrically perfect volume coils.

All scanners have a *body coil*, which fits just inside the gradient coils, as shown in Figure 13.1 (labeled as RF coils in the figure). Usually the body coil is a *birdcage resonator* as depicted in Figure 13.5(b). Most scanners come with a

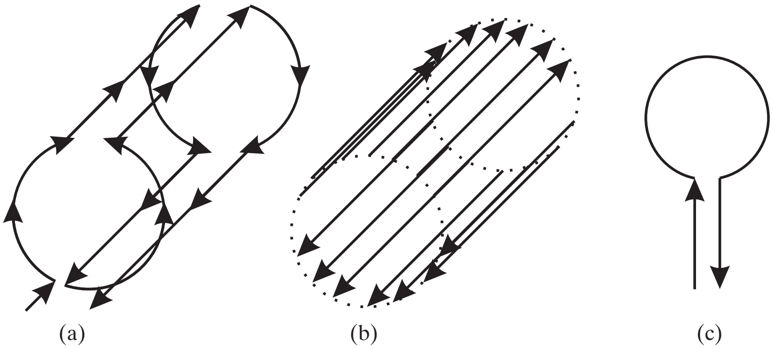


Figure 13.5
Different types of radio frequency coils: (a) saddle coil, (b) birdcage coil, and (c) surface coil.

head coil, which is a much smaller coil—birdcage, saddle coil, or Aldermann-Grant coil (not shown)—designed to fit closely around the human head. Other specialty volume coils, such as knee coils, neck coils, and small extremity coils can be purchased for most MRI scanners. Surface coils are individual loops [see Figure 13.5(c)] placed in close proximity to the organ being imaged, or they can be used in combination—so-called *phased arrays*—in order to improve their homogeneity and sensitivity over that of single coils.

During RF transmission, relatively large currents are produced in the coil elements from an RF amplifier, with a power requirement of approximately 2 kW for human imaging. Ideally, a transmission coil produces a relatively uniform B_1 field throughout the entire imaging volume. On reception, the coil must pick up very low amplitude magnetic fields, which produce very small voltages in the coil. Because transmission and reception require currents of much different amplitude in the coil, these functions are often split into two separate subsystems. In practice, a body coil located just inside the gradient coils is used to transmit RF into the patient, and another coil (e.g., head coil, surface coil, or phased array coil) positioned close to the volume of interest is used for signal reception. This is a very good way to increase the signal-to-noise ratio of the imaging signal, as will be discussed later. For both the transmit and receive coils, it is important to realize that only the transverse components of the magnetic field are used for imaging.

13.1.5 Scanning Console and Computer

The console [see Figures 13.1 and 13.2(b)] in a typical MRI system is used by the operator to select the scanning protocol, set the gating to the patient's electrocardiogram (ECG) or breathing (to synchronize acquisition to the appropriate periodic physiologic process), graphically select the orientation of the scan planes to image, review images obtained, and change variables in the pulse sequence to modify the contrast between tissues. The operator's console is connected to a compute engine, such as an array processor, which performs the image reconstruction. Current scanners reconstruct about 10–50 images per second, which is adequate for real-time scanning of slices. The speed of reconstruction is often limited by the data transfer rate from the receiver electronics to the array processor. Some scanners now have real-time image feedback much like ultrasound, so that the operator can use the real-time images from a single slice to maneuver the scan plane.

13.2 MRI Data Acquisition

Projection radiography encodes the spatial position of objects using the position of x-rays that hit a detector. Factors affecting the encoding of spatial position in radiography are the magnification caused by the diverging beam and the superposition of overlaying structures. In radiography, the superposition of structures is never "decoded;" radiologists learn to read radiographs with the knowledge that images are corrupted in this fashion. CT encodes the spatial position of objects by observing their effect in many projections, each from a different orientation. The position of an object in CT is actually spread across the observed data set in a rather complicated way. CT reconstruction algorithms decode the spatial position by filtering and integrating the data using a mathematical algorithm—for example, convolution backprojection.

So far, we understand only that MRI will use the gradient coils to encode spatial position. But how will this be done? How will the data be decoded? The "trick" is to use both the Larmor frequency and the phase of the transverse magnetization to encode spatial position. In this section, we present both frequency and phase encoding of spatial position and develop a fundamental MR image formation method.

13.2.1 Encoding Spatial Position

Before discussing the encoding of spatial position, we should ask, "What is the spatial coordinate system?" It is customary for $+z$ to be oriented along the direction of the $\mathbf{B}_0$ field and for the $+y$ direction to be oriented up. The $+x$ direction must therefore be horizontal and oriented in such a way to create a right-handed coordinate system. When the patient goes in head-first and supine, this means that the $+z$ direction is from the head to the feet; $+y$ is oriented posterior (back) to anterior (front); and $+x$ is oriented right to left, as shown in Figure 13.6. In this scenario, we see that if we were able to image a slice whose z-coordinate is constant, we would get a so-called *axial* image; we would get a *coronal* image by holding y constant; and we would get a *sagittal* image by holding x constant.

Although an MR scanner can create images at arbitrary locations and orientations, for simplicity, we will describe only the formation of an axial image.

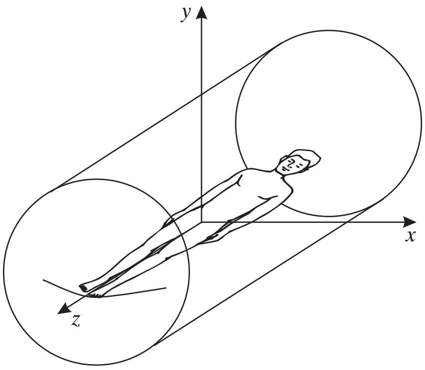


Figure 13.6
Laboratory coordinates in an MR scanner.

In this case, the z direction corresponds to the *through-plane* direction, and the x and y directions correspond to the *in-plane* directions. It is straightforward to acquire both coronal and sagittal images and also *oblique* and *double-oblique* images. An oblique image is one whose orientation is found by taking an image that is orthogonal to one of the three cardinal orientations and rotating about another cardinal axis. A double-oblique image has an orientation that requires a second rotation about a cardinal axis. The methods we develop for the formation of axial images will carry over with relatively little modification to the acquisition of images in these other orientations.

In order to encode spatial position using frequency, the Larmor frequency should vary as a function of spatial position. Recall that the frequency at which a magnetization vector $\mathbf{M}(t)$ precesses about a magnetic field $\mathbf{B}$ is given by $\nu = \gamma\!\!\!\!- B$, where $B = |\mathbf{B}|$. Now suppose that we turn on a constant gradient field $\mathbf{G} = (G_x, G_y, G_z)$ by applying a constant current to all three gradient coils. Then, the total magnetic field is given by (13.3), and the Larmor frequency is given by

$$\nu(\mathbf{r}) = \gamma\!\!\!\!- (B_0 + \mathbf{G} \cdot \mathbf{r}) \,, \tag{13.4}$$

where the dependence of Larmor frequency $\nu(\mathbf{r})$ on spatial position $\mathbf{r} = (x, y, z)$ is made explicit.

This concept, called *frequency encoding*, is the first of two concepts used to encode spatial position in MRI, the other being called *phase encoding*. Frequency encoding is used both for slice selection, as we describe in Section 13.2.2, and for reading out samples in Fourier space, as we describe in Section 13.2.3. Phase encoding is used to select a position in Fourier space, as described in Section 13.2.6.

EXAMPLE 13.2

It is very common to use only one gradient component when applying the concept of frequency encoding.

Question If a sample is put in a magnetic field with $B_0 = 1.5$ T, and a z-gradient with strength $G_z = 3$ G/cm is applied, what is the Larmor frequency for the protons on the $z = 0$ plane? If we are to image a slab with a thickness of 0.5 m centered at $z = 0$, what is the range of Larmor frequencies of the protons in the slab?

Answer In this case, $\mathbf{G}$ has only one nonzero component

$$\mathbf{G} = (0, 0, G_z) \,.$$

Using (13.4) gives

$$\nu(\mathbf{r}) = \gamma\!\!\!\!- (B_0 + \mathbf{G} \cdot \mathbf{r}) \tag{13.5a}$$

$$= \gamma\!\!\!\!- (B_0 + G_z z) \,. \tag{13.5b}$$

For protons, the gyromagnetic ratio is $\gamma\!\!\!\!- = 42.58$ MHz/tesla. So on the $z = 0$ plane the Larmor frequency is

$$\nu_0 = \gamma\!\!\!\!- B_0 = 63.87\,\text{MHz} \,.$$

With the z-gradient, the strength of the magnetic field is

$$B(z) = B_0 + G_z z \,.$$

Its range within the slab of sample is between

$$B_{\min} = B_0 - G_z \times 0.25\,\mathrm{m} = 1.5\,\mathrm{T} - 7.5\,\mathrm{mT}\,, \tag{13.6a}$$

$$B_{\max} = B_0 + G_z \times 0.25\,\mathrm{m} = 1.5\,\mathrm{T} + 7.5\,\mathrm{mT}\,. \tag{13.6b}$$

So, the range of the Larmor frequencies within the slab of sample is

$$63.55\,\mathrm{MHz} \le \nu \le 64.19\,\mathrm{MHz}\,.$$

■

13.2.2 Slice Selection

We have already studied several medical imaging techniques that image 2-D *slices* of the human body: CT, SPECT, PET, and ultrasound. In CT and ultrasound, the energy used to image the selected slice is restricted to the slice itself. In that way, the physiological property giving rise to image contrast arises from within only the selected slice—there is no other part of the body from which signal can arise. In SPECT and PET, the entire body is a potential source, but the observed signal is selected by physical and electronic collimation so that it belongs to a specific slice.

In MRI, the basic principles of both of these approaches can be used. Specifically, it is possible to excite only a selected slice so that the received signal can arise only from within the selected slice; it is also possible to excite the whole volume and then to extract images of selected slices. The first technique is called *2-D MR imaging* and the second is called *3-D MR imaging*. We confine our detailed presentation in this chapter to 2-D imaging and only comment briefly on 3-D imaging. The first step in 2-D imaging is to perform *slice selection*, the selective excitation of a nuclear spin system in a slice. As noted above, for convenience we describe the selection—and then imaging—of axial slices. The principles carry over quite directly to the imaging of arbitrary slices.

Principle of Slice Selection In Example 13.2, we found that application of the gradient $\mathbf{G} = (0, 0, G_z)$ yields a Larmor frequency that is a function of z. This spatially varying Larmor frequency can be written as

$$\nu(z) = \gamma\!\!\!{}^{-}(B_0 + G_z z)\,, \tag{13.7}$$

which is illustrated in Figure 13.7. If a pure sinusoidal RF excitation at a specific frequency were applied, then an infinitesimally thin slice of the body would undergo forced precession. That slice corresponds to all points whose z position has the specific Larmor frequency being excited. However, it is neither possible to create such a sinusoidal waveform in practice nor is it practically desirable. Instead, we create a waveform that excites a range of frequencies, which in turn excites a range of tissues, corresponding to a "thick slice" or a *slab*.

Figure 13.8 shows two slice selection scenarios. In the first, Figure 13.8(a), a z-gradient $G_z = G_A$ is used and a signal containing the range of frequencies $\nu \in [\nu_1, \nu_2]$ is generated. (We will discuss what this signal actually looks like below.) This combination causes a forced precession of all the spins around the neck area of the subject in this example. However, the slice is thicker than would

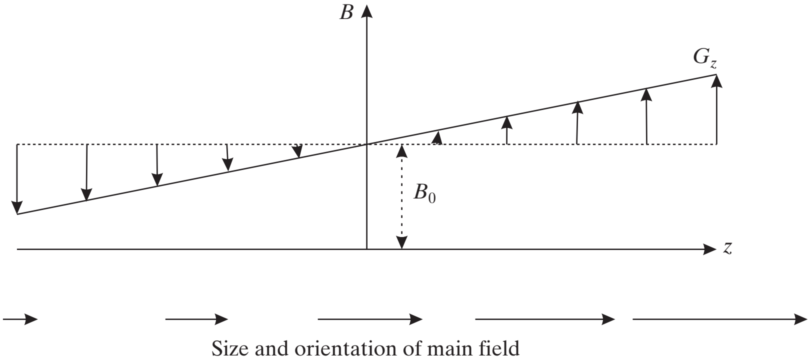


Figure 13.7
Effect on the main magnetic field from a z-gradient.

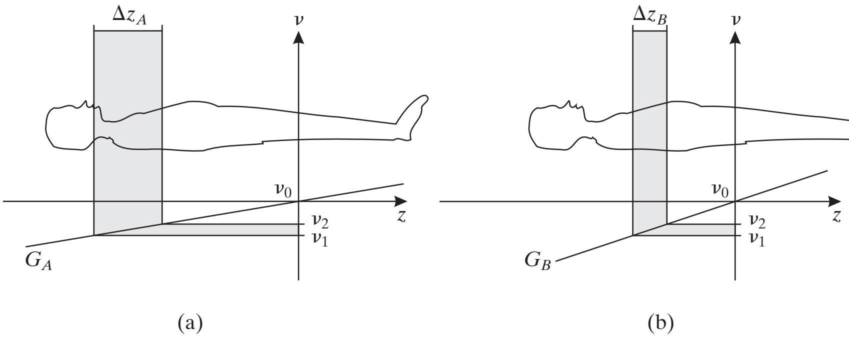


Figure 13.8
Slices are selected by application of a z-gradient and excitation over a specific frequency range. (a) Smaller and (b) larger gradients.

ordinarily be desirable. In the second scenario, Figure 13.8(b), a larger z-gradient $G_z = G_B$ is used and the same RF signal is applied. Because the z-gradient is larger in magnitude than that in Figure 13.8(a)—that is, $|G_B| > |G_A|$—the same RF waveform having frequencies in the range $[\nu_1, \nu_2]$ causes the forced precession of a thinner slice. Notice also that the slice is in a different position as well, at the lower chest in this case.

By studying Figure 13.8 we see that there are actually three parameters that are used together to select slices: *z-gradient strength* G_z, *RF center frequency*,

$$\overline{\nu} = \frac{\nu_1 + \nu_2}{2}, \tag{13.8}$$

and *RF frequency range*,

$$\Delta\nu = |\nu_2 - \nu_1| . \tag{13.9}$$

For example, assume that we want to image a thinner slice at the neck region using one of the two z-gradients in Figure 13.8. There are two ways to achieve this. Suppose we use the smaller z-gradient, G_A, as in Figure 13.8(a). We see that the RF center frequency is fine, but the RF frequency range would have to be made smaller in order to make the excited slice narrower. This is the first approach. The second approach uses the larger z-gradient, G_B, as in Figure 13.8(b). In this case, the RF frequency range is fine—it already produces a thin slice—but the RF center frequency must be made smaller in order to shift the slice from the chest region to the neck region.

With these three parameters, it is possible to control both *slice position* $\overline{z}$ and *slice thickness* Δz. Solving (13.7) for z, we find that the lowest and highest excited frequencies, ν_1 and ν_2, yield the slice boundaries,

$$z_1 = \frac{\nu_1 - \gamma\!\!\!\!-B_0}{\gamma\!\!\!\!-G_z}\,, \tag{13.10a}$$

$$z_2 = \frac{\nu_2 - \gamma\!\!\!\!-B_0}{\gamma\!\!\!\!-G_z}\,, \tag{13.10b}$$

where $\nu_1 = \nu(z_1)$ and $\nu_2 = \nu(z_2)$. Slice position $\overline{z}$ is therefore given by

$$\begin{aligned}\overline{z} &= \frac{z_1 + z_2}{2} && (13.11a)\\ &= \frac{\overline{\nu} - \nu_0}{\gamma\!\!\!\!-G_z}\,, && (13.11b)\end{aligned}$$

where (13.11b) follows after some algebra. Slice thickness Δz is given by

$$\begin{aligned}\Delta z &= |z_2 - z_1| && (13.12a)\\ &= \frac{\Delta\nu}{\gamma\!\!\!\!-G_z}\,. && (13.12b)\end{aligned}$$

EXAMPLE 13.3

Suppose that we desire a slice thickness of 2.5 mm and that the z-gradient strength is $G_z = 1$ G/cm.

Question What RF frequency range should be excited?

Answer Solve (13.12b) for the frequency range

$$\begin{aligned}\Delta\nu &= \gamma\!\!\!\!-G_z\Delta z\\ &= \gamma\!\!\!\!- \times 1 \text{ G/cm} \times 2.5 \text{ mm}\\ &= 4.258\,\frac{\text{kHz}}{\text{G}} \times 1\,\frac{\text{G}}{\text{cm}} \times 0.25 \text{ cm}\\ &= 1.06 \text{ kHz}\,.\end{aligned}$$

This gives the frequency range. The selection of the specific frequencies themselves—that is, ν_1 and ν_2—depends on the specific slice to be selected relative to the $z = 0$ origin. ■

We note that thinner slices have fewer nuclei, which makes the NMR signal smaller. At some point with progressively thinner slices, the received NMR signal—FID or echo—is too small to detect above the ambient antenna noise, and thinner slices are not practically achievable. In whole-body imaging today, slices thinner than about 0.8 mm require special pulse sequences and much longer imaging times.

Practical RF Waveforms We learned in the previous section that slice selection uses a constant gradient together with an RF excitation over a range of frequencies. We now consider what RF waveform to use in exciting the range of frequencies $[\nu_1, \nu_2]$. We desire a signal whose frequency content is

$$S(\nu) = A \operatorname{rect}\left(\frac{\nu - \overline{\nu}}{\Delta\nu}\right) . \tag{13.13}$$

According to Fourier transform theory (see Section 2.2.4 and Example 2.7), we see that the signal itself should be

$$s(t) = A\Delta\nu \operatorname{sinc}(\Delta\nu t)e^{j2\pi\overline{\nu}t} . \tag{13.14}$$

This analysis is valid provided both that the gradient is constant during RF excitation and that the RF excitation is short. An illustration of $s(t)$ is provided in Figure 13.9(a); its envelope is illustrated in Figure 13.9(b).

Let us consider the exact effect that the RF signal $B_1(t) = s(t)$ has on the spin system. Equation (12.31) gives the final tip angle α after an RF excitation pulse of duration τ_p and is repeated here:

$$\alpha = \gamma \int_0^{\tau_p} B_1^e(t)\, dt , \tag{13.15}$$

where $B_1^e(t)$ is the envelope of the RF excitation evaluated in the rotating coordinate system. For isochromats whose Larmor frequency is ν, the excitation signal in the rotating coordinate system is

$$B_1^e(t) = s(t)e^{-j2\pi\nu t} . \tag{13.16}$$

After some algebraic manipulation (see Problem 13.5), it can be shown that

$$\alpha(z) = \gamma A \operatorname{rect}\left(\frac{z - \overline{z}}{\Delta z}\right) , \tag{13.17}$$

which shows that the tip angles excited by the waveform in (13.14) define a perfect slab excitation, exactly as expected by the intuitive argument above. In principle, the value of constant A could be selected to produce an arbitrary tip angle within the slab, such as a 90° maximum signal excitation.

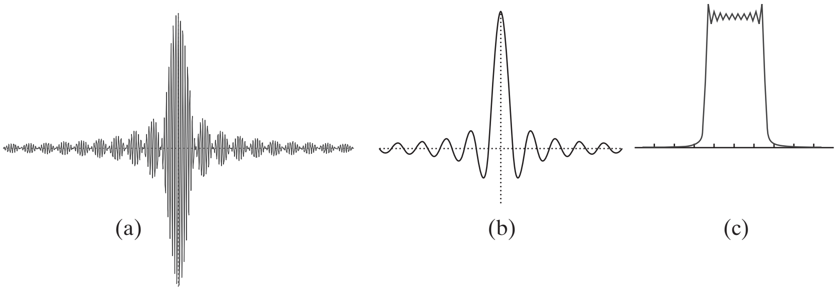


Figure 13.9
A slice selection waveform (a) and its envelope (b) and its Fourier transform (c).

Equation 13.17 is an ideal tip angle profile, which cannot be realized in practice because the sinc pulse in (13.14) would require an infinite amount of time to "play out." A more realistic scenario is created by truncating $s(t)$ and repeating the above analysis. This yields the following tip angle distribution

$$\alpha(z) = \gamma A \tau_p \mathrm{rect}\left(\frac{z-\bar{z}}{\Delta z}\right) * \mathrm{sinc}\left(\tau_p \bar{\gamma} G_z(z-\bar{z})\right) , \qquad (13.18)$$

where τ_p is the duration of the pulse (truncated beyond $[-\tau_p/2, \tau_p/2]$) and $*$ is convolution. An illustration of a slice profile resulting from a truncated sinc excitation is shown in Figure 13.9(c); it is essentially the Fourier transform of the slice selection waveform. Because of truncation, the slice profile is not perfectly rectangular in this example. This means that the edges of the slice will be somewhat blurry and there will be "ripples" caused by the convolution of the (ideal) rect profile with the sinc function associated with truncation. In general, there is a tradeoff between the duration and shape of the RF waveform and the profile edge "crispness" and ripples within the slice profile. A shorter excitation RF pulse multiplied by a Hamming window is generally desirable even though it will not produce a "crisp" rectangular slice profile.

Although it is useful to know what the slice profile is in terms of tip angle, it is somewhat more important to know what the signal strength will be as a function of z. Here, we recall that the amplitude of the transverse magnetization is proportional to $\sin\alpha$. In the case of small tip angle excitation, $\sin\alpha \approx \alpha$, which means that $|M_{x'y'}(t)|$ will (initially) have a rect profile as well. Often, the small tip angle approximation is used for initial design purposes all the way up to $\alpha = 90^\circ$ excitation. Fine-tuning of the actual excitation pulses can be made by using the Bloch equations to simulate all excitations.

EXAMPLE 13.4

From Example 13.3, a 2.5 mm thick slice can be excited using a gradient strength of $G_z = 1$ G/cm and an RF pulse bandwidth of 1.06 kHz.

Question What is the duration of the main lobe plus the first two side lobes (on either side) of the RF pulse required to achieve this slice thickness?

Answer The envelope of the waveform is $\mathrm{sinc}(\Delta\nu t)$. The time of the first zero t_1 satisfies

$$t_1 = 1/\Delta\nu = 943\ \mu s\,.$$

The duration T of main lobe plus the two nearest side lobes is four times this value. Therefore,

$$T = 3.77\ \mathrm{ms}\,.$$

A pulse that comprises only the main lobe and its two sidelobes is called a *two-period sinc approximation.* ■

Refocusing Gradients During RF excitation, the spin system within the excited slab is undergoing forced precession. The slice profile reveals differences in the final tip angles and hence implies different transverse magnetizations experienced at different z positions. Assuming reasonable design of the RF excitation waveform (e.g., some windowing to reduce spurious ripples), these differences are negligible in the final images. The strength of the gradient and RF waveform can be tuned so that the center of the slice experiences a desired tip angle, often 90° for imaging.

There is another effect that takes place during RF excitation, however, that is not negligible: *slice dephasing*. During forced precession, the spins at the "lower" edge of the slice are precessing slower than those at the "higher" edge simply because these spins have different Larmor frequencies. As a result of this, the spins become out of phase with each other across the slice. To a good approximation, the phase that is introduced is linear and is equal to $\gamma G_z(z-\bar{z})\tau_p/2$, where τ_p is the duration of the RF pulse. To see this mathematically, it is necessary to solve the Bloch equations, which we will not carry out here. We can get a sense that this is correct, however, by noticing that the greatest impact of the excitation is halfway through the pulse, since that is where the $B_t(t)$ field is strongest. If the entire RF pulse were concentrated at the center of the gradient pulse, then the phase would accumulate over duration $\tau_p/2$, rather than the whole pulse duration. Since the resulting FID or echo relies on in-phase precession, it is necessary to *rephase* the spins within the slice.

Spin rephasing is accomplished using a so-called *refocusing lobe* z-gradient waveform. One possible refocusing lobe waveform is simply a constant negative gradient of strength $-G_z$ for a duration of $\tau_p/2$, applied immediately after the initial RF excitation. Having no accompanying RF excitation, this negative gradient simply causes a change in the Larmor frequency as a function of z. The accumulated phase relative to that at $\bar{z}$ over the duration of the refocusing lobe is $\gamma(-G_z)(z-\bar{z})\tau_p/2$. Adding this to the phase accumulated over the initial RF excitation pulse yields zero, which means that the spins are now in phase over the slice and they will give an FID and can be used to form an echo in the usual way. The refocusing lobe can be shaped differently than a rectangle, however, since all that matters is that its integral is equal to half the integral of the slice selection gradient but with opposite sign. Most pulse sequences use refocusing lobes (with the required area) so that the pulse sequence can be as short as possible.

A Simple Pulse Sequence We now understand the basic elements of slice selection. A constant z-gradient is applied during which an RF waveform of duration τ_p is applied. After the RF waveform is completed, another gradient is applied to refocus the spins within the slice. After this, we should expect to find an FID arising from the excited spins in the slice that was selected.

Figure 13.10 shows a simplified pulse sequence, illustrating the concepts of slice selection. In this diagram, the gradient amplitudes are shown to change instantaneously as step functions, which is physically impossible given the minimum switching times and finite slew rates of the gradient coils in an MR scanner. In reality, the gradient pulses typically look like trapezoids with rising and falling edges having slopes equal in magnitude to the maximum slew rate of the gradient amplifiers. The mathematics becomes a bit more complicated when trapezoidal gradient pulses are used, but the principles of pulse sequence design are the same. For now we will continue to use the unrealizable rectangular pulses for didactic purposes.

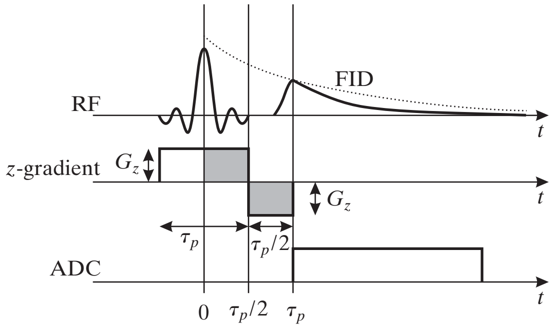


Figure 13.10
A simple pulse sequence implementing slice selection with refocusing.

At the completion of the refocusing gradient pulse, the phase angle of all magnetization vectors in the selected slice will be the same, and therefore the signals from these magnetization vectors will add constructively. This permits the reception of an FID and the creation of echoes in the usual way. If no dephasing were present across the selected slice, we would expect the FID to begin at the center of the RF pulse. As shown in Figure 13.10, this is where we place the time origin $t = 0$ in the slice selection pulse sequence. Because of dephasing, the appearance of the FID is delayed until near the conclusion of the refocusing lobe.

Although there may be brief FIDs or echoes throughout a pulse sequence, the location of the analog-to-digital converter (ADC) window in a pulse sequence indicates where data are sampled for the creation of an image. In Figure 13.10, the ADC is turned on immediately after slice refocusing in order to capture the FID.

13.2.3 Frequency Encoding

Basic Signal Model We learned in Chapter 12 [see (12.41b)] that the transverse magnetization for a uniform sample (and uniform magnetic field) after an α excitation is given by

$$M_{xy}(t) = M_{xy}(0^+)e^{-j(2\pi\nu_0 t-\phi)}e^{-t/T_2}\,, \tag{13.19}$$

where

$$M_{xy}(0^+) = M_z(0^-)\sin\alpha\,. \tag{13.20}$$

Suppose we have a heterogeneous isochromat within an excited slice. We model the spatial distribution of proton density, longitudinal relaxation, and transverse relaxation using functions of x and y only—that is, $P_D(x,y)$, $T_1(x,y)$, and $T_2(x,y)$—assuming that the slice is fairly thin so there is no z variation. This implies that there will be a spatial variation of transverse magnetization immediately after RF excitation, which can be written as $M_{xy}(x,y;0^+)$. The received signal is then an integral over the slice, given by

$$s(t) = A\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M_{xy}(x,y;0^+)e^{-j2\pi\nu_0 t}e^{-t/T_2(x,y)}\,dx\,dy\,, \tag{13.21a}$$

$$= e^{-j2\pi\nu_0 t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} AM_{xy}(x,y;0^+)e^{-t/T_2(x,y)}\,dx\,dy\,, \tag{13.21b}$$

where A is a constant representing many different gain terms arising from both physics and instrumentation, and ϕ is assumed to be zero without loss of generality.

There are a couple of details about (13.21) that we must make clear. First, we know from Chapter 12 that the FID decays more rapidly than T_2; in fact, it decays with time constant $T_2^* < T_2$. Therefore, we must view (13.21) either as an idealized signal model, or one that applies only for very short time intervals, where the difference in decay rates is negligible. Why not replace T_2 with T_2^* in (13.21)? While it is true that we would then get a more accurate representation of the true signal, we would find that it is not as easy to understand the generation of signals and images using spin echoes, a vitally important concept in MRI. We will see that the use of T_2 is most appropriate for spin-echo acquisitions, which is what we have been discussing so far. When we introduce gradient-echo acquisitions in Section 13.2.5, we will use T_2^*.

Second, it should be noted that $t = 0$ represents the center of the slice selection RF waveform, as shown in Figure 13.10. This is consistent with the time at which the spins would begin transverse and longitudinal relaxation if the pulses were made infinitesimally narrow. Third, it should be noted that this equation ignores the short time τ_p it takes for the FID to actually appear after the refocusing lobe of the slice select gradient. Nearly all practical imaging approaches record echoes rather than FIDs (as we shall see), so this small "dead period" is not usually important.

For clarity, let us define the *effective spin density* as

$$f(x, y) = AM(x, y; 0^+)e^{-t/T_2(x,y)}\,, \tag{13.22}$$

which represents the MR quantity that is being imaged here. The variable t is not included in $f(x, y)$ because it is assumed that the signal acquisition period is small relative to T_2. Using the definition given by (13.22) in (13.21), we find that the received signal is given by

$$s(t) = e^{-j2\pi\nu_0 t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy\,. \tag{13.23}$$

The received signal is always demodulated in MRI hardware, yielding the *baseband signal*

$$\begin{aligned} s_0(t) &= e^{+j2\pi\nu_0 t}s(t) && \text{(13.24a)}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy\,, && \text{(13.24b)} \end{aligned}$$

which is a constant, independent of x and y (and t, assuming a short data acquisition period). Equation (13.24) shows that this imaging procedure integrates out all the spatial dependency in the signal $s_0(t)$. Therefore, position is not encoded in this signal in any way other than the selective excitation of slice selection.

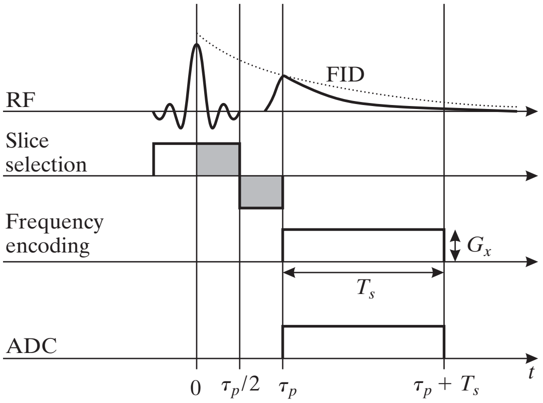


Figure 13.11
A simple pulse sequence implementing slice selection followed by frequency encoding.

Readout Gradient The first concept required for spatially encoding MR signals within the image plane is called *frequency encoding*. In frequency encoding, a gradient is turned on during the FID, as shown in Figure 13.11, causing the Larmor frequencies to be spatially dependent [see (13.4)]. (This same concept will apply to echoes, as we will see in a later section.) The direction of the frequency encoding gradient is called the *readout direction* because the signal that is "read out" (i.e., digitized when the ADC is turned on) is spatially encoded in that direction. The readout direction is arbitrary (except that it should be orthogonal to the slice selection gradient); for didactic purposes, we will associate it with the x direction. Accordingly, the Larmor frequencies during a frequency encode gradient are given by

$$\nu(x) = \bar{\gamma}(B_0 + G_x x)\,, \tag{13.25}$$

which should be compared with the Larmor frequencies during a slice selection gradient, given in (13.7).

The response during an FID under slice selection is still an integration of all the spins within the excited slice, but the Larmor frequency must be written as in (13.25). Therefore,

$$\begin{aligned} s(t) &= A \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M_{xy}(x,y;0^+) e^{-j2\pi(\nu_0+\bar{\gamma} G_x x)t} e^{-t/T_2(x,y)}\, dx\, dy\,, && (13.26\text{a}) \\ &= e^{-j2\pi\nu_0 t} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A M_{xy}(x,y;0^+) e^{-t/T_2(x,y)} e^{-j2\pi\bar{\gamma} G_x x t}\, dx\, dy\,. && (13.26\text{b}) \end{aligned}$$

Using the definition of effective spin density in (13.22), the demodulated (baseband) signal is given by

$$s_0(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) e^{-j2\pi\bar{\gamma} G_x x t}\, dx\, dy\,. \tag{13.27}$$

Equation (13.27) reveals an important concept in MR imaging and deserves a bit more explanation. In particular, the double integral on the right-hand side can be interpreted as a 2-D Fourier transform of $f(x, y)$, provided that the frequency variables are properly identified. First, we identify the spatial frequency variable in the x direction as

$$u = \gamma G_x t\,, \tag{13.28}$$

which has units of inverse length (typically cm^{-1}). We next realize that in order for (13.27) to match the 2-D Fourier transform, the spatial frequency variable in the y direction must be zero,

$$v = 0\,. \tag{13.29}$$

Denoting $F(u, v)$ as the 2-D Fourier transform of $f(x, y)$, we can now make the identity

$$F(u, 0) = s_0\left(\frac{u}{\gamma G_x}\right)\,, \tag{13.30}$$

which shows that the demodulated FID represents a certain "scan" of the 2-D Fourier space of the effective spin density.

In magnetic resonance imaging, Fourier space is usually referred to as *k-space*. This practice arises from the convention in physics where the wave number k represents a spatial frequency. Usually, the wave number has units of radians per unit length—that is, it is a radial frequency—and this was the convention in early MRI as well [see also the discussion around (10.10)]. More recently, MRI researchers associate the units of inverse length with k. In this case, the k-space variables can be identified with our Fourier frequencies,

$$k_x = u\,, \tag{13.31a}$$

$$k_y = v\,. \tag{13.31b}$$

Although somewhat unconventional, in order to avoid confusion between radial and cyclic frequencies and to maintain consistency throughout the book, we will use our customary Fourier frequencies u and v to describe MRI. When referencing a modern book on MRI, the identification in (13.31) will permit direct comparison.

Scanning Fourier Space The recognition that MR imaging can be interpreted as a "scanning" of 2-D Fourier space is a tremendously important simplifying and unifying concept. Like CT, which has the projection-slice theorem to tie 1-D projections to central slices in 2-D Fourier space, viewing pulse sequences as a type of scanning in Fourier space allows us to understand why a certain set of simple pulse sequences, like that in Figure 13.11, must be augmented or repeated with a changing parameter in order to acquire enough information to make a 2-D tomographic image.

For example, let us take a look at the Fourier information we obtain using the pulse sequence in Figure 13.11. From (13.30), this single readout provides Fourier information only on the horizontal u-axis in Fourier space, because this is where $v = 0$. Furthermore, since observations are taking place only where $0 < t < T_s$, we are only able to observe the positive u-axis over a particular interval. This Fourier scan, usually called a *Fourier trajectory*, is shown in Figure 13.12. It is quite evident from this diagram that only a small amount of

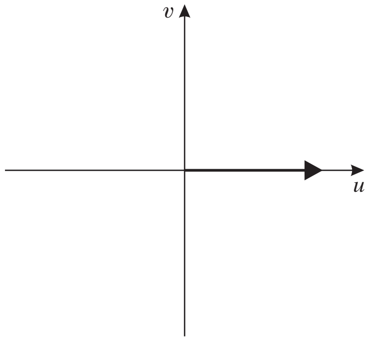


Figure 13.12
Fourier trajectory for a simple frequency-encoded FID readout.

Fourier space is scanned using this one small pulse sequence; somehow, we must find a way to scan more. How do we return to the origin again so that we might scan in some other direction? How do we scan in another direction?

In the next several sections, we study different techniques that will allow us to traverse Fourier space in a remarkably flexible manner, acquiring Fourier data as we "drive" through Fourier space. As in CT, the reconstruction algorithm that we use must match the particular geometry of the acquired data. MR is much more flexible in this regard than CT, and both rectilinear (as described above) and polar scanning are commonly used in MR imaging, requiring the fast Fourier transform algorithm and convolution backprojection, respectively, for reconstruction. The following sections are ordered for didactic purposes and do not necessarily present an order based on historical development or practical importance. Furthermore, some of the pulse sequences we present are never or rarely used in practice; therefore, we will specifically identify those that are widely used and worthy of more detailed study.

EXAMPLE 13.5

For a sample of hydrogen nuclei, we want to scan a line segment in the *k* space from $u = 0$, $v = 0$ to $u = 0.5\,\text{cm}^{-1}$, $v = 0$.

Question Assuming that the readout gradient has magnitude of $G_x = 1$ G/cm, how long should the duration of the readout gradient be?

Answer From (13.28), we have

$$u = \gamma\!\!\!- G_x t.$$

In order to scan the line segment from $u = 0$, $v = 0$ to $u = 0.5\,\text{cm}^{-1}$, $v = 0$, we must have

$$\gamma\!\!\!- G_x T_s = 0.5\,\text{cm}^{-1}.$$

The gyromagnetic ratio for ^{1}H is

$$\gamma\!\!\!- = 42.58\,\text{MHz/T} = 4.258\,\text{kHz/G}\,.$$

The duration of the readout gradient should thus be

$$T_s = \frac{0.5}{4.258 \times 10^3 \times 1} = 0.117\,\text{ms}\,.$$

■

13.2.4 Polar Scanning

Changing the direction of the Fourier space scan simply requires a different frequency-encoding gradient. In the previous section, the x direction was identified as the readout direction and an x-gradient was used to achieve this particular encoded FID signal. However, a more general gradient involving both an x and y component can be used to encode the Larmor frequency,

$$\nu(x, y) = \gamma\!\!\!-(B_0 + G_x x + G_y y)\,, \tag{13.32}$$

as shown in Figure 13.13(a). Following an analogous development as in the previous section [see (13.27)] leads to a baseband signal given by

$$s_0(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y) e^{-j2\pi\gamma\!\!\!-(G_x x + G_y y)t}\, dx\, dy\,. \tag{13.33}$$

From this, we identify the Fourier frequencies

$$u = \gamma\!\!\!- G_x t\,, \tag{13.34a}$$

$$v = \gamma\!\!\!- G_y t\,, \tag{13.34b}$$

and find that the implied Fourier trajectory is a ray emanating from the origin in the direction

$$\theta = \tan^{-1}\frac{G_y}{G_x}\,, \tag{13.35}$$

as depicted in Figure 13.13(b). Because polar Fourier data is related to to the object by the projection-slice theorem, this type of imaging is sometimes called *projection imaging*.

This more general scan strategy has the potential to be used in a complete imaging scenario. What seems to be required is the repetition of the basic pulse sequence enough times to cover Fourier space in polar rays. This is a good idea, except for a still unanswered question: How do we move back to the origin in order to start another polar scan? The simplest way is to wait for the transverse

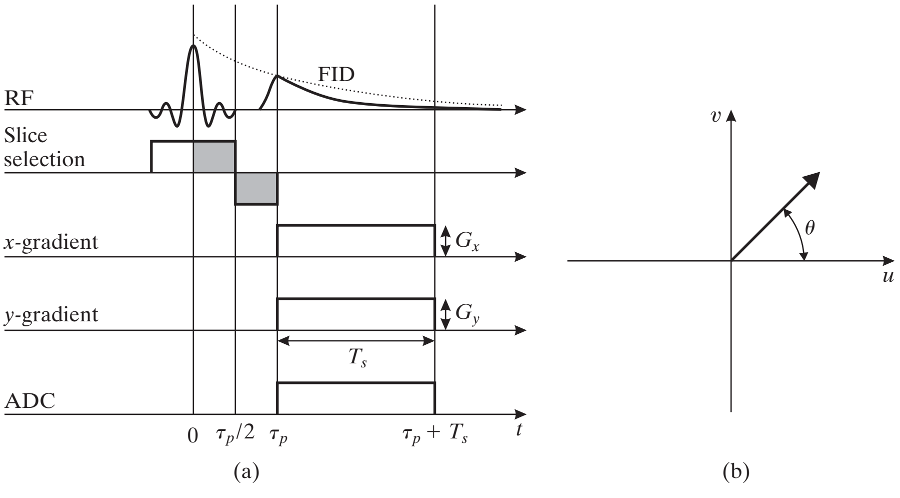


Figure 13.13
(a) A pulse sequence for arbitrary polar scan and (b) a Fourier trajectory for this polar scan.

magnetization to decay. There are two other useful concepts, however, and we begin to develop one of these in the next section.

13.2.5 Gradient Echoes

We studied the concept of spin echoes in Chapter 12, and we will return to those in a bit. First, we introduce a mechanism to create a different type of echo called a *gradient echo*. This idea can be readily connected to both the Fourier trajectories we have just introduced and the intuitive idea of spins realigning themselves, as in spin echoes and rephasing in slice selection.

Consider the pulse sequence shown in Figure 13.14(a). The first part of this pulse sequence is very recognizable. In fact, up through time $t = T_E - T_s/2$, the sequence looks like the simple frequency-encoded pulse sequence introduced above. In this case, however, the x-gradient is negative, which simply means that the trajectory in Fourier space from the origin moves in the negative u direction, rather than in the positive u direction that we studied above. This is illustrated by the A gradient lobe in Figure 13.14(a) and the A Fourier space trajectory in Figure 13.14(b).

The FID signal is ignored in this pulse sequence (as there is no ADC window during the FID). Instead, the purpose of the negative x-gradient is to move the Fourier space position onto the negative u-axis. In this way, when the positive readout gradient is applied immediately thereafter, as shown using the B lobe in Figure 13.14(a), the Fourier trajectory traverses back in the positive u direction, crossing the origin and continuing onto the $+u$-axis, as shown by the B trajectory in Figure 13.14(b).

It may not be immediately apparent why we should expect to see an echo peaking at time T_E. So, let us consider the effect of the negative x-gradient lobe with strength $-G_x$, applied immediately after slice selection. Of course, we realize that this lobe encodes spatial frequency, precisely as if it were a readout gradient. Therefore, spins are rotating faster on the $-x$-axis than they are on the $+x$-axis. These faster spins accumulate phase as

$$\phi(x,t) = \gamma \int_{\tau_p}^{t} -G_x x\, dt = -\gamma G_x x(t-\tau_p)\,, \quad \tau_p < t < \tau_p + T_s/2\,. \qquad (13.36)$$

This loss of phase coherence, by the way, occurs more rapidly than even T_2^* because of the applied gradient. At time $t = \tau_p + T_s/2$, a positive x-gradient with

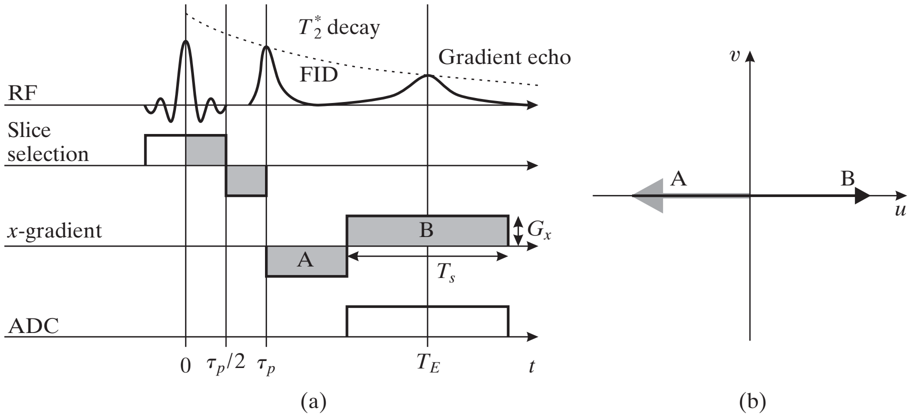


Figure 13.14
(a) A simple gradient-echo pulse sequence and (b) its Fourier space trajectory.

strength G_x is applied. During this applied gradient, the phase is accumulated according to

$$\phi(x,t) = -\gamma G_x x T_s/2 + \gamma \int_{\tau_p+T_s/2}^{t} G_x x dt \tag{13.37a}$$

$$= -\gamma G_x x T_s/2 + \gamma G_x x(t - \tau_p - T_s/2)\,. \tag{13.37b}$$

By direct substitution, we see that at time $t = \tau_p + T_s$, the accumulated phase is identically zero regardless of the actual value of G_x or the position x. This realignment of spins occurs at the center of the readout gradient, identified as the *echo time* T_E in Figure 13.14(a).

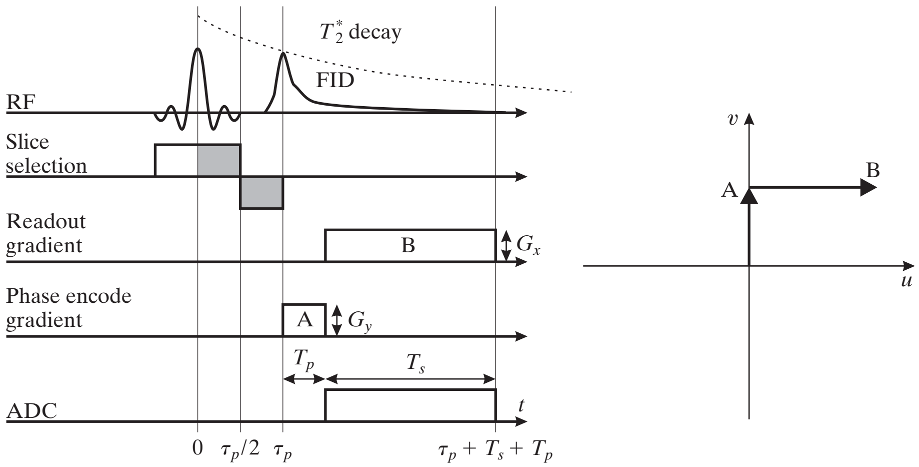
A gradient echo can therefore be viewed in two ways. It can be seen as the natural consequence of an increased signal at the Fourier origin when "driving around in Fourier space" or as a realignment of spins at $t = T_E$ due to a realignment of accumulated phase. There is one key feature of the gradient echo approach that encourages the "driving around in Fourier space" viewpoint rather than that of realignment of spins. That feature has to do with the dotted line in Figure 13.14(a), denoting the decay of the signal amplitude in this gradient echo pulse sequence. In particular, the signal strength decays with time constant T_2^* in the gradient echo approach, not with time constant T_2 as we saw in the spin echo approach in Chapter 12. The gradient echo refocuses the phase deviations that we have deliberately introduced through gradient application, not those deviations that are physically present due to static field inhomogeneities. This is a fundamental difference between gradient echoes and spin echoes, and it must always be kept in mind when designing or analyzing pulse sequences.

13.2.6 Phase Encoding

With frequency encoding and gradient echoes, it is clear that a wide variety of schemes (pulse sequences) can be designed to cover 2-D Fourier space and yield enough information to reconstruct a picture of the slice. There is a second important mechanism, however, that is used to encode spatial information in MRI: phase encoding. If we view frequency encoding as a mechanism to read out Fourier data in the u direction, then phase encoding is viewed as the mechanism to position our readout line in the v direction in Fourier space. After our discussion of frequency encoding and gradient echoes, phase encoding should seem very intuitive.

Basic Concept Consider the pulse sequence shown in Figure 13.15(a). The pulse sequence includes the usual slice-selective RF pulse followed by a refocusing gradient. The very next action is a y-gradient pulse (A in the figure) with strength G_y and duration T_p. This pulse achieves what is referred to as *phase encoding*, which can be interpreted as a polar scan in Fourier space in the vertical direction, as shown in Figure 13.15(b) (A in the figure). Although there is no readout during this pulse (because there is no ADC window at this time), from (13.33) we see that the phase accumulated during this pulse is given by

$$\phi_y(y) = -\gamma G_y T_p y\,. \tag{13.38}$$



(a) (b)

Figure 13.15
(a) A simple pulse sequence showing the phase encoding of an FID and (b) its Fourier trajectory.

The very next action in this pulse sequence is a standard x-oriented readout gradient [B in Figure 13.15(a)], which acquires data in Fourier space along the B trajectory in Figure 13.15(b). After incorporating this phase encoding step into (13.27), which is the imaging equation we previously derived for the FID readout pulse sequence, the baseband signal becomes

$$s_0(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi\bar{\gamma}G_x xt}e^{-j2\pi\bar{\gamma}G_y T_p y}\,dx\,dy\,. \tag{13.39}$$

Comparing this to the 2-D Fourier transform yields the following identifications:

$$u = \bar{\gamma}G_x t\,, \tag{13.40a}$$

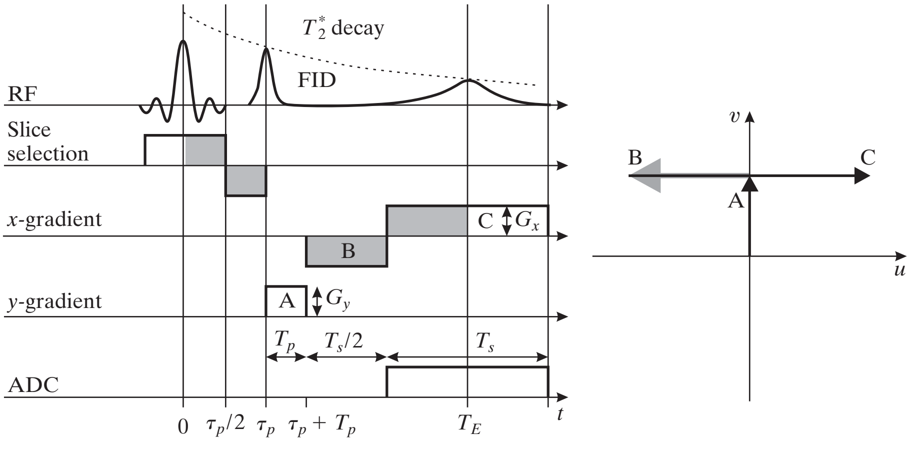
$$v = \bar{\gamma}G_y T_p\,, \tag{13.40b}$$

and

$$F(u,\bar{\gamma}G_y T_p) = s_0\left(\frac{u}{\bar{\gamma}G_x}\right), \quad 0 \le u \le \bar{\gamma}G_x T_s\,. \tag{13.41}$$

A 2-D Gradient Echo Pulse Sequence It is common to combine phase encoding, gradient echoes, and frequency encoding in a single pulse sequence. Figure 13.16(a) shows a pulse sequence identical to that of Figure 13.15(a), except that the FID readout has been replaced with a gradient echo readout. In Fourier space [Figure 13.16(b)], there are three distinct movements corresponding to the A, B, and C segments in the pulse sequence. Only the C segment corresponds to actual data collection. This pulse sequence is capable of acquiring an entire line of Fourier space from negative to positive frequencies. Analogous analysis to the previous pulse sequence reveals that the Fourier frequencies scanned are given by

$$F(u,\bar{\gamma}G_y T_p) = s_0\left(\frac{u}{\bar{\gamma}G_x}\right), \quad -\bar{\gamma}G_x T_s/2 \le u \le \bar{\gamma}G_x T_s/2\,. \tag{13.42}$$



(a) (b)

Figure 13.16
(a) A gradient echo pulse sequence and (b) its Fourier trajectory.

The pulse sequence depicted in Figure 13.16(a) would work, but it can be easily improved. A better pulse sequence would overlap the refocusing lobe, the phase-encoding gradient, and the gradient echo formation lobe. These, it turns out, can all be done at the same time, as their phase preparation processes are independent. The advantage is that the amplitude of the echo is significantly larger since T_E is shorter. This possibility is explored further in Problem 13.15.

EXAMPLE 13.6

A rectangular-shaped phase-encoding gradient is not practical because it has discontinuities at both ends (and large slew rates are required even to just approximate it). Since phase encoding depends on the area of the pulse rather than its specific shape, we can use waveforms with more practical shapes.

Question Suppose we want to use a sine-shaped gradient to achieve the same phase accumulation as a rectangular one with $G_y = 1$ G/cm and $T_p = 0.5$ ms. Assuming the same duration is required, what is the expression of the phase-encoding gradient and what is its maximum slew rate?

Answer The duration of the gradient is $T_p = 0.5$ ms, exactly half the period of a sine function. So the gradient can be expressed as

$$G_y(t) = G_{y\max} \sin(2\pi t)\,.$$

In order to achieve the same phase accumulation, we need

$$\begin{aligned} G_y T_p &= \int_0^{0.5\,\text{ms}} G_{y\max} \sin(2\pi t)\; dt \\ 0.5\ \text{G ms/cm} &= G_{y\max} \int_0^{0.5\,\text{ms}} \sin(2\pi t)\; dt \\ &= \frac{1}{\pi} G_{y\max}\,. \end{aligned}$$

So, the maximum value for the phase encoding gradient is $G_{y\,\max} = \pi/2$ G/cm. Therefore,

$$G_y(t) = \frac{\pi}{2} \sin(2\pi t) \text{ G/cm},$$

where t is in ms. The slew rate is the maximum value of $\left|\frac{dG_y(t)}{dt}\right|$, which can easily be computed to be

$$\text{SR} = \pi^2 \text{ G/cm/ms}.$$

■

13.2.7 Spin Echoes

We saw in Chapter 12 that spin echoes could be used to refocus spin systems that had been dephased due to T_2^* effects (static magnetic field inhomogeneities). By using spin echoes, an MR signal could be generated until it was completely lost due to T_2 effects (random perturbations of the magnetic field), which are not reversible. Because $T_2 > T_2^*$, the spin echo can produce a measurable signal long after the initial FID decays away. In this section, we will see how the spin echo can be used in an imaging sequence and its interpretation in Fourier space scanning.

Recall from Section 12.9 that a spin echo is generated by applying a 180° RF pulse shortly after the initial α pulse. Spins that were lagging behind are now ahead and vice versa. Consider what this means in terms of phase accumulation. Suppose we had deliberately applied a phase encode with gradient strength G_y just prior to the 180° pulse, which means that phases that had advanced due to their y position would now be lagging and vice versa. It would be as if we had applied a phase encode with the gradient strength $-G_y$. Now, suppose that we had prepared for a gradient echo by applying an x-gradient with strength $-G_x$ and then applied a 180° RF pulse. Again, the spins that were leading would now be lagging, and it would be as if we had applied an x-gradient with strength G_x. In both cases, the effect of the 180° pulse is to interchange the meaning of lead and lag, and this happens in both the x and y directions at the same time.

The effect of a 180° pulse in Fourier space can be understood by carefully interpreting some of our words from the previous paragraph. We stated that the position in Fourier space after the 180° pulse would be as if both the phase encode and the gradient echo preparation pulse had the opposite sign. Therefore, the 180° pulse causes a sudden change of sign of the frequencies corresponding to both the x and y directions, which is a reflection through the origin in Fourier space. If the position had been (u, v) prior to the pulse, it would now be $(-u, -v)$. This is a very rapid repositioning in Fourier space, costing only the time it takes to execute a 180° pulse.

Given this description, it is straightforward to construct a pulse sequence incorporating a spin echo. One possibility is shown in Figure 13.17. Section 1 in panel (a) performs the usual slice selection. Section 2 creates a phase encode, moving the Fourier frequency straight up from the origin, as shown in panel (b). Section 3 performs an x prephasing gradient, moving the Fourier location into the first quadrant, as shown in panel (b). Section 4 is a 180° RF pulse executed during a slice selection gradient. Refocusing of this slice selection gradient is not necessary because of a peculiarity of the 180° pulse. In particular, phase accumulated during the first half of the pulse is exactly balanced in the

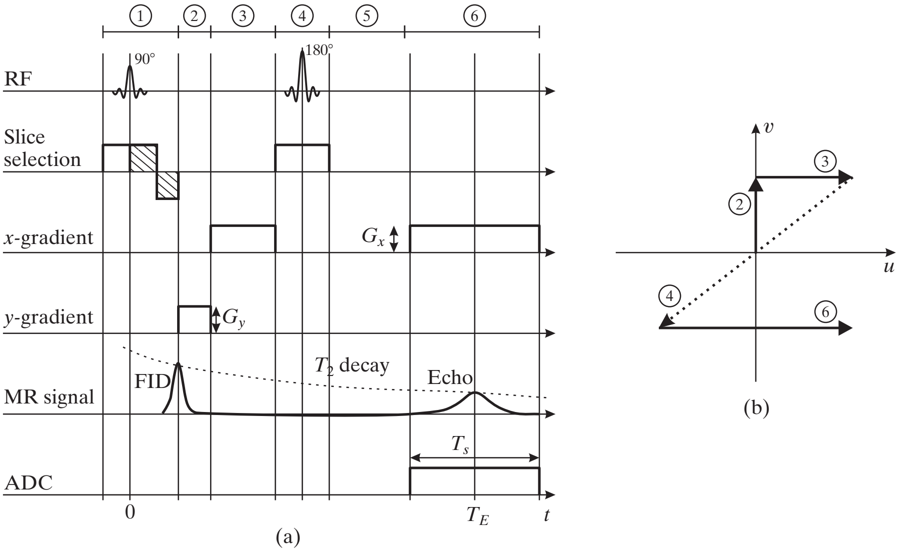


Figure 13.17
(a) A pulse sequence diagram for a spin echo image acquisition and (b) its Fourier space trajectory.

second half of the pulse because the spins are reversed. Section 5 is a required waiting period, a dead-time during which an echo is formed. Section 6 is the conventional frequency-encoded readout, scanning across Fourier space while collecting data.

EXAMPLE 13.7

Spin echo pulse sequences can be used to measure T_2 values of a sample of a homogeneous tissue. Consider Figure 13.17 and let $G_x = 0$.

Question What is the baseband signal at time $t = T_E$ for a sample of a homogeneous tissue?

Answer From (13.22) and (13.27), we can determine that the baseband MR signal is

$$s_0(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} AM_{xy}(x, y; 0^+)e^{-t/T_2(x,y)}e^{-j2\pi\gamma\!\!\!^{-} G_x xt}\,dx\,dy\,.$$

Since the sample is homogeneous, we have $M_{xy}(x, y; 0^+) = M_{xy}$ and $T_2(x, y) = T_2$. Since $G_x = 0$, the baseband signal at time $t = T_E$ is

$$s_0(T_E) = AM_{xy}e^{-T_E/T_2}\Delta A\,,$$

where ΔA is the area of the cross section of the sample in the x-y plane. Now, suppose we use the same pulse sequence with a different echo time T'_E to image the same sample. The baseband signal at time $t = T'_E$ measured is

$$s'_0(T'_E) = AM_{xy}e^{-T'_E/T_2}\Delta A.$$

The value of T_2 can be solved as

$$T_2 = \frac{T'_E - T_E}{\ln\left[\dfrac{s_0(T_E)}{s'_0(T'_E)}\right]}\,.$$

■

13.2.8 Pulse Repetition Interval

The gradient echo and spin echo pulse sequences have the potential to acquire all of 2-D Fourier space (at least in a sampled fashion). In order to do so, however, the basic pulse sequence must be repeated with different scan parameters. For example, the polar scan method must be repeated using different readout orientations θ. Both the gradient echo and spin echo techniques must be repeated using different phase-encoding gradient values G_y. The duration of the interval between such repetitions is called the *pulse repetition time* and is given the symbol T_R.

Our development of MR imaging equations (Section 13.3) will depend on whether the sequence is presumed to be either a *slow imaging sequence* or a *fast imaging sequence*. The slow imaging regime assumes $T_R \gg T_2$. In this case, the transverse magnetization has completely disappeared before application of the next α pulse, and there is no possibility of producing an echo from a previous excitation. The Fourier analysis of successive pulse sequences can correctly treat each one as if it starts from the Fourier origin. This is a little like ultrasound imaging, where T_R must be large enough so echoes cannot be received from previous pulses and thereby confused with echoes coming from present pulses.

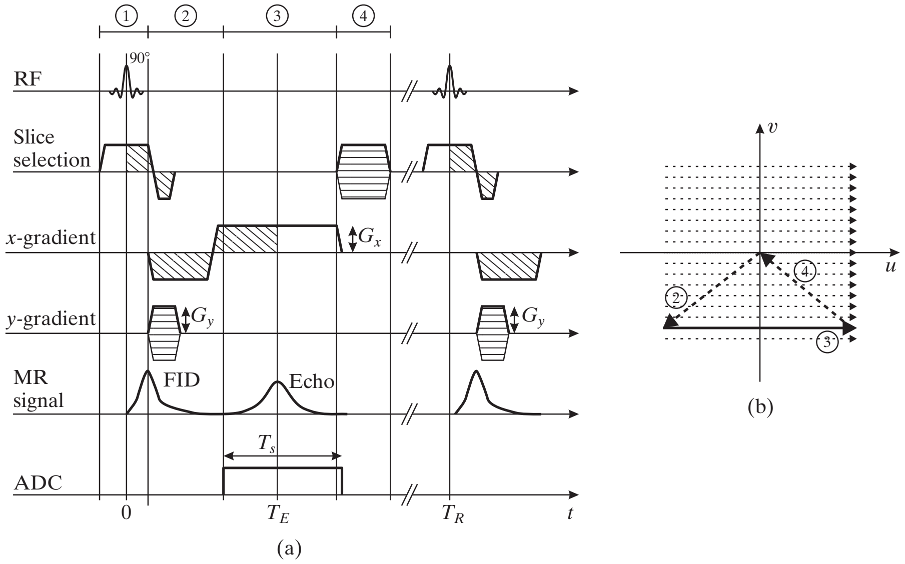
The fast imaging regime, however, breaks down this ‘‘barrier’’ using several different techniques, including the acquisition of multiple spin echoes or gradient echoes from a single initial RF excitation, and *spoiling* the transverse magnetization prior to subsequent excitation. Based on the previous discussion, we can concoct several schemes to acquire multiple spin echoes or gradient echoes. Basically, the ‘‘driving around in Fourier space’’ idea works fine, provided that we think correctly about compensating for spins that need to be rephased and continuing loss of signal due to T_2 and T_2^* effects. We do not have space to cover these techniques within the main body of the book, but we do offer several examples that can be worked out in the problems.

The idea of *spoiling* is something new. In this approach, a special additional gradient pulse called a *spoiler pulse* is used just prior to each successive excitation in order to guarantee that no echoes will form from previous excitations. The basic idea is to apply a z-gradient in order to dephase the spins in the z direction so that spins integrated over the thickness of the slice will add destructively and no signal can be produced. A different value of the spoiler gradient is used with each successive T_R interval so that there will be no possibility that the spins would actually be forced to line up again at some point later in the sequence. Gradient echo sequences that use this technique are called *spoiled gradient echo* (SPGR) image sequences.

13.2.9 Realistic Pulse Sequences

We now present three pulse sequences that might be implemented in practice. These pulse sequences are based on the principles developed in the previous section but show adherence to slew rate limitations, take advantage of overlap, and use alternate wave shapes where area is the only consideration.

Two-Dimensional Gradient-Echo Pulse Sequence A theoretically correct 2-D gradient-echo pulse sequence is shown in Figure 13.16. There are several



(a) (b)

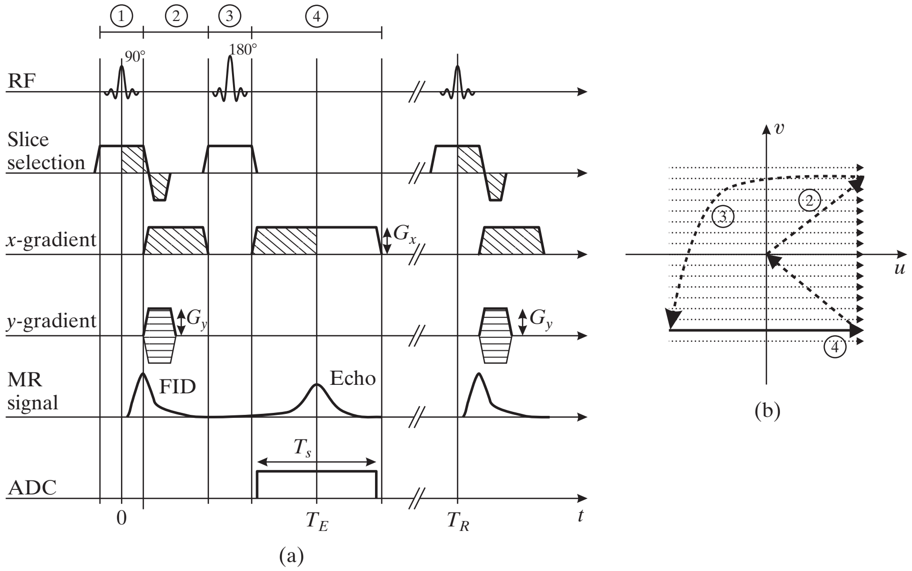
Figure 13.18
(a) A realistic gradient-echo pulse sequence and (b) its Fourier space trajectory.

aspects of this pulse sequence that are unrealizable, impractical, or undesirable. Figure 13.18 gives a more realistic pulse sequence for the following reasons. First, in this pulse sequence, all gradient waveforms are represented as trapezoidal pulses, which reflects the fact that there is a limited slew rate on the gradient amplifiers. Second, the slice selection refocusing lobe, the phase encode pulse, and the readout prefocusing pulse are all done at the same time. This does not cause any problem since these processes are independent. This practice saves time and puts the initial scan position in the correct location in Fourier space (although the exact Fourier trajectory taken from the origin to the final position might be quite complicated).

The third reason that Figure 13.18 is a more realistic pulse sequence is that it acknowledges the fact that the basic pulse sequence must be repeated with different phase encode values. The start of the second pulse sequence is shown after the breaks appearing on the time axes, and the repetition interval is explicitly shown on the pulse diagram as T_R. The breaks in the time axes indicate that time may elapse before starting the next excitation. The different phase encode values are indicated using a glyph that shows the basic waveform as a bold trapezoidal envelope and the other waveforms as lighter trapezoids. The Fourier diagram in Figure 13.18(b) indicates the Fourier trajectory in a single excitation as well as additional trajectories for multiple excitations. It is clear that this pulse sequence is capable of covering Fourier space after multiple excitations.

A fourth aspect of Figure 13.18 that is different from Figure 13.16 is the additional z-gradient during time period 4. This is the spoiler gradient discussed above, and it is often used to produce a faster imaging sequence. Like the phase encode gradient, the spoiler uses a different value with each repetition, as depicted by the bold and lighter lines.

Two-Dimensional Spin-Echo Pulse Sequence Having seen the elements that make up a more realistic 2-D gradient echo pulse sequence, the more realistic



(a)

(b)

Figure 13.19
(a) A realistic spin-echo pulse sequence and (b) its Fourier space trajectory.

spin-echo sequence shown in Figure 13.19 should come as no surprise. As in the gradient-echo sequence, gradient waveforms are shown as trapezoids, reflecting the actual limitations of MR gradient amplifiers. We depict a repetition of the basic pulse sequence using breaks in the time axes and indicating the pulse repetition interval T_R. There are also special glyphs to depict the use of multiple phase encodes over successive excitations.

The spin-echo sequence cannot use a spoiler since successive excitations need to create true echoes. Fast imaging techniques using spin echoes exist, and some discussion of these is presented in the problems.

Two-Dimensional Polar Imaging A realistic 2-D polar imaging pulse sequence is shown in Figure 13.20. We show a spin-echo sequence here, which is typical for polar imaging, but we could design a gradient-echo or spoiled gradient-echo sequence as well. As in both of the realistic sequences shown above, we have indicated the required repetition of the basic pulse sequence using both breaks in the time axes and glyphs that indicate the changing waveforms with each excitation.

The x- and y-gradients, which are used together for both phase encoding and frequency encoding (readout), in a polar pulse sequence must be chosen so that the correct orientation is scanned. In order to scan angle θ_n during the nth excitation, following (13.35), we see that the gradients must be selected as follows:

$$\theta_n = \tan^{-1} \frac{G_{y,n}}{G_{x,n}} . \tag{13.43}$$

13.3 Image Reconstruction

Since the data acquired from 2-D MR imaging pulse sequences can be interpreted as scans of Fourier space, the image reconstruction algorithm in MRI is the

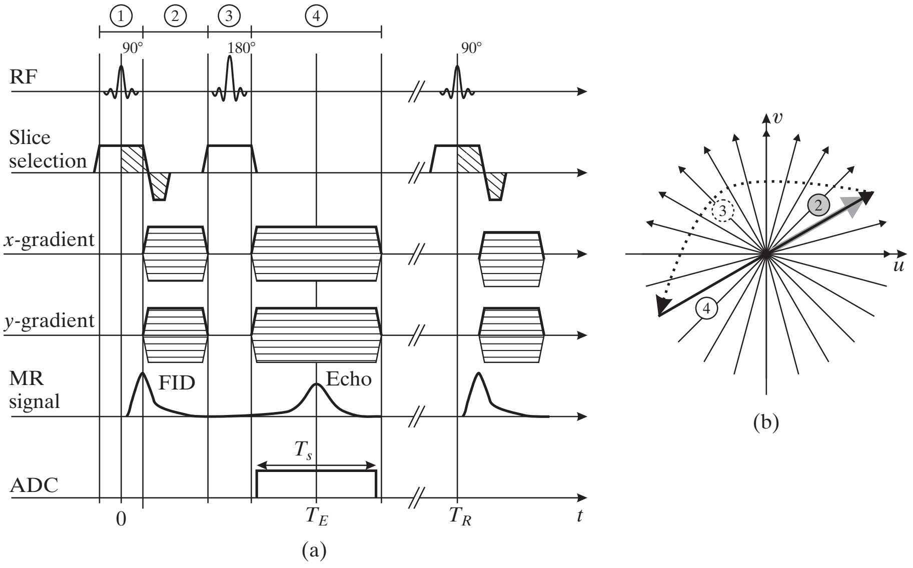


Figure 13.20
(a) A realistic spin-echo polar pulse sequence and (b) its Fourier space trajectory.

inverse Fourier transform. Since we have presented both rectilinear and polar data acquisition, we present the corresponding reconstruction algorithms.

13.3.1 Rectilinear Data

For conventional pulse sequences that acquire data in a rectilinear fashion, such as the gradient-echo or spin-echo sequences outlined above, the baseband signal is a temporal waveform that is dependent on the area A_y of the phase encode gradient (which can be negative). With reference to (13.39), the baseband signal can be written as

$$s_0(t, A_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) e^{-j2\pi\bar{\gamma} G_x x t} e^{-j2\pi\bar{\gamma} A_y y}\, dx\, dy\,, \tag{13.44}$$

where the dependence on both t and A_y has been made explicit. We identify the Fourier frequencies as [see (13.40)]

$$u = \bar{\gamma} G_x t\,, \tag{13.45a}$$

$$v = \bar{\gamma} A_y\,, \tag{13.45b}$$

which allows us to identify the Fourier transform of $f(x,y)$ as

$$F(u,v) = s_0\left(\frac{u}{\bar{\gamma} G_x}, \frac{v}{\bar{\gamma}}\right), \quad 0 \le u \le \bar{\gamma} G_x T_s\,. \tag{13.46}$$

An MR image is reconstructed as the inverse 2-D Fourier transform of $F(u,v)$, which from (13.46) can be written as

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s_0\left(\frac{u}{\bar{\gamma} G_x}, \frac{v}{\bar{\gamma}}\right) e^{+j2\pi(ux+vy)}\, du\, dv\,. \tag{13.47}$$

This equation is truly fundamental in MRI. Its simplicity, however, hides so many practical aspects of MRI that we can be easily lulled into a false sense of understanding. For example, we do not really know what $f(x, y)$ represents at this stage in our presentation. The meaning of $f(x, y)$ is not as conceptually straightforward as the linear attenuation coefficient in CT, the radioactivity concentration in SPECT, or even the reflectivity in ultrasound. In fact, $f(x, y)$ is determined by several tissue NMR properties as well as certain pulse sequence parameters. We will have more to say about this in Section 13.3.3.

Another aspect of (13.47) that must be addressed is the fact that both frequency variables, u and v, must be sampled. This leads to the practical requirement that (13.47) be computed using a 2-D discrete inverse Fourier transform algorithm [typically based on the fast Fourier transform (FFT)]. In this book, we do not require a background in digital signal processing, so we will not discuss the specifics of algorithm implementation. However, the fact that Fourier space is *sampled* in MRI is critical to image quality and will be discussed in Section 13.4.

EXAMPLE 13.8

A 2-D MR image is reconstructed from data collected by sampling Fourier space. Assume that data are acquired using a spin-echo pulse sequence (Figure 13.19), and a square region of Fourier space centered at $(u, v) = (0, 0)$ is sampled in a rectilinear fashion.

Question The parameters are set to be $T_R = 50$ ms, $T_s = 3$ ms, $G_x = 1$ G/cm, $\Delta G_y = 0.1$ G/cm, $\gamma\!\!\!\!- = 4.258$ kHz/G, and $T_p = 0.3$ ms. How many lines are acquired in Fourier space? How long does it take to acquire the entire image?

Answer The Fourier frequency is given by

$$u = \gamma\!\!\!\!- G_x t\,.$$

Since $G_x = 1$ G/cm and $T_s = 2$ ms, the range of acquired u is

$$u_{\max} - u_{\min} = \gamma\!\!\!\!- G_x T_s = 12.774\ \text{cm}^{-1}.$$

The square coverage in Fourier space requires $v_{\max} - v_{\min} = 12.774$ cm^{-1}. Since we have a fixed T_p and $\Delta G_y = 0.1$ G/cm, the adjacent lines in Fourier space are separated by

$$\Delta v = \gamma\!\!\!\!- \Delta G_y T_p = 0.128\,\text{cm}^{-1}.$$

In order to cover the square in Fourier space, we need to acquire

$$N = \frac{12.774\,\text{cm}^{-1}}{0.128\,\text{cm}^{-1}} = 100\,\text{lines}\,.$$

The imaging time is therefore

$$T = NT_R = 5\,\text{s}\,.$$

■

13.3.2 Polar Data

Pulse sequences that acquire polar data in Fourier space have baseband signals that depend on both time and orientation, $s_0(t, \theta)$, where θ is given by (13.35). In this case, the frequency variable is identified in polar coordinates as

$$\varrho = \gamma\!\!\!\!- \sqrt{G_x^2 + G_y^2}\ t\,. \tag{13.48}$$

From the development in CT, we know that the projection-slice theorem equates a "slice" of $F(u, v)$ to the Fourier transform of $g(\ell, \theta)$, which is a projection of $f(x, y)$. That is,

$$G(\varrho, \theta) = F(\varrho \cos\theta, \varrho \sin\theta)\,, \tag{13.49}$$

where

$$G(\varrho, \theta) = \mathcal{F}_{1\mathrm{D}}\{g(\ell, \theta)\}\,, \tag{13.50}$$

and

$$g(\ell, \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(\ell - x\cos\theta - y\sin\theta)\, dx\, dy\,. \tag{13.51}$$

The baseband signal is a polar scan in Fourier space, which yields

$$G(\varrho, \theta) = s_0\left(\frac{\varrho}{\gamma\!\!\!^{-}\sqrt{G_x^2 + G_y^2}}, \theta\right)\,. \tag{13.52}$$

From the 2-D Radon transform theory developed in Chapter 6, we know that $f(x, y)$ can be reconstructed using filtered backprojection [see (6.23)], which is given by

$$f(x, y) = \int_0^{\pi}\left[\int_{-\infty}^{\infty} |\varrho| G(\varrho, \theta) e^{j2\pi\varrho\ell}\, d\varrho\right]_{\ell = x\cos\theta + y\sin\theta} d\theta\,. \tag{13.53}$$

We know that convolution backprojection is an equivalent approach [see (6.26)]. Also, polar-to-rectangular interpolation methods could be used to form a rectilinear array of data, and the reconstruction method would thus be a standard 2-D inverse Fourier transform.

13.3.3 Imaging Equations

We introduced the concept of MRI *contrast mechanisms* in Section 12.10. We explained that the basic tissue parameters P_D, T_2, and T_1 are responsible for contrast and that the pulse sequence parameters—α, T_E, and T_R—can be manipulated to change the image contrast and, hence, the appearance of the reconstructed MR images. Table 13.1 summarizes our findings from the previous chapter describing the basic way to create contrast that is primarily dependent on P_D, T_2, and T_1. Please look back at Figure 12.10 to remind yourself how dramatically different the same cross section can look depending on the contrast mechanism used to create the image.

Although the list in Table 13.1 is valid, the given scanner parameters are not the only way to produce the target contrast. In particular, the changes in contrast due to tip angle α are not reflected in this table. To get a much more complete understanding of the possibilities, it is necessary to derive an *imaging equation* for a given pulse sequence, which we now develop. It should be noted that these three contrasts are not the only possible contrast mechanisms in MRI. Other key contrast mechanisms in MRI include T_2^*, flow, susceptibility, diffusion, and chemical shift. We provide an introduction to the susceptibility

TABLE 13.1

Contrast Generation in Basic MRI

Contrast	Scanner Parameters
P_D	Long T_R, read FID or use short T_E
T_2	Long T_R, $T_E \approx T_2$,
T_1	Read FID or use short T_E, $T_R \approx T_1$

and diffusion contrasts in Section 13.5. Other topics are left for independent study; a good starting point is the bibliography given at the end of this chapter.

We have established that MRI reconstructs an image of the effective spin density $f(x, y)$, which is given by

$$f(x, y) = AM(x, y; 0^+)e^{-t/T_2(x,y)}\,. \tag{13.54}$$

The first fact we need to realize is that t should be replaced by the time at which the signal is sampled. In the case of FID sampling $t = 0$, while for echo sampling $t = T_E$. Now consider a pulse sequence that issues a steady succession of α RF excitation pulses and assume that $T_R \gg T_2$ (which implies that the transverse magnetization is gone by the time each successive RF excitation is applied). If $T_R \approx T_1$ (at least for some tissues), then the sample will not be in equilibrium; that is, the longitudinal magnetization will not reach its equilibrium value before the next excitation pulse. But each pixel will establish a steady-state z magnetization in a certain sense. We now explore the concept of steady-state pulse sequences.

From (12.35), we know that the first α pulse will yield the longitudinal magnetization

$$M_z(t) = M_z(0^+)e^{-t/T_1} + M_0(1 - e^{-t/T_1})\,, \tag{13.55}$$

where it is understood in the context of imaging that M_0, M_z, and T_1 are functions of x and y within an excited image slab. If an α pulse is used, we have

$$M_z(0^+) = M_z(0^-)\cos\alpha\,, \tag{13.56}$$

and $M_z(0^-) = M_0$ for the first pulse. So,

$$M_z(t) = M_0 \cos\alpha e^{-t/T_1} + M_0(1 - e^{-t/T_1})\,, \tag{13.57}$$

which at the moment of the second excitation has the value

$$M_z(T_R) = M_0 \cos\alpha e^{-T_R/T_1} + M_0(1 - e^{-T_R/T_1})\,. \tag{13.58}$$

This longitudinal magnetization will be tipped by α in the next excitation and will then undergo further recovery.

It is straightforward to show that the magnetization just prior to the nth α pulse must obey

$$M_z(nT_R) = M_z([n-1]T_R)\cos\alpha e^{-T_R/T_1} + M_0(1 - e^{-T_R/T_1})\,. \tag{13.59}$$

Equation (13.59) is a difference equation that can be solved for its steady-state value $M_z^\infty(0^-)$. In particular, the steady-state value must obey

$$M_z^\infty(0^-) = M_z^\infty(0^-) \cos\alpha e^{-T_R/T_1} + M_0(1 - e^{-T_R/T_1}), \tag{13.60}$$

which, after some algebra, yields

$$M_z^\infty(0^-) = M_0 \frac{1 - e^{-T_R/T_1}}{1 - \cos\alpha e^{-T_R/T_1}}. \tag{13.61}$$

Looking back at the effective spin density [(13.54)], we realize that the longitudinal magnetization in (13.61) is tipped by α during RF excitation in order to generate a transverse component—the source of the NMR signal. Therefore, the reconstructed spin density for a steady-state pulse sequence can be written as

$$f(x, y) = AM_z^\infty(x, y; 0^-) \sin\alpha e^{-T_E/T_2(x,y)}, \tag{13.62}$$

where $T_E = 0$ in the case of FID imaging. Incorporating (13.61) yields

$$f(x, y) = AM_0 \sin\alpha e^{-T_E/T_2(x,y)} \frac{1 - e^{-T_R/T_1}}{1 - \cos\alpha e^{-T_R/T_1}}, \tag{13.63}$$

which is a very common imaging equation for MR imaging. It should always be remembered that (13.63) applies for steady-state imaging when $T_R \gg T_2$.

EXAMPLE 13.9

From (13.63), we can see that the parameters, T_E, T_R, and α, can be optimized to obtain the best contrast between different tissues.

Question Suppose that two tissues have the same proton density and T_2 value but have different T_1 values, T_1^{b} and T_1^{f}. What is the optimal T_R value that provides the best local contrast of effective spin density, given that $\alpha = \pi/2$?

Answer The tip angle is $\alpha = \pi/2$. So the effective spin density for two tissues are

$$\begin{aligned} f^{\mathrm{b}}(x, y) &= AM_0 e^{-T_E/T_2}(1 - e^{-T_R/T_1^{\mathrm{b}}}), \\ f^{\mathrm{f}}(x, y) &= AM_0 e^{-T_E/T_2}(1 - e^{-T_R/T_1^{\mathrm{f}}}). \end{aligned}$$

The local contrast of the effective spin density is

$$C = \frac{f^{\mathrm{f}}(x, y) - f^{\mathrm{b}}(x, y)}{f^{\mathrm{b}}(x, y)} = \frac{e^{-T_R/T_1^{\mathrm{b}}} - e^{-T_R/T_1^{\mathrm{f}}}}{1 - e^{-T_R/T_1^{\mathrm{b}}}}.$$

By assuming $T_R \gg T_1^{\mathrm{b}}$, we have

$$C \approx e^{-T_R/T_1^{\mathrm{b}}} - e^{-T_R/T_1^{\mathrm{f}}}.$$

By taking the derivative of C with respect to T_R and setting it to zero, we have

$$\frac{dC}{dT_R} = -\frac{1}{T_1^{\mathrm{b}}} e^{-T_R/T_1^{\mathrm{b}}} + \frac{1}{T_1^{\mathrm{f}}} e^{-T_R/T_1^{\mathrm{f}}} = 0.$$

The optimal T_R for best local contrast is therefore

$$\hat{T}_R = \frac{T_1^{\mathrm{b}} T_1^{\mathrm{f}}}{T_1^{\mathrm{b}} - T_1^{\mathrm{f}}} \ln\left(\frac{T_1^{\mathrm{f}}}{T_1^{\mathrm{b}}}\right).$$

■

13.4 Image Quality

In MRI, there are many issues that must be understood in order to get a good sense of image quality. In this section, we will address the concepts of sampling, resolution, noise, SNR, and artifacts. We will find that the issues are a bit more involved than in most medical imaging modalities, and we can pursue only the most basic concepts in this book. Further details can be found in several of the references listed in the bibliography at the end of this chapter.

13.4.1 Sampling

We now understand that MRI can be interpreted as a procedure that samples the 2-D spatial Fourier space of the effective spin density in a plane. The most basic part of this acquisition process is the sampling of the baseband signal $s_0(t)$—an FID or echo—during a readout gradient using an ADC. Suppose N_a samples are acquired T seconds apart during a readout that has gradient strength G_x. Then, the ADC will acquire data for duration

$$T_s = N_a T\,, \tag{13.64}$$

a variable that is identified in all of our pulse sequences. The sampling rate for this process is

$$f_s = \frac{1}{T}\,, \tag{13.65}$$

and it is given the name *receiver bandwidth* (RBW) in MRI. This is because the ADC uses an antialiasing filter that cuts off frequencies outside the interval $[-f_s/2, f_s/2]$ in order to avoid temporal aliasing. The width of this interval is the receiver bandwidth—that is, RBW $= f_s/2 - -f_s/2 = f_s$.

Recall that the readout gradient encodes the Larmor frequencies of the spin system in the x direction. Accordingly, on one side of the field of view (FOV), the spins have a higher frequency during the readout interval, and on the other side they have a lower frequency. We know that the received signal $s(t)$ is demodulated using the Larmor frequency $\nu_0 = \gamma\!\!\!^{-} B_0$, which corresponds to the center of the FOV. Therefore, relative to the center of the FOV, the baseband signal frequencies on one side of the FOV will have positive frequencies and the other side will have negative frequencies during a readout interval.

Since the sampling process uses an antialiasing filter, the sampled baseband signal cannot represent frequencies outside the interval $[-f_s/2, f_s/2]$. This means that spins with higher (negative or positive) frequencies will be invisible—the FOV is cut off in the x direction by the readout ADC. Therefore, the size of the FOV in the x direction FOV_x is specified by the spatial range of the spin system with Larmor frequencies within the RBW. Accordingly,

$$\gamma G_x \text{FOV}_x = f_s\,, \tag{13.66}$$

which leads to the relationship

$$\begin{aligned} \text{FOV}_x &= \frac{f_s}{\gamma G_x} && (13.67\text{a}) \\ &= \frac{1}{\gamma G_x T}\,, && (13.67\text{b}) \end{aligned}$$

where the last equality follows from (13.65). By comparing the denominator of (13.67a) with (13.40a), it is evident that the quantity $\gamma G_x T$ represents a "step" Δu in Fourier space in the readout direction. Accordingly,

$$\Delta u = \gamma G_x T \tag{13.68}$$

is the separation of samples in k space in the u direction corresponding to a sampled MRI readout.

Sampling in the v direction is determined by the sequence of phase encode gradients that are used. Typically, a step size in phase encode gradient area ΔA_y is selected, and any particular phase encode waveform has an area that is an integer multiple of ΔA_y. Accordingly, we find that

$$\Delta v = \gamma \Delta A_y \tag{13.69}$$

and

$$\begin{aligned} \text{FOV}_y &= \frac{1}{\gamma \Delta A_y} && (13.70\text{a}) \\ &= \frac{1}{\Delta v}\,. && (13.70\text{b}) \end{aligned}$$

Despite the analogous expressions for FOV in the readout and phase-encode directions, there is a key difference between them that requires careful interpretation and implementation. In the readout direction, an antialiasing filter is used to remove frequencies higher than those in the readout FOV. This has the effect of obliterating the signal coming from tissues outside the readout FOV. In the phase encode direction, however, there is no antialiasing filter; the Fourier space lines that are sampled are essentially "point samples" with respect to the phase encode direction. So, if tissues actually exist outside the calculated phase-encode FOV, they will be aliased and will cause wraparound in the phase encode direction in the reconstructed images.

The aliasing problem caused by a too-large step size in phase encode gradients is analogous to the concept of aliasing presented in Chapter 3. Here, we are sampling Fourier space rather than image space, and the aliasing causes wraparound in image space rather than in Fourier space. For example, in a conventional axial brain image in which the phase encode direction corresponds to the back-to-front direction, too-large phase-encode sampling will cause the back of the brain to appear at the front and vice versa.

EXAMPLE 13.10

The FOV of an MR image is related to the sampling steps in Fourier space. Suppose the reconstructed MR image has same resolution and size in both x and y directions.

Question If we set $T_s = 3$ ms, $G_x = 1$ G/cm, and $f_s = 85.33$ kHz, how many samples are acquired for each readout gradient? What is ΔA_y? What is the size of the FOV?

Answer The duration of a readout gradient is $T_s = 3$ ms, and RBW is $f_s = 85.33$ kHz. So there are $N_a = T_s f_s = 256$ samples taken for each readout gradient. The sampling period is $T = 1/f_s = 11.72\,\mu$s. Since the reconstructed image has same resolution and size in both directions, we have

$$\Delta v = \Delta u = \gamma\!\!\!^{-} G_x T = \gamma\!\!\!^{-} \Delta A_y \,.$$

This means that, $\Delta A_y = G_x T = 11.72\,\mathrm{G}\,\mu\mathrm{s/cm}$. The FOV has the same size in both directions, which is

$$\mathrm{FOV}_x = \frac{f_s}{\gamma\!\!\!^{-} G_x} = \frac{85.33\,\mathrm{kHz}}{4.258\,\mathrm{kHz/G} \times 1\,\mathrm{G/cm}} = 20.04\,\mathrm{cm}\,.$$

■

13.4.2 Resolution

Viewing MRI as a Fourier imaging method makes a basic discussion of image resolution fairly straightforward. We know that pulse sequences acquire data from a region in Fourier space using a sequence of readouts [see Figures 13.18(b), 13.19(b), and 13.20]. If there is any Fourier information outside the region acquired by the pulse sequence (i.e., higher frequency information), then it is not imaged and is reconstructed as if it were identically zero. Therefore, a pulse sequence inherently represents a low-pass filtering process applied to the underlying image. This limits the achievable resolution of MRI.

Let us consider the conventional acquisition approaches that use rectilinear scanning, as in Figures 13.18 and 13.19. The low-pass filter implied by these methods is a rectangle in Fourier space with dimensions

$$U = N_x \gamma\!\!\!^{-} G_x T\,, \tag{13.71a}$$

$$V = N_y \gamma\!\!\!^{-} \Delta A_y\,, \tag{13.71b}$$

where N_x and N_y are the number of readout and phase encode samples acquired, respectively. The implied low-pass filter is therefore given by

$$H(u,v) = \mathrm{rect}\left(\frac{u}{U}\right)\mathrm{rect}\left(\frac{v}{V}\right)\,, \tag{13.72}$$

and the PSF, which is the inverse Fourier transform of $H(u, v)$, is therefore given by

$$h(x, y) = UV \operatorname{sinc}(Ux) \operatorname{sinc}(Vy) \,. \tag{13.73}$$

It is sufficient to consider the main lobe of the 2-D sinc function in (13.73) in order to define resolution. As we have often done throughout this book, we will approximate the FWHM of a sinc function as one half of the interval between the first two zeros. Accordingly,

$$\mathrm{FWHM}_x = \frac{1}{U} = \frac{1}{N_x \bar{\gamma} G_x T} \,, \tag{13.74a}$$

$$\mathrm{FWHM}_y = \frac{1}{V} = \frac{1}{N_y \bar{\gamma} \Delta A_y} \,. \tag{13.74b}$$

Incorporating the definitions in (13.68) and (13.69) yields

$$\mathrm{FWHM}_x = \frac{1}{N_x \Delta u} \,, \tag{13.75a}$$

$$\mathrm{FWHM}_y = \frac{1}{N_y \Delta v} \,, \tag{13.75b}$$

which are called the *Fourier resolutions* of MRI.

When polar sampling is used, the concept of resolution is conceptually the same as that in CT (see Section 6.4.1). In CT, two effects degraded spatial resolution: detector size and the ramp filter window function $W(\varrho)$. In MR, the limited trajectories being scanned in Fourier space—not the detector size—are relevant in degrading resolution. Suppose $S(\varrho)$ represents a rectangular window describing the frequency scan (assumed to be symmetric about the origin) of each polar scan. Then the reconstructed effective spin density is given by

$$\hat{f}(x, y) = f(x, y) * \mathrm{h}(r) \,, \tag{13.76}$$

where $r = \sqrt{x^2 + y^2}$. The point spread function is circularly symmetric and given by the inverse Hankel transform [see (2.108)]

$$\mathrm{h}(r) = \mathcal{H}^{-1}\{S(\varrho) W(\varrho)\} \,. \tag{13.77}$$

It is important to distinguish what is sometimes called the *pixel size* from the Fourier resolution. Our formulas for reconstructing $f(x, y)$ are stated in terms of continuous variables—both spatial and Fourier variables—although we know that they will be implemented digitally. Without assuming prerequisite knowledge of digital signal processing, it is difficult to get too deeply into the details of implementation, but the concept of pixel size can be clarified easily enough.

Observations at discrete u and v locations will be obtained by MR scanning, and the reconstruction formula, (13.47) or (13.53), will be discretized to accept only those observed values. Similarly, the x and y locations at which $f(x, y)$ is computed will also be discretized. The separation of the x and y samples,

Δx and Δy, define the dimensions of a pixel, and together they define the pixel size. These dimensions are completely arbitrary. Although we may have 256 by 256 Fourier space samples, we can choose to reconstruct J by J pixels in the image space, and J can be absolutely anything. Of key importance, however, is that the underlying Fourier resolution does not change as a function of J, Δx, and Δy, but only as a function of the image acquisition Fourier window.

EXAMPLE 13.11

The resolution of an MR image is related to the k space coverage by Fourier transform. Suppose we want to acquire a 256×256 image, with resolution of $1\,\text{mm} \times 1\,\text{mm}$.

Question Suppose the readout gradient is $G_x = 1$ G/cm, what is the RBW f_s? What is ΔA_y? What is the k space coverage? If T_R is 50 ms, what is the imaging time?

Answer The Fourier resolution in the x direction is

$$\text{FWHM}_x = \frac{1}{N_x \bar{\gamma} G_x T} = 1\,\text{mm}\,.$$

So the RBW is

$$f_s = \frac{1}{T} = N_x \text{FWHM}_x \bar{\gamma} G_x = 109\,\text{kHz}\,.$$

The Fourier resolution in the y direction is

$$\text{FWHM}_y = \frac{1}{N_y \bar{\gamma} \Delta A_y} = 1\,\text{mm}\,.$$

So we have

$$\Delta A_y = \frac{1}{N_y \bar{\gamma} \text{FWHM}_y} = 9.17\,\text{G}\,\mu\text{s/cm}.$$

The k space coverage is $V = U = 1/\text{FWHM}_x = 10\ \text{cm}^{-1}$. The imaging time is $T = 50\,\text{ms} \times 256 = 12.8$ s. ■

13.4.3 Noise

Noise in MRI arises from statistical fluctuations of the signal—e.g., FID or echo—sensed by the receiver coils. The dominant source of this noise is *Johnson noise*, which is generically caused by the thermal agitation of electrons or ions in a conductor. In MRI, Johnson noise arises predominantly from the electrolytes in the patient, but it can also arise within the receive coil and the electronics attached to it. The variance of the received noise can be characterized by

$$\sigma^2 = \frac{2k\mathcal{T}R}{T_A}\,, \tag{13.78}$$

where k is Boltzmann's constant, $\mathcal{T}$ is temperature, R is the effective electrical resistance that the receiver coil "sees," and T_A is the total acquisition time. It is important to note that T_A is the sum of all the readout times during which data are collected; this time is less than the total time it takes to complete a given pulse sequence since there are many intervals in which the ADC is not active.

Normally, it is desirable to reduce noise as much as possible. We do not have control over temperature $\mathcal{T}$ (in ordinary human imaging scenarios, anyway), but we can control both R (to some extent) and T_A. The effective resistance in both the RF coil and the electronics is ordinarily much smaller than that of the body, and it can be ignored. Therefore, to reduce R, we need to reduce the amount of patient seen by the RF coil. We learned earlier that there are different RF coil designs (e.g., body coils, head coils, and surface coils). Therefore, choosing a coil whose sensitivity pattern encompasses a smaller volume will inevitably reduce the noise in the receiver. For example, it is highly desirable to use a head coil when imaging the head, since the body coil (attached within the bore of the scanner) would see a large part of the thorax, thereby contributing noise to the signal. Coils for the extremities are even smaller, and they are therefore better for imaging small body parts. Surface coils, while typically seeing a very small volume, have an uneven sensitivity pattern. In such coils the MR intensity will vary across the image, producing a "shading artifact" that is usually undesirable.

For a body or head coil, which can be approximated as a solenoid (a multiturn loop of wire), an equation for R can be worked out (see Problem 13.22). Suppose the solenoid has N turns per unit length, a total length of L, and each turn has radius r_0. We can then determine that

$$R = \frac{\pi^3 \mu_0^2 \nu_0^2 N^2 L r_0^4}{2\rho} \,. \tag{13.79}$$

Here, ν_0 is the Larmor frequency, $\mu_0^2 = 4\pi \times 10^{-7}$ weber/amp-meter is the permeability constant, and ρ is the resistivity of the body. From (13.78) and (13.79), it seems that noise can be reduced by decreasing (1) the solenoid's length, (2) the number of turns per unit length, (3) the solenoid's radius, or (4) ν_0. If we reduce the length of the solenoid, then its imaging volume is made smaller, which may be undesirable in volumetric imaging. If we reduce its radius, then only a smaller object will fit within the coil. This is essentially the difference between a head coil and a body coil, although either coil can be used to create an image of the head, a head coil has lower noise than a body coil and is therefore preferable.

According to (13.79), reducing the number of coils per unit length seems like a good idea, but it turns out that this will also have the effect of reducing the signal and will not actually improve the image quality. Finally, reducing ν_0 is easily accomplished by using less current in the superconducting magnet, which in turn reduces B_0 and hence $\nu_0 = \bar{\gamma} B_0$. It turns out, however, that the NMR signal will be reduced faster, causing an actual degradation of image quality. To understand these two facts, we need to understand the parameters affecting signal strength in MRI and then look at signal-to-noise ratio (SNR).

EXAMPLE 13.12

Noise is sometimes thought to be controlled by the sampling rate (RBW) and that higher sampling rates cause an increase in the image noise.

Question Find an expression for MRI noise variance in terms of the sampling rate (RBW) f_s.

Answer The total acquisition time is

$$T_A = MT_s\,,$$

where M is the number of readouts and T_s is the total time of a readout (where we assume that all readouts have the same duration in a given pulse sequence). Assuming there are N_a samples in each readout with a time separation of T per sample,

$$T_s = N_a T\,.$$

Since

$$f_s = \frac{1}{T}\,,$$

we can write

$$T_A = \frac{MN_a}{f_s}\,,$$

and using this in (13.78) yields

$$\sigma^2 = \frac{2k\mathcal{T}Rf_s}{MN_a}\,. \tag{13.80}$$

It therefore appears that increasing f_s leads directly to an increase in image noise. But what has really happened is that the larger f_s leads to a faster acquisition of N_a samples, which in turn reduces both T_s and T_A. Shorter overall acquisition time is what leads to higher noise. However, had we simultaneously changed the readout gradient G_x while increasing the sampling rate in such a way that T_s remained constant, the noise would have remained constant as well. Alternatively, if we doubled f_s but scanned Fourier space twice and then averaged the results, the total acquisition time would be the same and the image noise would have once again remain unchanged. Thus (13.80) should not be used, even though it is technically correct, as it hides the true mechanism governing image noise in MRI. ■

13.4.4 Signal-to-Noise Ratio

We are interested to know what factors can be adjusted to reduce the noise, while simultaneously keeping the signal strong. This concept is embodied in SNR, of course, and since we already have an expression for the noise, we need only find an expression for the signal. Under the assumption of a homogeneous sample and uniform RF field, (12.26) gives the result we need. For convenience, we repeat this equation for signal strength here:

$$|V| = 2\pi\nu_0 V_s M_0 \sin\alpha\, B^r\,. \tag{13.81}$$

If the receive coil is a solenoid (as in the previous section), then the reference field strength B^r at the center of the coil, produced by a unit current I_0 passing through the coil, is given by

$$B^r = \mu_0 I_0 N\,, \tag{13.82}$$

where N is the number of turns per unit length. Substituting (13.82) into (13.81) yields

$$|V| = 2\pi\nu_0 V_s M_0 \sin\alpha\, \mu_0 N\,, \tag{13.83}$$

where we have assumed $|I_0| = 1$, since it is unity reference current.

An expression for SNR is found using (13.78) and (13.83):

$$\mathrm{SNR} = \frac{|V|}{\sqrt{\sigma^2}} = \frac{2\pi \nu_0 V_s M_0 \sin\alpha \, \mu_0 N}{\sqrt{2k\mathcal{T} R/T_A}} . \tag{13.84}$$

Substituting for R using (13.79) and M_0 using (12.4), and working through some algebra yields

$$\mathrm{SNR} = \frac{\gamma \hbar^2}{\sqrt{4\pi k^3}} \frac{2\pi \nu_0 P_D \sqrt{\rho}}{r_0^2 \sqrt{L\mathcal{T}^3}} V_s \sin\alpha \sqrt{T_A} . \tag{13.85}$$

This expression has three basic terms in the product. The first fraction comprises physical constants, which cannot be changed or affected. The second term comprises terms that are related to the object or the system design and perhaps can be changed or selected. For example, objects with larger proton densities or resistivities, or ones that experience larger static magnetic fields (hence, higher Larmor frequencies) will yield larger SNRs. As well, lowering the temperature or making the coil smaller (radius or length) will increase the SNR.

The third term in (13.85) comprises terms that are directly selectable by the user in a given imaging scenario. Here, V_s represents the size of the sample represented by a single reconstructed signal value—that is, the pixel volume. We can directly select pixel volume through both slice selection, FOV selection, and choosing the size J of the reconstructed image. Also, we have already long recognized that choosing the tip angle $\alpha = \pi/2$ maximizes the received signal. Finally, we can improve SNR by increasing the amount of time T_A we spend gathering Fourier data.

13.4.5 Artifacts

The most common form of image distortion is geometric distortion or warping. This arises when the gradient strength is not uniform across the entire FOV; for example, if the amplitude of the x-gradient falls off as a function of increasing z, in a coronal image the patient will seem to "pinch in" at the extreme values of z. Because of the demands to image faster, physically shorter, faster switching gradients are now often used; hence, there is often considerable drop-off in the gradients at the edges of a larger FOV. Software is used to correct for this distortion at the time of image reconstruction. However, the slice shape and thickness will change with changing gradient strength, and this effect cannot be corrected in reconstruction.

Ghosts are one of the most common artifacts seen in MR images. They are most commonly caused from changes in the object between Fourier space acquisitions—for example, motion due to breathing, swallowing, tremor, heart beat or other flow effects. Because of these changes, the acquired data do not correspond to the same object, and reconstruction based on these data produces faint signals in one or more acquisitions that are not present in others—these are the ghosts. Ghosting artifacts are reduced by imaging in a breath-hold, by acquiring data in synchrony with the heartbeat (i.e., gating on the ECG signal) or by acquiring images exceptionally fast. However, there are disadvantages to these solutions. Breath-holding can be uncomfortable or impossible for the patient, gating increases the overall image acquisition time and is sometimes difficult to reliably acquire, and fast imaging always suffers from reduced SNR.

Wraparound artifacts are caused by aliasing in the phase encode direction. It has previously been determined that

$$\mathrm{FOV}_y = \frac{1}{\Delta v},$$

where Δv is the Fourier space separation of the measured Fourier scan lines in the phase encode direction. If the object does not fit within the FOV in the y direction, then it will wrap around in the y direction, in exact analogy to the frequency wraparound of a conventional sampled signal (see Chapter 3). The main difference here is that this sampling is done in only one of two Fourier directions (the phase encode direction) and the resulting replication leading to wraparound is in the spatial domain rather than the Fourier domain. This does not happen in the readout direction because an antialiasing filter is employed during readout.

To eliminate the wraparound artifact, it is necessary to eliminate the object that is outside the FOV. This can be done by rotating the scan plane to put the shorter dimension of the object in the phase encode direction, or by using small receiver coils whose sensitivity profile matches the FOV. Outer volume saturation pulses can be used to eliminate unwanted signals, but they are effective only for a short duration.

One method of reducing scan time is to eliminate phase-encode steps that sample the higher spatial frequencies in Fourier space. Unfortunately, this produces a low-resolution image and may also produce a "ringing" artifact. Gibbs ringing or the so-called "truncation artifact" is a classic problem with poorly designed low-pass filters. If the cutoff of the higher frequencies occurs too abruptly, the sinc-shaped point spread function causes large oscillating sidebands at edges in the image. One can reduce these artifacts by applying a filter that has a smooth transition zone as one approaches the region where the high-frequency data are not acquired.

In Chapter 12, we saw that chemical shift changes the Larmor frequency of nuclei in a spin system depending on what chemical environment they are in. For proton imaging, we are primarily interested in two species of protons—those in water and those in fat. At 1.5 T, the resonance frequency of protons in water is approximately 225 Hz higher than that of protons in the CH_2 components of fat. Therefore, since the position in x is encoded by frequency, fat protons will be positioned slightly different than those in water. There will therefore be a relative position shift in the readout direction between fatty tissues and nonfatty tissues. In extreme cases, this can cause the position of fat to appear in the wrong place relative to the water image, making anatomy look abnormal when it is in fact normal. A very common method used to reduce this artifact is to suppress the fat signal; the inversion recovery strategy (i.e., the STIR pulse sequence), described in Section 12.10, is one way to accomplish this. Finally, magnetic susceptibility differences can cause changes in the local magnetic field felt in the sample. Typically, this leads to signal "dropouts" at air/tissue interfaces; however, in the following section we will see how pulse sequences designed to image susceptibility itself can be quite useful.

13.5 Advanced Contrast Mechanisms

Up to this point, we have described tissue contrast mechanisms that are considered "basic" because they use only imaging pulses, inversion pulses, gradients, and spin echoes to manipulate the spin system of the sample. Here, we present more advanced mechanisms for establishing tissue contrast where, in some cases, the contrasts themselves are determined by other properties of the tissue, not just the basic NMR properties P_D, T_1 and T_2.

Susceptibility-Weighted Imaging Susceptibility causes a variation in the magnetic field from the static magnetic field, and it is possible to image this effect using *susceptibility-weighted imaging* (SWI). From (12.13), we see that the change in magnetic field due to susceptibility is given by

$$\Delta B(x, y) = \chi(x, y) B_0 \, , \tag{13.86}$$

where $\chi(x, y)$ is the susceptibility, shown here to depend on its spatial location within the imaging plane. This change in field strength will produce a change in Larmor frequency

$$\Delta \nu(x, y) = \bar{\gamma} \chi(x, y) B_0 \, , \tag{13.87}$$

and it follows that the transverse magnetization of an FID given in (12.33) must be modified by adding $\Delta\nu(x, y)$ to ν_0, yielding

$$M_{xy}(t) = M_0 \sin \alpha e^{-j(2\pi(\nu_0 + \Delta\nu(x,y))t - \phi)} e^{-t/T_2} \, . \tag{13.88}$$

Thus, the primary effect of susceptibility is to change the phase of the transverse magnetization in a spatially dependent fashion.

In an ideal world, we could measure the susceptibility by computing the phase of the complex transverse magnetization and simply solving for $\Delta\nu(x, y)$ assuming ν_0 and ϕ were known. While it is reasonable to assume that ν_0 is known (because of careful calibration), it turns out that there are many reasons that ϕ cannot be known in general. In fact, phase inhomogeneity is a ubiquitous problem in MRI, but it turns out that in many cases the phase inhomogeneity is dependent on spatial position and not on time. In this case, a repeated gradient echo experiment can be used to "tease out" the spatial distribution of susceptibility. Let us explore how to do this.

Consider the readout part of a standard gradient echo pulse sequence. From (13.25), we see that the spatially dependent Larmor frequency will be

$$\nu(x, y) = \bar{\gamma}(B_0 + \chi(x, y) B_0 + G_x x) \, , \tag{13.89}$$

when susceptibility is present. Standard signal demodulation only accounts for the nominal Larmor frequency, ν_0, so the baseband signal [see (13.27)] is left with a phase coming from susceptibility, as follows

$$s_0(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-j2\pi\bar{\gamma}(\chi(x,y)B_0 + G_x x)t} e^{j\phi_0(x,y)} \, dx \, dy \, , \tag{13.90}$$

where $\phi_0(x, y)$ represents the spatially dependent (inhomogeneous) phase that comes from a variety of sources.

Now, it was previously shown that in a properly designed gradient echo pulse sequence the phase is zero for all spatial positions x at the echo time T_E

(see Section 13.2.5). But that will not be the case here because of susceptibility and the underlying phase. From (13.90), we can see that the remaining phase will be

$$\phi(x,y) = 2\pi\bar{\gamma}\chi(x,y)B_0T_E + \phi_0(x,y)\,. \tag{13.91}$$

Performing this imaging experiment twice with two different values of T_E gives two equations and two unknowns, $\chi(x,y)$ and $\phi_0(x,y)$. Solving this for $\chi(x,y)$ yields

$$\chi(x,y) = \frac{\phi_1(x,y) - \phi_2(x,y)}{\omega_0(T_E^{(1)} - T_E^{(2)})}\,, \tag{13.92}$$

where $T_E^{(1)}$ and $T_E^{(2)}$ are the two echo times and $\phi_1(x,y)$ and $\phi_2(x,y)$ are the corresponding measured phases.

Functional Magnetic Resonance Imaging Functional magnetic resonance imaging (fMRI) is based on the fact that blood hemoglobin has two chemical states, oxygenated and deoxygenated, and these two molecules have different magnetic susceptibilities. As discussed in Section 12.4, diamagnetic materials produce a slight decrease in the local magnetic field, while paramagnetic materials produce a slight increase. Oxygenated hemoglobin is slightly diamagnetic and has virtually the same susceptibility as most body tissues, thus producing no contrast in a susceptibility-weighted image. However, deoxygenated hemoglobin is paramagnetic, which provides contrast in a susceptibility-weighted image. Since active neurons in the brain use oxygen at a higher rate than inactive neurons, there tends to be a higher concentration of deoxygenated hemoglobin in areas of high brain activity. This means that there will be a higher fraction of paramagnetic molecules in these regions, which produces a change in the phase of the measured signal [see (13.90)]. This is referred to as the *blood oxygenation level dependent* (BOLD) signal that is measured in fMRI.

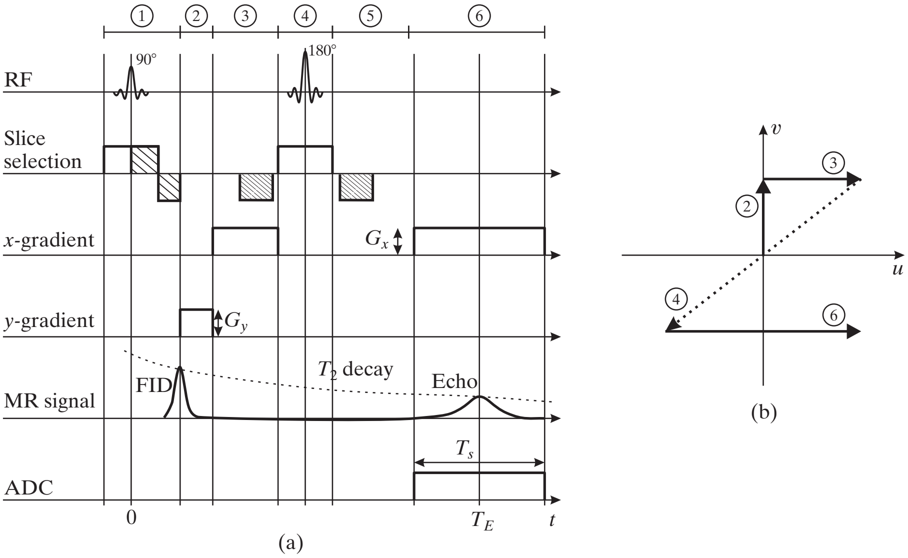
Because deoxygenated hemoglobin molecules are present in concentration within the blood, their paramagnetic property yields inhomogeneities in the Larmor frequency at the microscopic scale, which in turn reduces T_2^*. Thus, rather than directly measuring a change in local susceptibility (e.g., using the two-echo SWI method described above), fMRI is typically accomplished by using a pulse sequence that is sensitive to T_2^*—that is, a gradient-echo pulse sequence. The typical experiment first acquires an image when the subject is at rest and then acquires a second image while the subject is carrying out a mental task. When the task image is subtracted from the resting state image, the parts that are brighter indicate those regions of the brain containing higher concentrations of deoxygenated hemoglobin and are therefore more involved in the particular mental task.

Because of the difference in susceptibility of deoxygenated hemoglobin, the venous system in the brain has a different susceptibility than other tissues in the brain. Therefore, SWI, described above, can be used to image the veins in the brain. As well, because of the quantitative relationship between the phase of the measured signal and the actual susceptibility $\chi(x,y)$, it is possible to relate the measured susceptibility to the actual oxygen saturation in the blood.

This provides a much more direct and quantitative approach to fMRI, but it is technically very challenging and is being explored in research labs today.

Diffusion Magnetic Resonance Imaging We know that the ability to generate an NMR signal depends on an in-phase transverse magnetization within the sample (which is considered to a voxel in MRI). Both constants T_2 and T_2^* have the effect of dephasing the transverse magnetization and thereby reducing the induced signal. It turns out that *diffusion*—that is, the random movement of molecules, particularly prevalent in liquids—is another mechanism that can cause such dephasing. To see how this happens, consider turning on a z-gradient G_z after a 90-degree excitation pulse. We know that the Larmor frequencies across the slice (really a "slab" because there is some slice thickness) will be different and therefore dephasing across the slice will begin immediately. But if, in addition, there are molecules moving around in three dimensions, the phases at any given z position within the slice will take on additional randomness (beyond that determined by T_2) which cannot be recovered, even with a spin echo. Thus, if molecular diffusion is large, we would expect to see a significant drop in the received FID or echo signal relative to that of the signal in the absence of diffusion. We should therefore be able to measure local diffusion in tissues using MRI.

The usual way to measure diffusion in MRI is to use a *bipolar pulse* just after the excitation pulse and before the readout gradient, as shown in Figure 13.21. The first gradient in a bipolar pulse is long enough (and large enough) to permit diffusion within the tissues to cause measurable signal loss (exactly as described above). The second gradient reverses the effects of fixed dephasing across the image slice caused by the first gradient. In the illustrated pulse sequence, the bipolar pulses are applied in the slice-encode direction, so we can think of the second gradient as serving the same purpose as the refocusing lobe in slice selection. Although this measurement can be carried out using a gradient echo pulse sequence, the additional time required for diffusion



(a)

(b)

Figure 13.21
(a) A pulse sequence diagram for a diffusion-weighted image acquisition. The bipolar gradient pulse that encodes diffusion is shown in the "Slice Selection" gradient waveform during times 3 and 5. (b) The Fourier space trajectory of this pulse sequence.

encoding tends to cause excess signal losses due to T_2^*. So, Figure 13.21 depicts a spin echo pulse sequence with the 180-degree pulse occurring between the two bipolar pulses. The spin reversal caused by the 180-degree pulse is the reason for a spin echo and also the reason that the sign of second bipolar gradient is the same as that of the first gradient.

The simplest way to use a diffusion-weighted image is to simply subtract two images, one acquired without a bipolar pulse (a reference image) and the other acquired with a bipolar pulse. It is important to keep the timing of all the pulses to be the same, however, so that the T_2-weighting of the two images arises from the same echo time T_E. Since the diffusion-weighted image will have a smaller signal where there is substantial diffusion, the resultant subtracted image will be brighter where there is higher diffusion.

Although the presentation above captures the essence of diffusion-weighted imaging, actual use of the sequence and image data resulting from this sequence is far more involved in practice. For example, use of the bipolar pulse in the slice selection gradient represents only a single direction of diffusion sensitivity; in practice, these pulses can be applied to all three gradient coils at the same time in order to give an arbitrary direction of sensitivity. Because some tissues (most notably, the white matter of the human brain) show directionally dependent diffusion, the use of multiple directions can be used to compute a *diffusion tensor* at each voxel. Such an approach is called *diffusion tensor imaging* (DTI). Eigenanalysis of this tensor yields, among other things, a *principal eigenvector*, which characterizes the direction of greatest diffusion, and a *fractional anisotropy*, which characterizes the degree of directionality of the underlying tissue. The *mean diffusivity*, a single number representing the overall diffusivity regardless of direction, is also computed from the diffusion tensor at each voxel.

13.6 Summary and Key Concepts

MRI systems are very useful in medical imaging because they provide high-resolution anatomic information noninvasively, and they can image a variety of tissue properties that relate to the underlying composition and molecular environment of the tissues. Such systems represent large versions of NMR systems with significant additions to form images. In this chapter, we presented the following key concepts that you should now understand:

1. An *MRI scanner* consists of five principal components: the main magnet, a set of switchable gradient coils, RF coils, pulse sequence and receive electronics, and a computer.
2. The most common magnet is a *cylindrical superconducting magnet* with field strength ranging from 0.5 to 7 T.
3. The *gradient coils* produce the change in local magnetic field necessary to encode spatial location in the MR signal.
4. The *RF coils* or *resonators* receive the MR signals and may be large (e.g., a body coil) or small (e.g., a surface coil).
5. Manipulation of the gradient coils produces *frequency and phase encoding of location*; manipulation of the RF excitation pulse sequence produces *varying image contrast*.

6. *MR data* arise from scans of Fourier space; *MR image reconstruction* is based on the inverse 2-D Fourier transform and yields the distribution of *effective spin density*.
7. *MR image quality* depends on *contrast* (which itself depends on intrinsic tissue parameters and the choice of pulse sequence), *sampling*, and *noise*.
8. Careful control of the spin systems using both RF excitations and gradient pulses can produce images that encode information about *susceptibility*, including the BOLD effect in hemoglobin captured by *functional MRI*, and *diffusion*.

Further Reading

Bernstein, M.A., King, K.F., and Zhou, X.J. *Handbook of MRI Pulse Sequences*, Burlington, MA: Elsevier Academic Press, 2004.

Brown, M.A. and Semelka, R.C. *MRI Basic Principles and Applications*, 4th ed. Hoboken, NJ: Wiley, 2010.

Bushberg, J.T., Seibert, J.A., Leidholdt Jr., E.M., and Boone, J.M. *The Essential Physics of Medical Imaging*, 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2012.

Bushong, S.C. *Magnetic Resonance Imaging*," 3rd ed. St. Louis, MO: Mosby, 2003.

Haacke, E.M., Brown, R.W., Thompson, M.R., and Venkatesan, R. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. New York, NY: Wiley, 1999.

Liang, Z.P., and Lauterbur, P.C. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*. Piscataway, NJ: IEEE Press, 2000.

Macovski, A. Noise in MRI. *Magnetic Resonance in Medicine* 36 (1996): 494–97.

Stark, D., Bradley, Walter G., and Bradley, William G. (Eds.) *Magnetic Resonance Imaging*. St. Louis, MO: Mosby, 1999.

Problems

MR Imaging Instrumentation

13.1 A uniform magnetic field in the z direction is applied to a sample. When an x-gradient is applied, what change is made to the magnetic field?

13.2 Briefly explain the function of RF coils.

Encoding Spatial Position and MR Imaging Equation

13.3 You intend to image one slice ($\bar{z} = 5$ cm, $\Delta z = 1$ cm) of an off-centered cube having width 10 cm as shown in Figure P13.1. You are given $G_z = 1$ G/mm and $\bar{\gamma} = 4.258$ kHz/G.

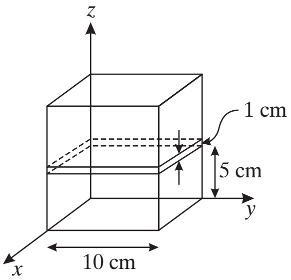


Figure P13.1
See Problem 13.3.

(**a**) Find the bandwidth (in Hz) of the RF waveform needed to perform the slice selection.

(**b**) Give a mathematical expression for the RF waveform $B_1(t)$ (in the rotating frame) that is needed to perform the slice selection.

13.4 (**a**) What is the slice thickness (defined as the full width at half maximum) if a slice-selection gradient amplitude of 1 G/cm is used with a Gaussian shaped RF pulse that has a shape given by $A(t) = A_0 \exp\{-t^2/\sigma^2\}$, where σ = 1 ms.

(**b**) What is the slice thickness if we cut the gradient amplitude in half?

Suppose we change the shape of the RF pulse so that σ is reduced by a factor of two.

(**c**) What is the new slice thickness (with a 1 G/cm gradient)?

(**d**) What else is affected by this change?

13.5 An RF signal is given by

$$s(t) = A\Delta\nu \,\mathrm{sinc}(\Delta\nu t) e^{-j2\pi\bar{\nu}t}.$$

Show that the tip angle distribution is given by (13.17).

13.6 Explain why a refocusing gradient is usually needed for slice selection, and why its duration is half the duration of the RF excitation when its waveform is a constant negative gradient of strength $-G_z$.

13.7 Substituting (13.22) into (13.44) and assuming T_2 is constant yields

$$\begin{aligned} s_0(t, A_y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Ae^{-t/T_2(x,y)} M(x, y; 0^+) e^{-j2\pi\gamma G_x x t} e^{-j2\pi\gamma A_y y}\, dx\, dy \\ &= Ae^{-t/T_2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M(x, y; 0^+) e^{-j2\pi\gamma G_x x t} e^{-j2\pi\gamma A_y y}\, dx\, dy\,. \end{aligned}$$

What is the effect of the term e^{-t/T_2} on the reconstructed image? (Assume that the pulse sequence in Figure 13.15 is used.)

13.8 Suppose a point object is moving along the x-axis with the following trajectory: $x(t) = x_o + vt$.

(**a**) Calculate the phase shift that is induced in the transverse magnetization after the application of the gradient waveform shown in Figure P13.2.

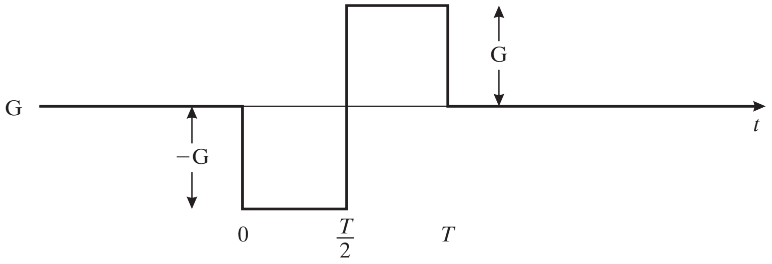


Figure P13.2
See Problem 13.8(a).

(**b**) Calculate the phase shift induced by the waveform shown in Figure P13.3. This is called a *flow compensation pulse*.

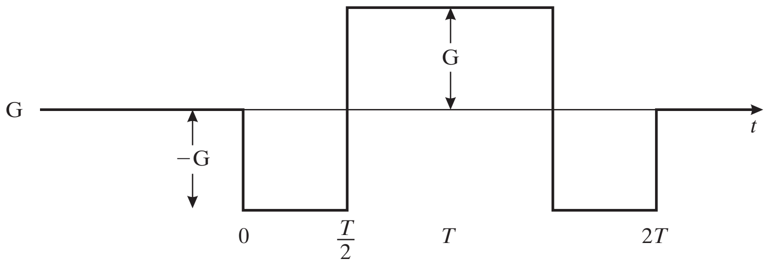


Figure P13.3
See Problem 13.8(b).

(**c**) Repeat the calculations for the above two waveforms for the trajectory given by $x(t) = x_o + vt + \frac{1}{2}at^2$.

(**d**) What gradient waveform could you use as an *acceleration compensation* pulse?

(**e**) Is it possible to design a gradient waveform that will produce phase shifts that are independent of acceleration but are dependent on the velocity? Explain.

13.9 Suppose that a slice has been selected through a 5 × 5 × 5 cm cube. We wish to produce a projection of this slice with the application of a readout gradient. The pulse sequence in Figure P13.4 is used to produce the profile.

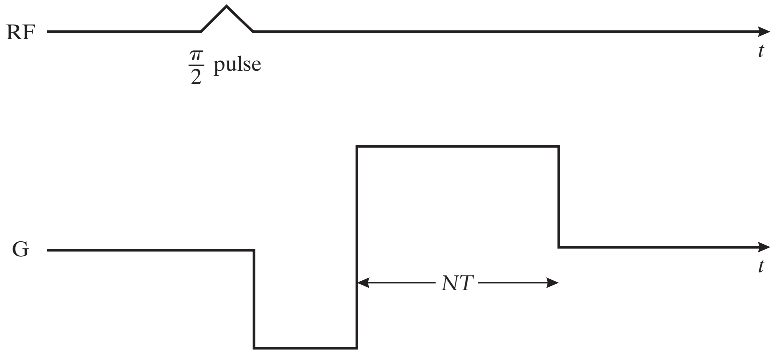


Figure P13.4
Problem 13.9

Suppose the gradient strength is G = 1 G/cm and $N = 256$ sample points are taken in the total sampling time $NT = 10$ ms (recall that T is the time between sampling points in the readout direction).

(**a**) What is the spatial extent (i.e., the FOV) of the profile? What is the bandwidth of the received signal?

(**b**) How many pixels does the profile of the object span?

(**c**) Suppose we set $G = 0.5$ G/cm, $N = 256$, and $NT = 20$ ms. How does this affect (a) and (b)?

(**d**) Repeat (a) to (c) for the pulse sequence shown in Figure P13.5.

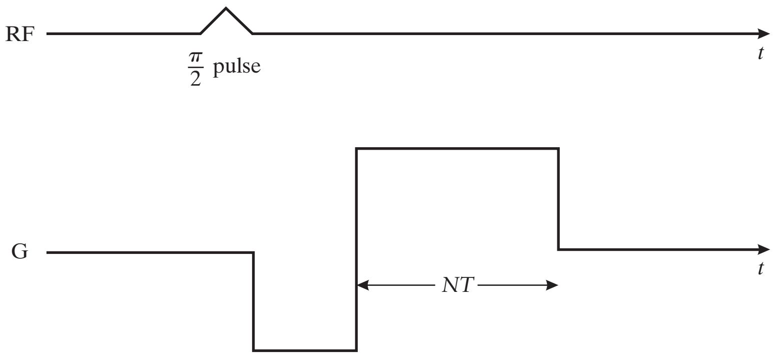


Figure P13.5
See Problem 13.9(d).

13.10 In order to produce an actual image of T_1, as opposed to one that is merely T_1-weighted, it is necessary to take at least two separate images and use them to compute a pixel-by-pixel estimate of T_1.

(**a**) Show that the imaging equation in (13.63) can be written as

$$\frac{f}{\sin\alpha} = e^{-T_R/T_1}\frac{f}{\tan\alpha} + AM_0 e^{-T_E/T_2}(1 - e^{-T_R/T_1})\,.$$

(**b**) Assume that three images f_1, f_2, and f_3 with tip angles α_1, α_2, and α_3, respectively, are acquired. Show that the points $(f_i/\tan\alpha_i,\ f_i/\sin\alpha_i)$, $i = 1, 2, 3$, should lie on a line.

(**c**) Using the fact shown in part (b), find a closed-form expression for an estimate of T_1 given only two images f_1 and f_2 with different tip angles α_1 and α_2.

Sampling the Frequency Space

13.11 Suppose a pulse sequence uses the x- and y-gradients shown in Figure P13.6. Draw its Fourier space trajectory, and carefully label the axes.

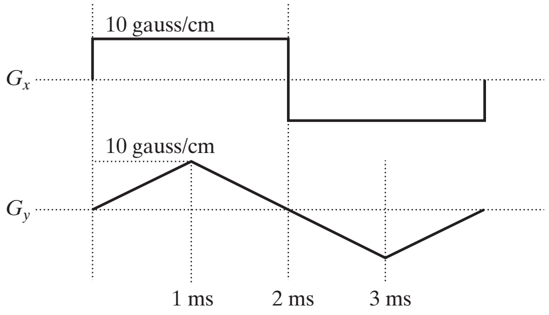


Figure P13.6
The x- and y-gradients in a pulse sequence for Problem 13.11.

13.12 Consider the pulse sequence shown in Figure 13.20. Suppose we want to sample a radial line from $(-0.25, -0.5)$ mm^{-1} to $(0.25, 0.5)$ mm^{-1} and measure 128 points along the line. Ignore the ramps at the ends of the gradients.

(**a**) If the duration of the x- and y-gradients in the interval labeled 2 are 0.1 ms, what are the strengths of the gradients?

(**b**) If $T_s = 10$ ms, what are the readout gradients? What is the sampling rate?

13.13 Draw the pulse sequence that scans two lines of Fourier space after an excitation RF pulse. Start scanning Fourier space at $(-1, 0.5)$ mm^{-1} and end at $(-1, 0.4)$ mm^{-1} as shown in Figure P13.7. Assume that the RF pulse is not slice-selective. Label your timing diagram including the amplitudes of the gradients.

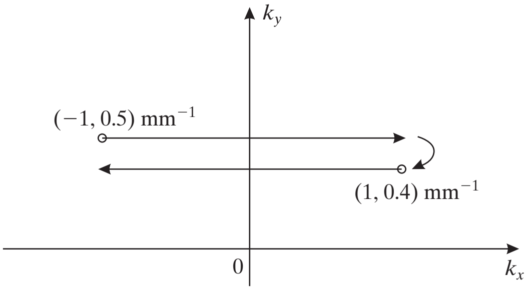


Figure P13.7
See Problem 13.13.

13.14 We have been asked to phase encode two point objects offset in y. Let us first turn to an example in which we have two point samples in the imaging plane that both have the x coordinate $x = x_o$, but they have different y coordinates. In this case, the signal is the sum of the individual signals from each object; therefore, we cannot determine their individual y coordinates from a single measurement such as that shown in Figure 13.15 and described by (13.42). Show that we can obtain *two* measurements, each measurement having a different value of the phase-encoding gradient amplitude G_y, to solve for y the position of the two point objects.

13.15 Explain why the refocussing lobe, the phase-encoding gradient, and the gradient echo formation lobe can all overlap in time. (Refer to Figure 13.18.)

MR Image Reconstruction

13.16 2-D projection imaging (i.e., polar scanning) was an early MR imaging technique that is enjoying a comeback because it is very fast. Consider the pulse sequence shown in Figure P13.8 and let

$$G_x = G\cos\theta\,,$$
$$G_y = G\sin\theta\,.$$

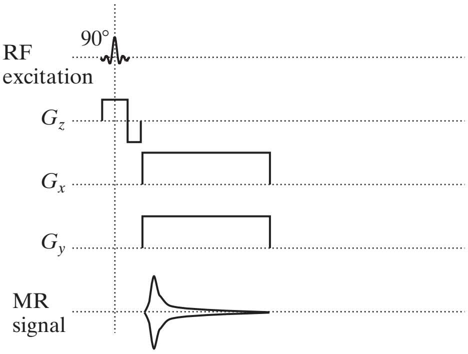


Figure P13.8
Pulse sequence for projection imaging. See Problem 13.6.

(**a**) Find an expression for the signal given θ.

(**b**) If the cross section being imaged is $f(x, y) = A\delta(x - 1, y) + B\delta(x, y + 1)$, sketch the recorded signal for $\theta = 0$, and $\theta = 90°$.

(**c**) How can one obtain an image of the cross section using this pulse sequence?

13.17 We wish to image two "point objects" that are sitting in slightly different static fields due to magnet imperfections. Point object 1 is in a field B_o and point object 2 is in a field $\mathrm{B}_o + \Delta\mathrm{B}$. Describe why this is a problem if we are going to use the 2-D projection method of MRI.

13.18 Consider the profile of an object in the readout direction. Show that the inverse Fourier transform of the baseband signal $s_0(t)$ given in (13.27) is a projection of the object $M_{xy}(x, y; 0^+)$ onto the x-axis convolved with the Fourier transform of the T_2 decay envelope $\mathcal{F}^{-1}\{e^{-t/T_2}\}$. (Assume that T_2 does not depend on x and y.)

MR Image Quality

13.19 We want to acquire an image of a sagittal slice (parallel to y-z plane).

(**a**) In what directions should the slice selection gradient, phase-encoding gradient, and frequency-encoding gradient be?

(**b**) How can one prevent aliasing in the phase-encoding and frequency-encoding directions?

13.20 We want to reconstruct a 2-D MR image with size of 256×256 pixels.

(**a**) What parameter(s) determine the spatial extent in the phase-encoding direction?

(**b**) What parameter(s) determine the spatial extent in the readout direction?

(**c**) What parameter(s) determine the spatial resolution in the phase-encoding direction?

(**d**) What parameter(s) determine the spatial resolution in the readout direction?

13.21 A cross section of a sample of protons within a 25.6 cm × 25.6 cm square is being imaged by an MR scanner. We want to reconstruct an image with matrix size 256×256. We use the Nyquist sampling rate in the u and v directions.

(a) What is the extent of the object in the diagonal direction? How many samples are there in each (main) diagonal direction?

(b) What is the Fourier space sampling rate for the diagonal directions?

13.22 A solenoid coil with N turns per unit length, a length of L, and a radius of r_0 is used to image an object of radius r_0 with uniform resistivity of ρ. The coil is excited with a current $I(t) = \cos(2\pi \nu_0 t)$.

(a) What is the average power dissipated in the object? What is the voltage induced for a cylindrical shell of radius r?

(b) Given that the differential conductance of a cylindrical shell with length L, radius r, and resistivity ρ is $dG = Ldr/(2\pi r\rho)$, show that the effective electrical resistance of a solenoid is given by (13.79).

13.23 What repetition time T_R should we use to generate the maximum signal difference between the two tissues whose T_1 values are T_1^a and T_1^b? What T_R value will give the maximum signal difference to noise ratio? (Assume that you can use a T_E of zero and that all transverse magnetization has decayed before each $\pi/2$ pulse.)

13.24 The resolution of optical imaging methods is equal to the Rayleigh limit, which is one half of the wavelength of the optical frequency being used. If MRI were performed using radio waves propagating into the body, what would be the predicted resolution for a conventional magnet?

13.25 The location of "fatty" tissues in an MR image is displaced relative to "watery" tissues because of chemical shift. Explain how chemical shift will affect the appearance of images containing both fat and water and develop an approach to "suppress" the fat image using the knowledge that $T_1(\text{H}_2\text{O}) \approx 4T_1(\text{fat})$.

13.26 It is often useful to consider what happens to image SNR when only a single imaging variable (or small group of imaging variables) is changed. Consider making each of the changes in the following parts and determine how the SNR changes as a result.

(a) $G_x \to 0.5G_x$.

(b) $N_y \to 2N_y$.

(c) $f_s \to 2f_s$ (assuming that $N_x \to 2N_x$ as well).

(d) $f_s \to 2f_s$ (assuming that N_x stays constant).

Applications, Extensions, and Advanced Topics

13.27 Consider the imaging problem in Problem 13.3. Assume the effective spin density within the cube is constant and equal to 1.

(a) Refer to Figure P13.1 and sketch the 2-D function $f(x, y)$ that you will image. Also sketch $|F(u, 0)|$.

You use the gradient echo pulse sequence shown in Figure P13.9 to image the slice with gradient $G_x = 0.5$ G/mm and $G_y = 0$.

(b) Find the duration (in seconds) of the x-gradient preceding the readout gradient in order to collect data in the range $-0.4\ \text{cm}^{-1} \leq u \leq 0.4\ \text{cm}^{-1}$.

(c) How long (in seconds) after the gradient G_x changes sign will the gradient echo occur and why?

(d) After reconstructing the data collected during the acquisition, will we able to get a perfect reconstruction of $g(\ell, 0^\circ)$, where $g(\ell, \theta)$ is the Radon transform of $f(x, y)$? Explain.

13.28 Consider a constant magnetic field $B_0 = 1.5$ T is applied to an object. The gyromagnetic ratio of the object is $\gamma = 2\pi \times 4,258$ (rad/s)/G. Suppose the following RF pulse is applied

$$B_1(t) = A\Delta f \operatorname{sinc}(\Delta f t) e^{j2\pi\nu_0 t} \operatorname{rect}\left(\frac{t}{\tau_p}\right),$$

where $\Delta f = 4.258 \times 10^4$ Hz, $A = 2$ G/Hz, $\tau_p = 2$ ms, and ν_0 is the Larmor frequency.

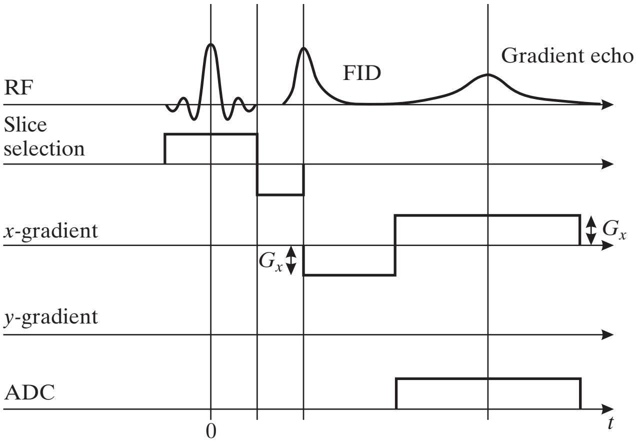


Figure P13.9 Pulse sequence for Problem 13.27.

(a) Find ν_0. Write down an equation to calculate the tip angle.

(b) Assume τ_p is long enough so that $B_1(t)$ could be approximated as

$$B_1(t) = A\Delta f \operatorname{sinc}(\Delta f t) e^{j2\pi\nu_0 t}.$$

What is the spectrum of $B_1(t)$? Sketch it.

Suppose a slice selection gradient is applied in z direction while $B_1(t)$ is applied and $G_z = 2$ G/cm.

(c) Where is this slice centered? What is the slice thickness?

(d) If you want to select another slice *adjacent* to the current slice (moving in the +z direction) with the same slice thickness, how should you change the RF pulse? Assume G_z stays the same.

(e) Right after the RF pulse in (d), what is the maximum phase difference of the spins inside the current slice? How does one rephase the spins?

(f) Suppose you want to sample k-space using a sampling trajectory that starts as shown in Figure P13.10. Assume that $|G_x| = 2.5$ G/cm and $|G_y| = 1.5$ G/cm. Sketch a pulse sequence to achieve this part of the trajectory. Indicate the acquired signal, the readout gradient, the phase-encoding gradient, and the slice selection gradient.

(g) For the sampling indicated in (e), if FOV_x must be 50 cm, what is the smallest sampling rate f_s?

13.29 Suppose we wish to image an object in all three dimensions, that is, we want our final image to be an $N \times N \times N$ array of voxels with isotropic dimensions. However, our ability to produce RF pulses is limited to short bursts of unmodulated RF; therefore, we cannot use slice selection.

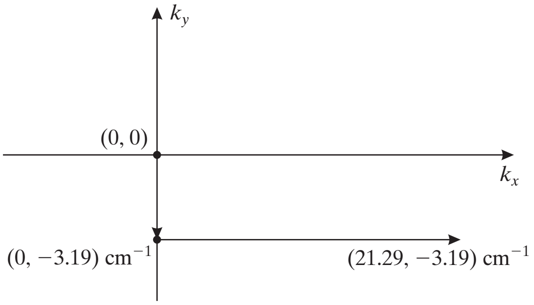


Figure P13.10
See Problem 13.28.

(a) Can you think of a method to the object? (Hint: look at the phase encoding section of the book and extend these ideas.)

(b) What would the equation for the signal be?

(c) How would you reconstruct the image?

13.30 Consider a generalized signal acquisition scenario for spatial encoding. Show that given N point objects with amplitudes A_j (where the A_j are real numbers), we can solve for their y coordinates y_j from the N equations derived from N separate signal acquisitions that have different phase-encoding gradient amplitudes G_y^m $(m = 0, 1, 2 \cdots N)$.

Index

A

A-mode, 11, 380–381
abdomen, 145
absorption
  acoustic, 344
absorption coefficient, 345
accuracy, 55, 88–92
  diagnostic, 89
  quantitative, 89
ACD, *See* annihilation coincidence detection
acoustic dipole, 356
acoustic energy flux, 341
acoustic gel, 371
acoustic impedance, 371
acoustic intensity, 341, 342
acoustic pressure, 348
acoustic wave, 336
  acoustic pressure, 337
  backward-traveling wave, 339
  characteristic impedance, 338
  coefficient of nonlinearity, 348
  forward-traveling wave, 339, 341
  frequency, 339, 350
  inward-traveling wave, 341
  longitudinal wave, 336
acoustic wave *(Continued)*
  mode conversion, 344
  nonlinear wave propagation, 347–349
  outward-traveling wave, 341
  particle displacement, 337
  particle velocity, 337
  plane wave, 338–340
  power, 342
  propagation, 367
  shear wave, 336
  shock wave, 349
  spherical wave, 340–341
acoustics, 335–366
active shimming, 442
ADC, *See* analog-to-digital converter
AEC, *See* x-ray tube, automatic exposure control
air
  as contrast agent, 143
air kerma, 128
airgap, 145
algebraic reconstruction technique, *See* reconstruction, iterative, algebraic reconstruction technique
aliasing, 81, 83
  anti-aliasing filter, 85, 475

aluminum, 144, 148, 149, 194
  density, 140
amino acid transport, 293
amorphous selenium, 152
amorphous silicon, 152
amplitude attenuation factor, 345, 346
analog-to-digital converter, 455, 475
Anger camera, *See* gamma camera
Anger, Hal, 6
angiography, 8
angular momentum, 413, 414
annihilation coincidence detection, 299, 300, 309
anode, 115
ART, *See* reconstruction, iterative, algebraic reconstruction technique
artifacts, 55, 86–87
  beam hardening artifact, 87
  motion artifact, 87
  ring artifact, 87
  star artifact, 87
atom
  atomic number, 107, 119, 147, 240
  binding energy, 119, 142, 240
    curve of, 241
  disintegration, 243
  effective atomic number, 119
  electron, 241
  electron cloud, 116
  excitation, 110, 114
  ground state, 108
  ion, 117
  ionization, 109, 115
  isobar, 240
  isomer, 240
  isotone, 240
  isotope, 240
  mass defect, 240, 241
  mass number, 107, 240
  neutron, 241, 245
  nucleon, 107, 240, 241
atom *(Continued)*
  nuclide, 107, 240
  planetary, 107
  proton, 241, 245
  radionuclide, 107, 240
  shell, 246
  unified atomic mass unit, 240
attenuation
  acoustic, 344–346
  differential, 118
  electromagnetic radiation, 120
  mass attenuation coefficient, 129
  tissue dependent, 120
attenuation coefficient, 345
attenuation correction, *See* reconstruction, attenuation correction
Auger electron, 117
avalanche photodiode, 304
axial, 4
azimuth, 393

B

B-mode, *See* ultrasound imaging, B-mode scan
background, 60
backprojection, 201–202, 207
  fan-beam, 211
  image, 201, 203
  summation, 202
bar phantom, 68
barium, 142, 143
beam hardening, 139, 214, 221
beam softening, 127
beamforming, 389
Becquerel, 243
BGO, *See* bismuth germanate
bias, 89
BIBO stability, 31
binding energy, 119, 142, 240, 241
  electron, 241
  nuclear, 242

biopsy needle, 334
bismuth germanate, 195, 301
bladder, 143, 235
Bloch equations, 410, 413, 425–426, 453
Bloch, Felix, 6
blood, 105
  flow, 293
  hemoglobin, 415, 485
  oxygenation level dependent, 485
Bohr, 107
BOLD, *See* blood oxygenation level dependent
Boltzmann's constant, 413, 479
bone, 222, 235
  cancellous, 105
  cortical, 105
  marrow, 102
  marrow edema, 408
  mineral density, 194
boron, 245
bowel, 143
Bq, *See* Becquerel
brain, 101, 222
breast, 154, 332
  cancer, 154
  compression, 154
breath-hold, 193
bremsstrahlung radiation, *See* radiation, bremsstrahlung
broad beam geometry, *See* x-ray, broad beam geometry
Bucky, Gustav, 144
bulk modulus, 336
Butterworth filter, 307

C

cadmium tungstate, 195
cadmium-zinc-telluride, 264, 297
calcium tungstate, 147, 166
cancer
  breast, 154
carbon, 245, 293, 415
Cassen, Benedict, 6
CBP, *See* reconstruction, convolution backprojection
center of mass, 265
central limit theorem, 72
cerebrospinal fluid, 429
cerium-doped lutetium oxyorthosilicate (LSO), 301
cesium iodide, 195, 264
characteristic radiation, 114, 139
chemical shift, 416, 435, 483
  shielding constant, 416
chest, 105, 145
circular polarization, 420
circular symmetry, 43–47
coincidence
  random, 301
coincidence detection, 294
  line, 293, 299
  scatter, 301
collimation
  grid, 144
  grid ratio, 144
  source, 188
collimator, 141, 256
  all-purpose, 296
  converging, 257
  diverging, 257
  effective hole length, 274
  efficiency, 276–278
  holes, 276
  line, 195
  parallel-hole, 256
  pinhole, 257
  sensitivity, 276–278
  septa, 275
  septum, 256
comb function, 19–20, 81
combined PET/CT, 303
combined PET/MRI, 303–304
combined SPECT/CT systems, 297–298
compensation filter, 141–142

compressibility, 336
Compton, 116-118, *See also* scattering, Compton
  edge, 262
  event, 119
  photon, 117
  plateau, 261, 302
  scattering, 116, 173–174, 258
  energy discrimination, 261
  scatter-to-primary ratio, 174
computed radiography, 149
computed tomography, 9–10, 101, 186–234
  aliasing, 221
  artifacts, 221–222
  beam hardening, 221–222
  collimation, 194, 196
  cone beam, 213
  CT number, 198
  data processing, 193
  dual source, 195
  dual-energy, 194–195, 222
  electron beam, 192
  fan-beam, 194
  fan-beam projections, 208
  filtration, 194
  gantry, 9
  generations, 188, 191–193, 199
  helical, 9, 187
    CT, 193, 196, 212–213, 303
  image quality, 213–222
  interpetrous lucency artifact, 222
  metal artifact, 222
  motion artifact, 222
  multiple-row detector (MDCT), 9, 187, 193, 196
  multislice, *See* computed tomography, multiple-row detector
  noise, 216–221
  patient table, 9
  point spread function, 215
  pulsed mode, 195
computed tomography *(Continued)*
  reference intensity, 197
  resolution, 214–216
  scan time, 192
  scanner, 105, 186
    cost, 193
    gantry, 196, 212
    good design, 219
    patient table, 196, 212
    slip ring, 196
  signal-to-noise ratio (SNR), 219–221
  slice thickness, 196
  slipring, 193
  spiral, *See* computed tomography, helical
  streak artifact, 222
computer vision, 194
computer-aided diagnosis, 154
contrast, 54–61, 135, 143, 174, 219
  local, 60
contrast agent, 105, 119, 142–143, 170
conventional radiography, *See* projection radiography
conversion efficiency, 147
convolution, 26
  associativity property, 28
  commutativity property, 28
  distributivity property, 28
  equation, 26, 42
  implementation, 43
  integral, 26
  theorem, 39
convolution backprojection, *See* reconstruction, convolution backprojection
coordinate system
  laboratory, 305
copper, 194
  characteristic photons, 140
Cormack, Allan, 5
coronal, 4
coronary angiography, 237
coronary artery, 237

critical angle, 343
Crooke's tube, 5, 106
cross section
  axial, 305, 447
  conventional radiographic view, 305
  coronal, 447
  double-oblique, 448
  oblique, 448
  sagittal, 447
cryostat, 441
CSF, *See* cerebrospinal fluid
CT number, *See* computed tomography, CT number
CT scanner, *See* computed tomography, scanner
cumulative distribution function, *See* random variable, distribution function
curie, 243
cutoff frequency, 43
cyclotron, 249
CZT, *See* cadmium-zinc-telluride

D

Damadian, Raymond, 6
dB, *See* decibels
de Hevesy, George, 6
decibels, 80
delta function, 17
delta ray, 114
density, 336
depth of penetration, 384
detective quantum efficiency, 171–173
detector
  area detector, 214
  area of, 311
  block, 300
  calibration, 197
  coincidence, 299
  coincident, 10
  collimation, 127, 190
  computed tomography, 195–196
  efficiency, 166, 171, 190, 219, 271, 278, 282

detector *(Continued)*
  energy-sensitive, 144
  movement, 297
  ring, 300
  solid-state, 195, 196
  spacing, 219
  width, 219
  x-ray, 8
  xenon gas, 195
DEXA, *See* dual-energy x-ray absorptiometry
diagnostic accuracy, 90
diagnostic radiography, 7
diaphragm, 103, 140
  Potter-Bucky, 145
diffusion, 486
diffusion tensor imaging, 487
  fractional anisotropy, 487
  mean diffusivity, 487
  principal eigenvector, 487
digital radiography, 8, 149–153
  computed radiography, 149
  photostimulable imaging plate, 149
Dirac function, 17
discrete random variable, 72–74
discretization, 80
disintegrations per second, 243
distortion, 55, 88
  magnification, 88
  shape distortion, 88
  size distortion, 88
distribution function, 70
Doppler
  effect, 349, 351
  frequency, 351
  motion monitor, 353
  pulse–echo mode, 352
  shift, 349, 351
  velocimeter, 353
dose, 113
  effective, 130
  x-ray, 140

dose equivalent, 129
dps, *See* disintegrations per second
DQE, *See* detective quantum efficiency
Driffield, Vero Charles, 168
dual-energy x-ray absorptiometry, 194
dynamic focusing, 390
dynode, 258

E

EBCT, *See* computed tomography, electron beam
ECG, *See* electrocardiogram
echo time, 427, 429
echocardiography, 331
ECT, *See* emission computed tomography
Edison, Thomas, 147
effective dose, 130
effective energy, 170, 197
efficiency, *See* quantum efficiency and detective quantum efficiency
electrocardiogram, 267, 269, 446
electromagnetic energy, 6
electromagnetic radiation, *See* radiation, electromagnetic
electron
  Auger, 117
  beam, 192
  binding energy, 109
  bound, 109
  collisional transfer, 114
  Compton
    kinetic energy, 118
  delta ray, 114
  density, 119
  energetic, 114
  free, 107, 109
  kinetic energy, 114
  radiative transfer, 114
  rest mass, 117
  shells, 108
electron capture, 245
electron volt, 7, 109, 241
electron-hole pairs, 264
elevation, 393
emission computed tomography, 238, 255, 294
emission imaging
  event, 256
emission tomography
  activity concentration, 314
  attenuation and scatter, 319–320
  attenuation correction, 307–309
  contrast, 320–321
  image quality, 317–321
  noise, 321
  signal-to-noise ratio, 321
  spatial resolution, 318–319
energy, 48
energy density
  acoustic, 341
  kinetic, 341
  potential, 341
energy fluence, 121
equivalent aluminum, 140
Ernst, Richard, 6
error function, 71
eV, *See* electron volt
excitation, 110
exponential signal, 22–23
exposure, 128, 171
extremities, 145

F

f-factor, 129
fan-beam, 191
  geometry, 208
  reconstruction formula, 211
Faraday induction, 417
  principle of reciprocity, 417
fast Fourier transform, 43, 459, 471
fat, 343, 344
FBP, *See* reconstruction, filtered backprojection

FDA, *See* Food and Drug Administration
FDG, *See* radiotracer, fluorodeoxyglucose (FDG)
fetal imaging, 331
FFT, *See* fast Fourier transform
fibular, 103
FID, *See* nuclear magnetic resonance, free induction decay
field of view (FOV), 188, 213, 444, 475
field pattern, 353, 372
  beamwidth, 354
  definition, 358
  diffraction formulation, 355–361
  far field, 354
  focusing, 372
  Fraunhofer approximation, 360–361, 374, 379–380
  Fraunhofer region, 354
  Fresnel approximation, 354, 358–360, 374, 379
  Fresnel region, 354
  geometric approximation, 354, 378–379
  geometric region, 354
  main lobe, 354
  paraxial approximation, 358
  plane wave approximation, 357–358
  received signal, 356–357
  side lobe, 354
filament, 136–139
film, 5, 105, 166–168, 263
  darkroom, 148
  gamma, 167, 168
  H&D curve, 168
  latent image, 167
  latitude, 167, 168
  optical density, 167
  optical opacity, 167
  optical transmissivity, 167
  radiographic, 147
  radiographic cassette, 147
    contaminant, 148
film *(Continued)*
  size, 147
  speed, 147, 168
  spoil exposure, 147
filter
  anti-aliasing, 85
  ideal low-pass, 43
filtered backprojection, 205, 472
filtration, 139
FLAIR, *See* fluid attenuated inversion recovery
flip angle, *See* tip angle
fluence rate, *See* x-ray, photon fluence rate
fluence, *See* x-ray, photon fluence
fluid attenuated inversion recovery, 433
fluorescence, 146
fluorescent, 147
fluorine, 250, 293
fluoroscopy, 5, 7, 136, 148
  dynamic, 5
  static, 5
fMRI, *See* functional magnetic resonance
focusing cup, 137
Food and Drug Administration, 444
foot, 103, 158
forced precession, 420, 449, 453
Fourier method, 204
Fourier resolution, 478, 479
Fourier scanning
  180° pulse, 465
  rectilinear, 477
Fourier space, 458
  aliasing, 477
  sampling, 471, 477
  scanning, 458–460
  trajectory, 458
Fourier transform, 31–36, 203, 457
  conjugate symmetry property, 37
  conjugation property, 37
  convolution property, 38
  fast, 43, 459, 471
  inverse, 32, 472
  linearity property, 36

Fourier transform *(Continued)*
  magnitude, 32
  Parseval's theorem, 40
  phase, 32
  product property, 39
  properties, 36
  rotation property, 38
  scaling property, 37
  separability property, 40
  slice, 203
  translation property, 37
FOV, *See* field of view
Fraunhofer approximation, *See* field pattern, Fraunhofer approximation
free fatty acid metabolism, 293
frequency, 32, 112
  spatial, 203
frequency encoding
  gradient, 460
frequency response, 41
Fresnel approximation, *See* field pattern, Fresnel approximation
full width at half maximum (FWHM), 62–63, 273, 275, 279, 478
function, 16
  antisymmetric, 37
  symmetric, 37
functional magnetic resonance imaging, 407
fundamental frequency, 22
FWHM, *See* full width at half maximum

G

gadolinium oxyorthosilicate (GSO), 301
gallium, 263
gamma camera (Anger camera), 6, 10, 255, 256, 293, 305
  crystal, 278
  detector system, 279
  discriminator circuit, 265
  efficiency, 258
gamma camera *(Continued)*
  energy resolution, 279–280
  event positioning, 264–266
  gating circuit, 262–263
  imaging equation, 269
  multiple, 295
  nonuniformity, 296
  positioning logic, 260
  pulse height analyzer, 260–262, 279
  pulse pileup, 282
  resolution, 258
  rotating, 294
  scintillator, 275
  sodium iodide crystal, 299
  uniformity, 278–279
  Z-pulse, 260, 265
gamma ray, 106, 124, 245, 246
Gaussian random variable, 71
GCV, *See* grid, conversion factor
glucose, 293
GM, *See* gray matter
gradient echoes, 461–462
gray, 128
gray matter (GM), 429
grid, 144–145
  conversion factor, 145
  crosshatch, 144
  focused, 144
  focused grid, 144
  frequency, 144
  stationary, 145
grid frequency, *See* grid, frequency
GSO, *See* gadolinium oxyorthosilicate
gyromagnetic ratio, 411

H

H&D curve, 168
half value layer (HVL), 124
half-life, 244, 293
Hamming window, 453
hand, 101

Hankel transform, 43–47, 215
  inverse, 478
heart, 102
  apex, 332
  chambers, 332
  horizontal long axis, 238
  mitral valve, 334
  short axis, 238
  vertical long axis, 238
helical computed tomography, 9, 187
helix
  pitch, 193, 212
hemoglobin, 415, 485
  deoxygenated, 485
  oxygenated, 485
Hounsfield unit, 187, 198
Hounsfield, Godfrey, 5
HU, *See* Hounsfield unit
Hurter, Ferdinand, 168
HVL, *See* x-ray, half value layer
hydrogen
  in fat, 416

I

image, 16
image contrast, 173
image quality, 54–99
image reconstruction, *See* reconstruction
imaging
  anatomical, 10
  computed tomography, 9
  cross sectional, 5
  dynamic, 5
  functional, 10
  magnetic resonance, 12
  nuclear medicine, 10, 144
  projection, 8
  projection radiography, 7
  radionuclide, 10
  reflection imaging, 7
imaging *(Continued)*
  static, 5
  ultrasound, 11
imprecision, 89
impulse, 16–19, 340
  function, 17
  response, 25
in-plane direction, 448
indicator function, 357
intensifying screen, 146–147, 166
  blurring, 165–166
  conversion efficiency, 147
  speed, 147
intensity, 121, 157
  reflectivity, 344
  transmittivity, 344
interpetrous lucency artifact, 222
invasive techniques, 1
inverse Radon transform, 186, 214
inverse square law, 312
iodine, 142, 249
ion, 107, 109
ionization, 107–110
  chamber, 127, 138
ionizing radiation, 101
  attenuation, 101
  bremsstrahlung, 137
  characteristic, 137
  electromagnetic
    interactions, 116
  particulate
    collisional transfer, 137
    radiative transfer, 137
  properties, 113–120
  shield, 118
  transmission, 101
irradiance, 167
isocenter, 208
isochromat, 416, 452, 455
isotropic, 58
iterative reconstruction, *See* reconstruction, iterative

J

Jacobian, 210
Johnson noise, 479

K

K-edge absorption, 142
k-space, *See* Fourier space
Kaczmarz, 314
kerma, 128
kidney, 143, 249, 332
Klein-Nishina formula, 119
knee, 102
kVp, *See* x-ray tube, peak tube voltage

L

laminogram, 202
lanthanum bromide, 302
Larmor frequency, 414–416, 419, 447, 452, 454, 457, 460, 475, 480, 486
  spatial dependence, 448, 449
laser scanner, 149
latent image, 146
Lauterbur, Paul, 6
lead, 145
lead zirconate titanate (PZT), 368
least squares, 315
LET, *See* radiation, linear energy transfer
light, 195
light box, 167
likelihood function, 316
line, 198
  angle, 198
  lateral position, 198
line impulse, 19, 61
line integral, 197, 198
line of stability, *See* nuclide, line of stability
line pairs per millimeter, 68
line spread function, 61–62
linear attenuation coefficient, 306
linear attenuation coefficient, *See* x-ray, linear attenuation coefficient
linear energy transfer (LET), 128
linear polarization, 420
linear tomography, *See* tomography, linear
linearity, 24
liquid helium, 441
liver, 340, 343, 344
local contrast, 60–61, 169
longitudinal magnetization, 416, 424, 432, 473
longitudinal relaxation time, 455
LOR, *See* positron emission tomography, line of response
Lorentz force, 443
lp, *See* line pairs per millimeter
LSF, *See* line spread function
LSI systems, *See* systems, linear shift-invariant
LSO, *See* cerium-doped lutetium oxyorthosilicate
luminescence, 146
lungs, 102, 143

M

M-mode, 11, 381
mA, *See* x-ray tube, tube current
magnesium oxide, 146
magnet
  active shielding, 442
  fringe field, 442
  passive shielding, 442
  passive shimming, 442
  shimming, 415, 442
  superconducting, 441
magnetic field
  external, 412
  fluctuations, 415
  macroscopic, 411
  microscopic, 411
  static, 412, 414

magnetic resonance imaging, 7, 12–13, 407, 410, 439–496
  T_1-weighted contrast, 430–432
  T_2-weighted contrast, 430
  antenna noise, 451
  artifacts, 482–483
  baseband signal, 456, 463, 466
  contrast agent, 407
  contrast mechanism, 472
  coordinate system, 447
  diffusion imaging, dMRI, 13
  diffusion-weighted imaging, 486–487
  echo time, 462
  echo-planar imaging, 12
  eddy current, 444
  field of view (FOV), 476
  field sensitivity, 445, 480
  Fourier resolution, 478, 479
  Fourier trajectory, 458
  frequency encoding, 448, 455–460
  functional imaging, fMRI, 12
  functional magnetic resonance, 485–486
  geometric distortion, 482
  ghosts, 482
  gradient amplitude, 443
  gradient strength, 450
  gradient-echo imaging, 456
  image quality, 475–483
  imaging equations, 472–475
  inversion pulse, 432
  inversion recovery, 432–433
  inversion time, 433
  noise, 479–481
  noise reduction, 480
  observed signal, 417
  peripheral nerve stimulation, 444
  phase accumulation, 462
  phase coherence, 461
  phase encoding, 448
  point spread function, 478
  polar data, 471–472
  polar sampling, 478

magnetic resonance imaging *(Continued)*
  polar scanning, 460
  practical RF waveform, 452–453
  proton density–weighted contrast, 429–430
  pulse repetition interval, 467
  pulse sequence, 407, 413
  readout direction, 457
  readout gradient, 457–458, 475
  received signal, 456
  receiver bandwidth, 475, 476
  refocusing gradient, 453–454
  refocusing lobe, 454
  resolution, 477–479
  RF center frequency, 450
  RF frequency range, 450
  ringing artifact, 483
  sampling, 475–477
  shading artifact, 480
  signal-to-noise ratio, 481–482
  slab, 449
  slice dephasing, 454
  slice position, 451
  slice profile, 453
  slice selection, 449–455
  slice thickness, 451
  spectroscopy, 12
  spin density, *See* spin density
  spin rephasing, 454
  spin-echo imaging, 456
  spoiled gradient echo (SPGR), 433, 467
  spoiler pulse, 467
  superconducting magnet, 13
  susceptibility-weighted imaging (SWI), 484–485
  tissue contrast, 429
  wraparound artifacts, 483
magnetic resonance scanner
  Aldermann-Grant coil, 446
  birdcage resonator, 445
  body coil, 445, 480
  console, 446
  field strength, 441

magnetic resonance scanner *(Continued)*
  gantry, 407
  gradient amplifier, 444
  gradient coil, 442–445
    shielding coil, 444
    slew rate, 444, 445
  head coil, 446, 480
  magnet, 441–442, 444
  phased array, 446
  pulse sequence, 408
  radio frequency coil, 445–446
  receiver coil, 407
  saddle coil, 444
  surface coil, 445, 480
  system components, 439–441
  volume coil, 445
magnetic susceptibility, 415
  diamagnetic, 415, 485
  ferromagnetic, 415
  paramagnetic, 415, 485
magnetization vector, 412, 448
  equation of motion, 413
  longitudinal component, 416
  transverse component, 416
magnification, 88, 160, 162, 165
main lobe, 354
mammography, 7, 136, 137, 144, 148, 154
  full field digital, 154
  paddle, 154
  stereo digital, 154
Mansfield, Peter, 6
marrow, 105
mAs, *See* x-ray tube, milliampere-second
mass
  relativistic, 110
mass attenuation coefficient, 129
matrix
  inverse, 315
maximum likelihood, 316
MDCT, *See* computed tomography, multiple-row detector
medical imaging, 1–4
  emission imaging, 7, 235, 239
  functional imaging, 235
  history, 5–6
  modality, 1, 7
  multimodal imaging, 13
  Nobel Prize, 5, 6
  physical signals, 6–7
  projection radiography, 7
  signals and systems, 2
  transmission imaging, 7
metastable, *See* nuclide, metastable
microcalcification, 154
mirror, 148
ML, *See* maximum likelihood
ML-EM, *See* reconstruction, iterative, maximum likelihood by expectation maximization
modulation, 56
modulation transfer function, 56–60, 166
  of system cascade, 66
molybdenum, 137, 251
motion tomography, 7
MTF, 57, 59
multiple sclerosis, 408
multislice computed tomography, *See* computed tomography, multiple-row detector (MDCT)

N

narrow beam geometry, *See* x-ray, narrow beam geometry
narrowband pulse, 355–356
  complex envelope, 355
  complex signal, 355
  envelope, 355
narrowband signal, 376
National Council on Radiation Protection and Measurement (NCRP), 140
NCRP, *See* National Council on Radiation Protection and Measurement

NEC, *See* positron emission tomography, noise equivalent count rate
negative predictive value, 91
neper, 345
neuron, 485
neuroradiology, 8
neuroreceptor density, 293
neutrino, 245
niobium-titanium wire, 441
nitrogen, 293
NMR, *See* nuclear magnetic resonance
noise, 55, 69–76, 169
  photons, 169
  power spectrum, 78
  thermal vibrations, 69
  white, 78
  wide-sense stationary, 78
noise equivalent count rate (NEC), 321
noninvasive techniques, 1
normal equation, 315
NPS, 78
nuclear magnetic resonance (NMR), 6, 7, 12, 407, 410–438
  T_2^*, 423
  basic signal, 418
  Bloch equations, 410, 413, 425–426, 453
  contrast mechanisms, 429–433
  echo, 423
  equilibrium, 424, 473
  Faraday induction, 7
  forced precession, 420, 449, 453
  free induction decay, 422
  inversion pulse, 421
  isochromat, 416, 452, 455
  longitudinal relaxation, 423
  longitudinal relaxation time (T_1), 424
  multiple echoes, 428
  nuclear magnetism, 411
  nuclear spin system, 411
  pi over two pulse, 421
  precession, 7, 12, 419
  pulse sequence, 427
nuclear magnetic resonance (NMR) *(Continued)*
  radio frequency excitation, 419–421
  refocusing, 423
  relaxation, 422–425
  sample volume, 418
  short TI inversion recovery (STIR), 433
  spin echoes, 426–428
  spin system, 410
  steady-state, 424
  transverse relaxation, 422, 426
  transverse relaxation time (T_2), 422, 455
nuclear medicine, 10, 235–329
  biodistribution, 235
  dynamic studies, 268
  hot spot, 235
  rectilinear scanner, 6
nuclear spin system, 7
nucleon
  binding energy, 241
nucleus
  angular momentum, 410
  charge, 410
  spin, 410
  spin quantum number, 412
  strong force, 242
nuclide, 107, 240
  line of stability, 242
  metastable, 246
  radioactive, 108, 242
  stable, 242
Nyquist rate, 268
Nyquist sampling theorem, 84, 268

O

object
  sinusoidal, 56
oblique, 4
obliquity, 157
obstetrics, 332
optical density, 167
optical transfer function, 41

organ boundaries, 335
orthopedic surgery, 105
OSEM, *See* reconstruction, iterative, ordered subsets expectation maximization
OTF, *See* optical transfer function
oxygen, 250, 293, 415, 485

P

pair production, 116
parallel ray geometry, 188, 198, 199
Parseval's theorem, 40
partial differential equation, 338
partial volume effect, 319
particulate radiation, 106, 110–112, 245
pdf, *See* random variable, density function
PDF, *See* random variable, distribution function
pelvis, 103
perfusion imaging
  myocardial, 237
permeability constant, 480
Perrier, Carlo, 6
PET, *See* positron emission tomography
phase encoding, 462–465
phosphor, 146, 148
phosphorescence, 147
phosphorescent, 147
photocathode, 258
photoelectric effect, 116–117, 123, 166, 195, 258, 319
photoelectron, 117
photomultiplier tube, 149, 195, 256, 258, 259, 275, 279, 300
  anode, 259
  dynode, 259, 280
  output pulse, 260
  photocathode, 258
  window, 258
  Z-pulse, 261, 262
photon, 112, 169, 172
  absorption, 118
  annihilation, 246, 293
photon *(Continued)*
  attenuation, 124
  Compton, 117
  energy, 112, 119, 170
  fluence, 121
  light, 147
  monoenergetic, 123
  quanta, 169
  scintillation, 260
  x-ray, 147
photon fluence, 271
photon fluence rate, 270, 271, 306, 310
photons, 69
photopeak, 261, 279, 301
piezoelectric crystal, 367
piezoelectric effect, 368
pitch, *See* helix, pitch
pixel, 16, 267
pixel size, 478
planar scintigraphy, 238, 251, 255–292
  acquisition modes, 266–269
  activity concentration, 272
  collimator, 270
  collimator resolution, 273–275
  Compton scattering, 118
  contrast, 281
  dynamic frame mode, 268
  effective resolution, 276
  factors affecting count rate, 281–282
  frame mode, 280
  image contrast, 279
  image formation, 264–272
  image quality, 272–282
  imaging equation, 269–272
  intrinsic resolution, 275–276
  list mode acquisition, 266
  mean pixel count, 272
  multiple-gated acquisition, 268, 269
  noise, 280–281
  patient motion, 281
  planar source, 271–272
  point spread function, 274

planar scintigraphy *(Continued)*
  resolution, 273–276
  sensitivity, 276–278
  signal-to-noise ratio, 280
  solid state camera, 264
  spatial distortion, 279
  static frame mode, 267
  total counts, 268
  uniform flood, 278
  whole-body acquisition, 268
Planck's constant, 112, 117
plane waves, 338
plastic, 144, 146
PMF, *See* random variable, probability mass function
point, 16
point impulse, 17, 23, 24, 199
  scaling property, 19, 20
  sifting property, 18, 199, 203, 272
point spread function (PSF), 25, 275
  absolutely integrable, 31
  circularly symmetric, 56, 478
  Gaussian, 31
  isotropic, 61
Poisson counting process, 248
Poisson random variable, *See* random variable, Poisson
positive predictive value, 91
positron, 245
  annihilation, 298
positron decay, 245
positron emission tomography (PET), 10, 238, 255, 298–304, 309–312
  3D mode, 301
  attenuation correction, 313
  coincidence line, 299
  detector ring, 300
  energy discrimination, 301
  glucose metabolism, 238
  image formation, 309–313
  image reconstruction, 312
  imaging equation, 310–312
positron emission tomography (PET) *(Continued)*
  instrumentation, 298–304
  line of response (LOR), 309–310
  noise equivalent count rate, 321
  random coincidence, 320
  septa, 301
  time window, 299, 320
  time-of-flight, 302
positron range, 318
positron range function, 318
Potter-Bucky diaphragm, 145
power, 48
precession, 414
pressure
  reflectivity, 344
  transmittivity, 344
prevalence, 91
probability, 70
  Compton event, 119
  photoelectric event, 119
probability density function, 70
probe, *See* ultrasound probe
projection, 5, 9, 105, 135, 186, 197, 199, 200
  cone beam, 193
  fan-beam, 191, 193, 208
  parallel-ray, 188, 198, 199
  redundant, 201
  truncated, 213
projection image, 3
projection radiography, 7–9, 101, 135–185
  beam divergence, 157–158
  chest x-ray, 135
  density maintenance formula, 156
  depth-dependent blurring, 160
  depth-dependent magnification, 160–161
  exposure time, 135
  extended source, 162–165
  film, 10
  geometric effects, 155–162
  imaging equation, 162, 164
  inverse square law, 155–157, 165

projection radiography *(Continued)*
  obliquity, 157
  path length, 158–160
  patient positioning, 158
  radiation exposure, 135
  resolution, 163
  shading artifact, 159
  source magnification, 164
projection-slice theorem, 203–204, 215, 472
proton density, 413, 455
PSF, *See* point spread function
pulse height, 259
pulse height analyzer, 260–262, 279
  acceptance window, 262
  offset window, 263
pulse repetition interval, 429
pulse sequence
  2-D gradient echo, 463
  bipolar pulse, 486
  fast imaging sequence, 467
  FLAIR sequence, 433
  gradient echo, 485
  low-pass filtering process, 477
  slice selection, 454–455
  slow imaging sequence, 467
  spin echo, 487
  spoiled gradient echo (SPGR), 467
  steady state, 473
  STIR sequence, 433
Purcell, Edward, 6
PVDF, *See* ultrasound transducer, polyvinylidene fluoride
PZT, *See* lead zirconate titanate

Q

quality factor, 129
quanta, 69
quantum efficiency, 171–172
quantum mottle, 69, 169
quantum physics, 412
quartz, 369

R

rad, 128
radiation, 107
  absorbed dose, 128
  bremsstrahlung, 114, 115, 122
  characteristic, 114, 139
  detection
    scintillation, 258
  dose, 128, 170
  dose equivalent, 129
  dosimetry, 127–131
  effective dose, 130
  electromagnetic, 112–113, 245, 319
  exposure, 128, 171
  f-factor, 129
  infrared, 114
  inverse square law, 128, 270, 271, 306
  ionizing, 106, 107, 331
  linear energy transfer (LET), 128
  particulate, 106, 110–112, 245
  positron, 298
  quality factor, 129
  specific ionization, 128
  units, 127
radiative transfer, 114
radio frequency coil, 417
radio frequency field, 446
radio wave, 11, 417
radioactive decay, 239, 245
radioactive decay law, 243–245
radioactivity, 243, 244
  alpha decay, 245
  beta decay, 245
  curie, 243
  daughter atom, 242
  decay, 242
  decay constant, 244, 270
  decay factor, 244
  decay law, 243–245, 247
  disintegrations per second, 243
  effective half-life, 249

radioactivity *(Continued)*
  half-life, 244, 293
  isomeric transition, 245
  parent atom, 242
  statistics of decay, 247
radiogenic carcinogenesis, 130
radiograph, 135
  planar, 5
radiographic cassette, 147–148
radiography
  conventional, 135
  planar, 5
radionuclide, 107, 240, 242, 244
radiopharmaceutical, *See* radiotracer
radiotracer, 6, 7, 10, 235, 239, 249–251, 293
  activity concentration, 270
  biodistribution, 269
  cyclotron, 249
  energy, 249
  fluorodeoxyglucose (FDG), 238, 250
  gamma emitter, 255
  half-life, 249
  iodine, 10
  low-energy, 271
  monoenergetic, 250
  positron emitter, 255
  redistribution, 268
  safe, 249
  uptake, 268
  washout, 268
radium, 243
radon, 243
Radon transform, 186, 199, 312, 472
  convolution property, 214
  inverse, 186, 214
  sinogram, 200–201, 309
ramp filter, *See* reconstruction, ramp filter
random coincidence, 320
random process
  Poisson counting process, 248
random variable, 70–76
  continuous, 70
  density function, 70
  discrete, 72–74
  distribution function, 70
  expected value, 70, 73
  Gaussian, 71
  independent, 75, 217
  mean, 70, 217, 247
  mean value, 73
  Poisson, 73, 170, 216, 217, 247, 280, 316
  probability mass function, 73, 247
  random deletions, 216
  Rayleigh, 379
  standard deviation, 71, 169
  transformation, 217
  uniform, 71
  variance, 70, 73, 170, 217
range, 393
range equation, 378
rayl, 338
Rayleigh, John, 6
RBW, *See* magnetic resonance imaging, receiver bandwidth
receiver coil
  electrical resistance, 479
reconstruction, 3, 105, 206–207, 293, 446
  attenuation correction, 294
  cone beam, 213
  convolution backprojection, 206, 217, 312, 313, 459, 472
  fan-beam, 208–212
  Feldkamp algorithm, 213
  filtered backprojection, 205, 313
  forward problem, 314
  forward projection process, 315
  Fourier method, 204–205
  Fourier rebinning, 309, 312
  iterative, 213, 313–317
    algebraic reconstruction technique, 314–316
    convergence, 315

reconstruction *(Continued)*
        forward model, 317
        initial guess, 314
        maximum likelihood by expectation maximization, 316–317
        ordered subsets expectation maximization (OSEM), 317
        sensitivity factor, 317
        stopping rule, 314
    parallel ray, 198–207
    ramp filter, 205, 206, 214, 217, 307, 313, 318
    time-of-flight constraint, 302
    window, 219
    windowing, 206
rect function, 20–22, 38, 161, 169, 196, 318
reflection coefficient, 347
reflectivity, 392
reformat, 105
relativity
    theory of, 110
relaxation, 422–425
    matrix, 425
rem, 129
repetition time, 381
resistivity, 480
resolution, 55, 61–69
    energy, 262
    modulation transfer function, 63–65
    spatial, 61
    spectral, 61, 69
    temporal, 61, 69
    test pattern, 68
    tool, 68
resolution cell, 378
resolution tool, 68
resolving time, *See* scintillator, dead time
resonator, *See* magnetic resonance scanner, radio frequency coil
restriction, 139–141
RF pulse, 419
ribs, 103
roentgen, 128
Roentgen, Wilhelm, 5, 106, 146, 147
Rose model, 80
rotating frame, 419

S

sagittal, 4
sampling, 16, 80–86
    aliasing, 81, 83, 212
    coarse sampling, 80
    fine sampling, 80
    function, 19–20, 81
    Nyquist sampling period, 84
    rate, 475
    rectangular, 80
    sampling frequency, 80
    sampling period, 80
    sampling theorem, 84
scanning slit, 145–146
scatter rejection, 145
scatter-to-primary ratio, 174
scatterer, 357, 367, 378
scattering
    acoustic, 344, 347
    angle, 117, 118
    backscatter, 118
    Compton, 116–118, 140, 144, 145, 188, 192, 196, 275, 278, 319
    reduce effects of, 144
scintigraphy, 10
scintillation, 239, 258
scintillation crystal
    cut, 300
scintillator, 195
    cesium iodide, 152
    dead time, 282, 320
    decay time, 301
    sodium iodide, 256
SCR, *See* silicon-controlled rectifier
Segre, Emilo, 6

selenium
  amorphous, 152
sensitivity, 89
separable signals, 23
separable systems, 30–31
septum, 256, 275
shadow, 135
shield, 118
shift invariance, 25–27
shoulder, 101
SI, *See* radiation, specific ionization
side lobe, 354
sifting property, 18, 199, 203, 272
signal, 15, 16, 76
  band-limited, 83
  bounded, 31
  circularly symmetric, 44
  comb and sampling functions, 19–20
  complex exponential, 22
  continuous, 15, 16
  discrete, 15
  energy, 48
  exponential and sinusoidal signals, 22–23
  input, 24
  line impulse, 19
  mixed, 15
  output, 24
  period, 24
  periodic, 23–24
  point impulse, 16–19, 340
  power, 48
  rect function, 20–22, 38, 161, 169, 196, 318
  representation, 16
  sampling function, 20
  sawtooth, 349
  separable, 23
  sinc function, 20–22, 215
  sinusoid, 349, 350, 418
  sinusoidal, 22, 339
  smoothing, 43
  total power, 48
  visualization, 16
signal-to-noise ratio (SNR), 76–80, 169, 174, 214, 219
  amplitude, 77
  background, 169
  differential, 79
  frequency-dependent power, 79
  power, 77, 78
  target, 169
silicon
  amorphous, 152
silicon-controlled rectifier, 138
similar triangle, 160
sinc function, 20–22, 215
single photon emission computed tomography (SPECT), 10, 255, 294–298, 304–309
  bone scan, 295
  cardiac, 295
  dual head system, 295
  instrumentation, 294–296
  multihead system, 296
  myocardial perfusion, 237
  patient table, 10
  sensitivity, 295
  specialized geometry systems, 297
single photon emission tomography, 251
sinogram, 200–201, 206, 207, 309
sinusoidal signal, 22–23
slab
  thin, 118, 123, 126
Snell's law, 342
SNR, *See* signal-to-noise ratio
sodium iodide, 124, 195, 256, 258, 264
soft tissue, 101
solenoid, 480
solid-state photodiode, 195
sound, 335
  propagation, 335
sound wave, *See* acoustic wave
source magnification, 164
specific ionization, 128
specificity, 89
speckle, 378, 392

SPECT, *See* single photon emission computed tomography
spectral line, 122
spectral resolution, 61, 69
spectrum, 32
  bremsstrahlung, 115
  line, 122
  magnitude, 33
  phase, 33
  power, 33
  x-ray, 122
speed of sound, 336, 348
SPGR (spoiled gradient echo), 433, 467
spin density
  effective, 456, 473
spin echo, 426, 465–466
spin-lattice relaxation, *See* nuclear magnetic resonance, longitudinal relaxation
spin-spin relaxation, *See* nuclear magnetic resonance, transverse relaxation
spiral computed tomography, *See* computed tomography, helical
spot size, 162, 164
stability, 31
standard deviation, 71, 169
strain, 367
streak artifacts, 222
STIR (short TI inversion recovery), 433
superposition, 165
superposition integral, 25
surgery, 1
SWI, *See* magnetic resonance imaging, susceptibility-weighted imaging
system
  subsystem cascade, 65–68, 318
systems, 15, 24
  BIBO stable, 31
  cascade connection of, 28, 65
  connection of, 28
  continuous, 24
  continuous-to-continuous, 16, 24
  continuous-to-discrete, 16
systems *(Continued)*
  frequency response, 41
  linear, 24–25
  linear shift-invariant (LSI), 26
  optical transfer function, 41
  parallel connection of, 28
  point spread function, 25
  separable, 30–31
  serial connection of, 28
  shift invariant, 25–27
  stable, 31
  superposition integral, 25
  transfer function, 41

T

target, 60
technetium, 6, 124, 240, 243, 251, 261, 278
temperature, 413, 479
temporal resolution, 61, 69
TGC, *See* time-gain compensation
thallium, 251, 278
through-plane direction, 448
thyroid, 142, 249
tibia, 103
time-gain compensation (TGC), 375
time-of-flight PET (TOF-PET), 302
tip angle, 418, 421, 429, 452, 472
  small tip angle approximation, 453
titanium, 148
titanium dioxide, 146
TOF-PET, *See* time-of-flight PET
tomogram, 186
tomographic image, 4
tomography, 9
  linear, 154
tomosynthesis, 154
torque, 414
total power, 48
transducer, 11, 335
  beam divergence, 362
  field pattern, 353–362

transducer *(Continued)*
  focusing, 361–362
    electronic, 361
    lens, 361
transducer beam
  steering, 386
transducer, *See* ultrasound transducer
transfer function, 41
transistor
  thin film, 152
transmittivity, 162, 165
transverse magnetization, 416–418, 446, 447, 453, 455, 484, 486
  dephasing, 422
  phase, 417
  spoiling, 467
transverse relaxation time, 422, 455
true-negative fraction, 89
true-positive fraction, 89
tube current, 136, 138
tube voltage, 137–139
tungsten, 137

U

ultrasound, 335
  frequency, 335
  fundamental frequency, 349
  harmonic frequencies, 348, 349
  propagation, 335
  wave equation, 336–341
ultrasound beam
  steered, 372
ultrasound imaging, 11–12, 331, 367–405
  A-mode scan, 11, 380–381
  A-mode signal, 377
  B-mode scan, 11, 381–385
  basic pulse-echo signal equation, 358
  compound B-mode scan, 382
  compound scanning, 395–396
  depth of penetration, 384
  displaced aperture, 396

ultrasound imaging *(Continued)*
  Doppler effect, 334
  Doppler imaging, 11, 349–353
  dynamic focusing, 389
  echo, 11, 335
  embryo, 12
  envelope detection, 376
  frame rate, 385
  grating lobe, 383
  harmonic, 372
  imaging modes, 380–385
  instrumentation, 367–374
  M-mode scan, 11, 381
  pulse repetition rate, 384–385
  pulse-echo, 374–377
  pulse-echo equation, 374–377
  pulse-echo mode, 367
  range equation, 378
  real-time, 334
  reflection, 331
  resolution cell, 378
  scan conversion, 383
  slice thickness, 393–394
  speckle, 12, 332
  steering and focusing, 386–392
  system, 12, 331, 392
  three-dimensional, 391–392
  transducer, 12
  transducer motion, 377–380
ultrasound probe, 370, 372–374
  1.5D array, 393
  beamforming, 389
  delay element, 386
  dynamic focusing, 390
  electronic scanner, 373–374
  focusing, 387
  linear array, 373
  linear scanner, 383
  matching layer, 371
  mechanical scanner, 373
  mechanical sector scanner, 383
  phased array, 373, 387

ultrasound probe *(Continued)*
  phased array sector scanner, 383
  rocker, 373
  steering, 387
  wobbler, 373
ultrasound transducer, 367–372
  bandwidth, 372
  broadband, 371–372
  damping, 371
  damping and matching, 370–371
  polyvinylidene fluoride (PVDF), 369
  quartz, 369
  receiving constant, 369
  resonance, 369–370
  shape, 369
  shock excitation, 370
  transmitting constant, 369
uniform random variable, 71
uniformity, 278–279
unit disk, 45

V

vacuum, 148
vein, 235
voxel
  size, 419

W

wave equation
  plane wave, 338
  propagation, 341
  spherical, 340
  three-dimensional, 338
wave number, 339, 458
wavelength, 112, 339
white matter (WM), 408, 429
whole body scan, 9
WM, *See* white matter

X

x-ray
  added filtering, 140
  attenuation, 123, 142, 147, 188
    integral form, 125
  beam, 136
  beam hardening, 139, 214, 221
  beam softening, 127
  beam strength, 121
  broad beam geometry, 127
  characteristic, 115
  chest, 101, 186
  diverging beam, 155
  dose, 127, 170
  effective energy, 123, 139, 216, 222
  energy fluence, 121
  energy fluence rate, 280
  equivalent monoenergetic beam, 122
  exposure, 167
  filtration, 158
  flux, 157
  half value layer (HVL), 124
  image intensifier, 148–149
    anode, 149
    dynode, 149
    photocathode, 148
  intensity, 121, 157
  linear attenuation coefficient, 124, 147, 154, 194, 197, 274, 308
    energy dependence, 125
  monoenergetic, 197
  narrow beam geometry, 123, 125
  penetration, 136
  photon fluence, 121
  photon fluence rate, 121, 122
  polyenergetic, 122, 139, 170, 197, 222
  spectrum, 139, 155, 170, 194, 197, 222

x-ray imaging
  cone beam, 9
x-ray systems
  mobile, 8
x-ray tube, 8, 104, 115, 136, 140, 192, 194
  added filtering, 139
  anode, 137, 139, 140, 158, 170, 192
    rotating, 138
  anode heel effect, 158
  automatic exposure control, 138
  cathode, 136, 158
  dielectric oil, 139
  exposure, 138
  filament, 136–139
  filament current, 148, 170
  filtration, 139
  focusing cup, 137
  glass housing, 139
  heat, 138
  housing, 140
x-ray tube *(Continued)*
  inherent filtering, 139
  milliampere-second, 138
  peak tube voltage, 137, 170
  restriction, 139–141
  spot size, 162, 164
  thermionic emission, 136
  tube current, 136, 138
  tube voltage, 137–139
x-rays, 5, 106, 136
XRII, *See* x-ray, image intensifier

Y

yttrium, 195

Z

Z, *See* atom, atomic number