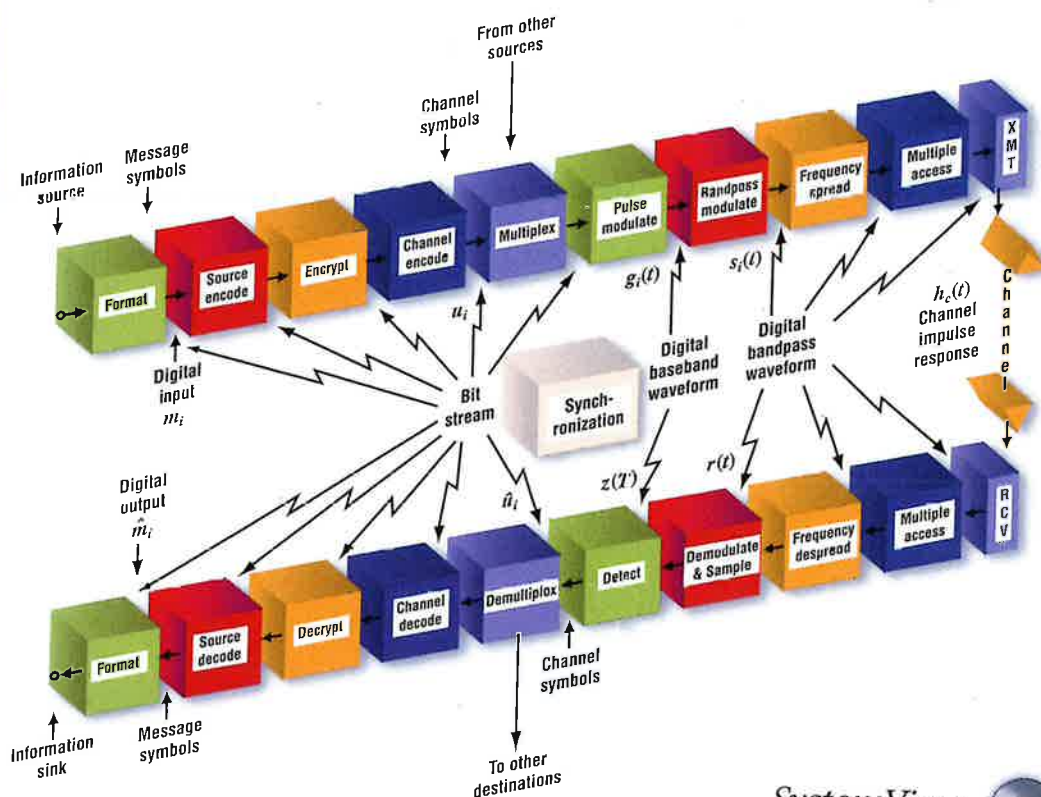


Digital Communications

SECOND EDITION

Fundamentals and Applications



SystemView
BY ELANIX

BERNARD SKLAR

DIGITAL COMMUNICATIONS

Fundamentals and Applications

Second Edition

BERNARD SKLAR

*Communications Engineering Services, Tarzana, California
and
University of California, Los Angeles*



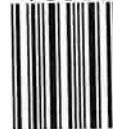
Prentice Hall P T R
Upper Saddle River, New Jersey 07458
www.phptr.com

ISBN 0-13-084788-7



9 780130 847881

90000



Library of Congress Cataloging-in-Publication Data

SKLAR, BERNARD

Digital communications, 2/e/Bernard Sklar—2nd ed.

Includes index and bibliographical references.

ISBN: 0-13-084788-7

1. Digital communications. I. Title.

TK5103.7.855 2000

00-7476

621.38'0413

Editorial/Production Supervision: *Rose Kernan*

Publisher: *Bernard Goodwin*

Marketing Manager: *Bryan Gambrel*

Manufacturing Manager: *Alexis Heydt*

Cover Design Director: *Jerry Votta*

Cover Designer: *Nina Scuderi*

Interior Formatting: *Pine Tree Composition*



© 2001 Prentice Hall PTR

Prentice-Hall, Inc.

Upper Saddle River, New Jersey 07458

Prentice Hall books are widely used by corporations and government agencies for training, marketing, and resale.

The publisher offers discounts on this book when ordered in bulk quantities.

For more information, contact:

Corporate Sales Department

Prentice Hall PTR

1 Lake Street

Upper Saddle River, NJ 07458

Phone: 800-382-3419; FAX: 201-236-7141

E-mail: corpsales@prenhall.com

All products or services mentioned in this book are the trademarks or service marks of their respective companies or organizations. Screen shots reprinted by permission from Microsoft Corporation.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 Tenth Printing

Reprinted with corrections September, 2004.

ISBN 0-13-084788-7

Prentice-Hall International (UK) Limited, *London*

Prentice-Hall of Australia Pty. Limited, *Sydney*

Prentice-Hall Canada Inc., *Toronto*

Prentice-Hall Hispanoamericana, S.A., *Mexico*

Prentice-Hall of India Private Limited, *New Delhi*

Prentice-Hall of Japan, Inc., *Tokyo*

Pearson Education Asia Pte. Ltd.

Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

*To my wife Gwen, and our children,
Debra, Sharon, and Dean, and to
the memory of my mother and father,
Ruth and Julius Sklar*

respec-
oration.

without

Contents

PREFACE

xvii

1 SIGNALS AND SPECTRA

1

- 1.1 Digital Communication Signal Processing, 3
 - 1.1.1 *Why Digital?*, 3
 - 1.1.2 *Typical Block Diagram and Transformations*, 4
 - 1.1.3 *Basic Digital Communication Nomenclature*, 11
 - 1.1.4 *Digital versus Analog Performance Criteria*, 13
- 1.2 Classification of Signals, 14
 - 1.2.1 *Deterministic and Random Signals*, 14
 - 1.2.2 *Periodic and Nonperiodic Signals*, 14
 - 1.2.3 *Analog and Discrete Signals*, 14
 - 1.2.4 *Energy and Power Signals*, 14
 - 1.2.5 *The Unit Impulse Function*, 16
- 1.3 Spectral Density, 16
 - 1.3.1 *Energy Spectral Density*, 17
 - 1.3.2 *Power Spectral Density*, 17
- 1.4 Autocorrelation, 19
 - 1.4.1 *Autocorrelation of an Energy Signal*, 19
 - 1.4.2 *Autocorrelation of a Periodic (Power) Signal*, 20
- 1.5 Random Signals, 20
 - 1.5.1 *Random Variables*, 20
 - 1.5.2 *Random Processes*, 22
 - 1.5.3 *Time Averaging and Ergodicity*, 25
 - 1.5.4 *Power Spectral Density of a Random Process*, 26
 - 1.5.5 *Noise in Communication Systems*, 30

v

- 1.6 Signal Transmission through Linear Systems, 33
 - 1.6.1 *Impulse Response*, 34
 - 1.6.2 *Frequency Transfer Function*, 35
 - 1.6.3 *Distortionless Transmission*, 36
 - 1.6.4 *Signals, Circuits, and Spectra*, 42
- 1.7 Bandwidth of Digital Data, 45
 - 1.7.1 *Baseband versus Bandpass*, 45
 - 1.7.2 *The Bandwidth Dilemma*, 47
- 1.8 Conclusion, 51

2 FORMATTING AND BASEBAND MODULATION

55

- 2.1 Baseband Systems, 56
- 2.2 Formatting Textual Data (Character Coding), 58
- 2.3 Messages, Characters, and Symbols, 61
 - 2.3.1 *Example of Messages, Characters, and Symbols*, 61
- 2.4 Formatting Analog Information, 62
 - 2.4.1 *The Sampling Theorem*, 63
 - 2.4.2 *Aliasing*, 69
 - 2.4.3 *Why Oversample?* 72
 - 2.4.4 *Signal Interface for a Digital System*, 75
- 2.5 Sources of Corruption, 76
 - 2.5.1 *Sampling and Quantizing Effects*, 76
 - 2.5.2 *Channel Effects*, 77
 - 2.5.3 *Signal-to-Noise Ratio for Quantized Pulses*, 78
- 2.6 Pulse Code Modulation, 79
- 2.7 Uniform and Nonuniform Quantization, 81
 - 2.7.1 *Statistics of Speech Amplitudes*, 81
 - 2.7.2 *Nonuniform Quantization*, 83
 - 2.7.3 *Companding Characteristics*, 84
- 2.8 Baseband Modulation, 85
 - 2.8.1 *Waveform Representation of Binary Digits*, 85
 - 2.8.2 *PCM Waveform Types*, 85
 - 2.8.3 *Spectral Attributes of PCM Waveforms*, 89
 - 2.8.4 *Bits per PCM Word and Bits per Symbol*, 90
 - 2.8.5 *M-ary Pulse Modulation Waveforms*, 91
- 2.9 Correlative Coding, 94
 - 2.9.1 *Duobinary Signaling*, 94
 - 2.9.2 *Duobinary Decoding*, 95
 - 2.9.3 *Precoding*, 96
 - 2.9.4 *Duobinary Equivalent Transfer Function*, 97
 - 2.9.5 *Comparison of Binary with Duobinary Signaling*, 98
 - 2.9.6 *Polybinary Signaling*, 99
- 2.10 Conclusion, 100

3 BASEBAND DEMODULATION/DETECTION

104

- 3.1 Signals and Noise, 106
 - 3.1.1 *Error-Performance Degradation in Communication Systems*, 106
 - 3.1.2 *Demodulation and Detection*, 107
 - 3.1.3 *A Vectorial View of Signals and Noise*, 110
 - 3.1.4 *The Basic SNR Parameter for Digital Communication Systems*, 117
 - 3.1.5 *Why E_b/N_0 Is a Natural Figure of Merit*, 118
- 3.2 Detection of Binary Signals in Gaussian Noise, 119
 - 3.2.1 *Maximum Likelihood Receiver Structure*, 119
 - 3.2.2 *The Matched Filter*, 122
 - 3.2.3 *Correlation Realization of the Matched Filter*, 124
 - 3.2.4 *Optimizing Error Performance*, 127
 - 3.2.5 *Error Probability Performance of Binary Signaling*, 131
- 3.3 Intersymbol Interference, 136
 - 3.3.1 *Pulse Shaping to Reduce ISI*, 138
 - 3.3.2 *Two Types of Error-Performance Degradation*, 142
 - 3.3.3 *Demodulation/Detection of Shaped Pulses*, 145
- 3.4 Equalization, 149
 - 3.4.1 *Channel Characterization*, 149
 - 3.4.2 *Eye Pattern*, 151
 - 3.4.3 *Equalizer Filter Types*, 152
 - 3.4.4 *Preset and Adaptive Equalization*, 158
 - 3.4.5 *Filter Update Rate*, 160
- 3.5 Conclusion, 161

4 BANDPASS MODULATION AND DEMODULATION/DETECTION

167

- 4.1 Why Modulate? 168
- 4.2 Digital Bandpass Modulation Techniques, 169
 - 4.2.1 *Phasor Representation of a Sinusoid*, 171
 - 4.2.2 *Phase Shift Keying*, 173
 - 4.2.3 *Frequency Shift Keying*, 175
 - 4.2.4 *Amplitude Shift Keying*, 175
 - 4.2.5 *Amplitude Phase Keying*, 176
 - 4.2.6 *Waveform Amplitude Coefficient*, 176
- 4.3 Detection of Signals in Gaussian Noise, 177
 - 4.3.1 *Decision Regions*, 177
 - 4.3.2 *Correlation Receiver*, 178
- 4.4 Coherent Detection, 183
 - 4.4.1 *Coherent Detection of PSK*, 183
 - 4.4.2 *Sampled Matched Filter*, 184
 - 4.4.3 *Coherent Detection of Multiple Phase Shift Keying*, 188
 - 4.4.4 *Coherent Detection of FSK*, 191

- 4.5 Noncoherent Detection, 194
 - 4.5.1 *Detection of Differential PSK*, 194
 - 4.5.2 *Binary Differential PSK Example*, 196
 - 4.5.3 *Noncoherent Detection of FSK*, 198
 - 4.5.4 *Required Tone Spacing for Noncoherent Orthogonal FSK*, 200
- 4.6 Complex Envelope, 204
 - 4.6.1 *Quadrature Implementation of a Modulator*, 205
 - 4.6.2 *D8PSK Modulator Example*, 206
 - 4.6.3 *D8PSK Demodulator Example*, 208
- 4.7 Error Performance for Binary Systems, 209
 - 4.7.1 *Probability of Bit Error for Coherently Detected BPSK*, 209
 - 4.7.2 *Probability of Bit Error for Coherently Detected Differentially Encoded Binary PSK*, 211
 - 4.7.3 *Probability of Bit Error for Coherently Detected Binary Orthogonal FSK*, 213
 - 4.7.4 *Probability of Bit Error for Noncoherently Detected Binary Orthogonal FSK*, 213
 - 4.7.5 *Probability of Bit Error for Binary DPSK*, 216
 - 4.7.6 *Comparison of Bit Error Performance for Various Modulation Types*, 218
- 4.8 M-ary Signaling and Performance, 219
 - 4.8.1 *Ideal Probability of Bit Error Performance*, 219
 - 4.8.2 *M-ary Signaling*, 220
 - 4.8.3 *Vectorial View of MPSK Signaling*, 222
 - 4.8.4 *BPSK and QPSK Have the Same Bit Error Probability*, 223
 - 4.8.5 *Vectorial View of MFSK Signaling*, 225
- 4.9 Symbol Error Performance for M-ary Systems ($M > 2$), 229
 - 4.9.1 *Probability of Symbol Error for MPSK*, 229
 - 4.9.2 *Probability of Symbol Error for MFSK*, 230
 - 4.9.3 *Bit Error Probability versus Symbol Error Probability for Orthogonal Signals*, 232
 - 4.9.4 *Bit Error Probability versus Symbol Error Probability for Multiple Phase Signaling*, 234
 - 4.9.5 *Effects of Intersymbol Interference*, 235
- 4.10 Conclusion, 236

5 COMMUNICATIONS LINK ANALYSIS

242

- 5.1 What the System Link Budget Tells the System Engineer, 243
- 5.2 The Channel, 244
 - 5.2.1 *The Concept of Free Space*, 244
 - 5.2.2 *Error-Performance Degradation*, 245
 - 5.2.3 *Sources of Signal Loss and Noise*, 245

- 5.3 Received Signal Power and Noise Power, 250
 - 5.3.1 *The Range Equation*, 250
 - 5.3.2 *Received Signal Power as a Function of Frequency*, 254
 - 5.3.3 *Path Loss is Frequency Dependent*, 256
 - 5.3.4 *Thermal Noise Power*, 258
- 5.4 Link Budget Analysis, 259
 - 5.4.1 *Two E_b/N_0 Values of Interest*, 262
 - 5.4.2 *Link Budgets are Typically Calculated in Decibels*, 263
 - 5.4.3 *How Much Link Margin is Enough?* 264
 - 5.4.4 *Link Availability*, 266
- 5.5 Noise Figure, Noise Temperature, and System Temperature, 270
 - 5.5.1 *Noise Figure*, 270
 - 5.5.2 *Noise Temperature*, 273
 - 5.5.3 *Line Loss*, 274
 - 5.5.4 *Composite Noise Figure and Composite Noise Temperature*, 276
 - 5.5.5 *System Effective Temperature*, 277
 - 5.5.6 *Sky Noise Temperature*, 282
- 5.6 Sample Link Analysis, 286
 - 5.6.1 *Link Budget Details*, 287
 - 5.6.2 *Receiver Figure of Merit*, 289
 - 5.6.3 *Received Isotropic Power*, 289
- 5.7 Satellite Repeaters, 290
 - 5.7.1 *Nonregenerative Repeaters*, 291
 - 5.7.2 *Nonlinear Repeater Amplifiers*, 295
- 5.8 System Trade-Offs, 296
- 5.9 Conclusion, 297

6 CHANNEL CODING: PART 1

304

- 6.1 Waveform Coding and Structured Sequences, 305
 - 6.1.1 *Antipodal and Orthogonal Signals*, 307
 - 6.1.2 *M-ary Signaling*, 308
 - 6.1.3 *Waveform Coding*, 309
 - 6.1.4 *Waveform-Coding System Example*, 313
- 6.2 Types of Error Control, 315
 - 6.2.1 *Terminal Connectivity*, 315
 - 6.2.2 *Automatic Repeat Request*, 316
- 6.3 Structured Sequences, 317
 - 6.3.1 *Channel Models*, 318
 - 6.3.2 *Code Rate and Redundancy*, 320
 - 6.3.3 *Parity Check Codes*, 321
 - 6.3.4 *Why Use Error-Correction Coding?* 323

- 6.4 Linear Block Codes, 328
 - 6.4.1 Vector Spaces, 329
 - 6.4.2 Vector Subspaces, 329
 - 6.4.3 A (6, 3) Linear Block Code Example, 330
 - 6.4.4 Generator Matrix, 331
 - 6.4.5 Systematic Linear Block Codes, 333
 - 6.4.6 Parity-Check Matrix, 334
 - 6.4.7 Syndrome Testing, 335
 - 6.4.8 Error Correction, 336
 - 6.4.9 Decoder Implementation, 340
- 6.5 Error-Detecting and Correcting Capability, 342
 - 6.5.1 Weight and Distance of Binary Vectors, 342
 - 6.5.2 Minimum Distance of a Linear Code, 343
 - 6.5.3 Error Detection and Correction, 343
 - 6.5.4 Visualization of a 6-Tuple Space, 347
 - 6.5.5 Erasure Correction, 348
- 6.6 Usefulness of the Standard Array, 349
 - 6.6.1 Estimating Code Capability, 349
 - 6.6.2 An (n, k) Example, 351
 - 6.6.3 Designing the (8, 2) Code, 352
 - 6.6.4 Error Detection versus Error Correction Trade-Offs, 352
 - 6.6.5 The Standard Array Provides Insight, 356
- 6.7 Cyclic Codes, 356
 - 6.7.1 Algebraic Structure of Cyclic Codes, 357
 - 6.7.2 Binary Cyclic Code Properties, 358
 - 6.7.3 Encoding in Systematic Form, 359
 - 6.7.4 Circuit for Dividing Polynomials, 360
 - 6.7.5 Systematic Encoding with an $(n - k)$ -Stage Shift Register, 363
 - 6.7.6 Error Detection with an $(n - k)$ -Stage Shift Register, 365
- 6.8 Well-Known Block Codes, 366
 - 6.8.1 Hamming Codes, 366
 - 6.8.2 Extended Golay Code, 369
 - 6.8.3 BCH Codes, 370
- 6.9 Conclusion, 374

7 CHANNEL CODING: PART 2

381

- 7.1 Convolutional Encoding, 382
- 7.2 Convolutional Encoder Representation, 384
 - 7.2.1 Connection Representation, 385
 - 7.2.2 State Representation and the State Diagram, 389
 - 7.2.3 The Tree Diagram, 391
 - 7.2.4 The Trellis Diagram, 393
- 7.3 Formulation of the Convolutional Decoding Problem, 395
 - 7.3.1 Maximum Likelihood Decoding, 395

- 7.3.2 *Channel Models: Hard versus Soft Decisions*, 396
- 7.3.3 *The Viterbi Convolutional Decoding Algorithm*, 401
- 7.3.4 *An Example of Viterbi Convolutional Decoding*, 401
- 7.3.5 *Decoder Implementation*, 405
- 7.3.6 *Path Memory and Synchronization*, 408
- 7.4 *Properties of Convolutional Codes*, 408
 - 7.4.1 *Distance Properties of Convolutional Codes*, 408
 - 7.4.2 *Systematic and Nonsystematic Convolutional Codes*, 413
 - 7.4.3 *Catastrophic Error Propagation in Convolutional Codes*, 414
 - 7.4.4 *Performance Bounds for Convolutional Codes*, 415
 - 7.4.5 *Coding Gain*, 416
 - 7.4.6 *Best Known Convolutional Codes*, 418
 - 7.4.7 *Convolutional Code Rate Trade-Off*, 420
 - 7.4.8 *Soft-Decision Viterbi Decoding*, 420
- 7.5 *Other Convolutional Decoding Algorithms*, 422
 - 7.5.1 *Sequential Decoding*, 422
 - 7.5.2 *Comparisons and Limitations of Viterbi and Sequential Decoding*, 425
 - 7.5.3 *Feedback Decoding*, 427
- 7.6 *Conclusion*, 429

8 CHANNEL CODING: PART 3

436

- 8.1 *Reed-Solomon Codes*, 437
 - 8.1.1 *Reed-Solomon Error Probability*, 438
 - 8.1.2 *Why R-S Codes Perform Well Against Burst Noise*, 441
 - 8.1.3 *R-S Performance as a Function of Size, Redundancy, and Code Rate*, 441
 - 8.1.4 *Finite Fields*, 445
 - 8.1.5 *Reed-Solomon Encoding*, 450
 - 8.1.6 *Reed-Solomon Decoding*, 454
- 8.2 *Interleaving and Concatenated Codes*, 461
 - 8.2.1 *Block Interleaving*, 463
 - 8.2.2 *Convolutional Interleaving*, 466
 - 8.2.3 *Concatenated Codes*, 468
- 8.3 *Coding and Interleaving Applied to the Compact Disc Digital Audio System*, 469
 - 8.3.1 *CIRC Encoding*, 470
 - 8.3.2 *CIRC Decoding*, 472
 - 8.3.3 *Interpolation and Muting*, 474
- 8.4 *Turbo Codes*, 475
 - 8.4.1 *Turbo Code Concepts*, 477
 - 8.4.2 *Log-Likelihood Algebra*, 481
 - 8.4.3 *Product Code Example*, 482
 - 8.4.4 *Encoding with Recursive Systematic Codes*, 488
 - 8.4.5 *A Feedback Decoder*, 493

381

- 8.4.6 *The MAP Decoding Algorithm*, 498
- 8.4.7 *MAP Decoding Example*, 504
- 8.5 Conclusion, 509
- Appendix 8A The Sum of Log-Likelihood Ratios, 510

9 MODULATION AND CODING TRADE-OFFS

520

- 9.1 Goals of the Communications System Designer, 521
- 9.2 Error Probability Plane, 522
- 9.3 Nyquist Minimum Bandwidth, 524
- 9.4 Shannon–Hartley Capacity Theorem, 525
 - 9.4.1 *Shannon Limit*, 528
 - 9.4.2 *Entropy*, 529
 - 9.4.3 *Equivocation and Effective Transmission Rate*, 532
- 9.5 Bandwidth Efficiency Plane, 534
 - 9.5.1 *Bandwidth Efficiency of MPSK and MFSK Modulation*, 535
 - 9.5.2 *Analogies Between Bandwidth-Efficiency and Error Probability Planes*, 536
- 9.6 Modulation and Coding Trade-Offs, 537
- 9.7 Defining, Designing, and Evaluating Digital Communication Systems, 538
 - 9.7.1 *M-ary Signaling*, 539
 - 9.7.2 *Bandwidth-Limited Systems*, 540
 - 9.7.3 *Power-Limited Systems*, 541
 - 9.7.4 *Requirements for MPSK and MFSK Signaling*, 542
 - 9.7.5 *Bandwidth-Limited Uncoded System Example*, 543
 - 9.7.6 *Power-Limited Uncoded System Example*, 545
 - 9.7.7 *Bandwidth-Limited and Power-Limited Coded System Example*, 547
- 9.8 Bandwidth-Efficient Modulation, 555
 - 9.8.1 *QPSK and Offset QPSK Signaling*, 555
 - 9.8.2 *Minimum Shift Keying*, 559
 - 9.8.3 *Quadrature Amplitude Modulation*, 563
- 9.9 Modulation and Coding for Bandlimited Channels, 566
 - 9.9.1 *Commercial Telephone Modems*, 567
 - 9.9.2 *Signal Constellation Boundaries*, 568
 - 9.9.3 *Higher Dimensional Signal Constellations*, 569
 - 9.9.4 *Higher-Density Lattice Structures*, 572
 - 9.9.5 *Combined Gain: N-Sphere Mapping and Dense Lattice*, 573
- 9.10 Trellis-Coded Modulation, 573
 - 9.10.1 *The Idea Behind Trellis-Coded Modulation (TCM)*, 574
 - 9.10.2 *TCM Encoding*, 576
 - 9.10.3 *TCM Decoding*, 580
 - 9.10.4 *Other Trellis Codes*, 583

- 9.10.5 *Trellis-Coded Modulation Example*, 585
- 9.10.6 *Multi-Dimensional Trellis-Coded Modulation*, 589
- 9.11 *Conclusion*, 590

10 SYNCHRONIZATION

598

- 10.1 *Introduction*, 599
 - 10.1.1 *Synchronization Defined*, 599
 - 10.1.2 *Costs versus Benefits*, 601
 - 10.1.3 *Approach and Assumptions*, 602
- 10.2 *Receiver Synchronization*, 603
 - 10.2.1 *Frequency and Phase Synchronization*, 603
 - 10.2.2 *Symbol Synchronization—Discrete Symbol Modulations*, 625
 - 10.2.3 *Synchronization with Continuous-Phase Modulations (CPM)*, 631
 - 10.2.4 *Frame Synchronization*, 639
- 10.3 *Network Synchronization*, 643
 - 10.3.1 *Open-Loop Transmitter Synchronization*, 644
 - 10.3.2 *Closed-Loop Transmitter Synchronization*, 647
- 10.4 *Conclusion*, 649

11 MULTIPLEXING AND MULTIPLE ACCESS

656

- 11.1 *Allocation of the Communications Resource*, 657
 - 11.1.1 *Frequency-Division Multiplexing/Multiple Access*, 660
 - 11.1.2 *Time-Division Multiplexing/Multiple Access*, 665
 - 11.1.3 *Communications Resource Channelization*, 668
 - 11.1.4 *Performance Comparison of FDMA and TDMA*, 668
 - 11.1.5 *Code-Division Multiple Access*, 672
 - 11.1.6 *Space-Division and Polarization-Division Multiple Access*, 674
- 11.2 *Multiple Access Communications System and Architecture*, 676
 - 11.2.1 *Multiple Access Information Flow*, 677
 - 11.2.2 *Demand Assignment Multiple Access*, 678
- 11.3 *Access Algorithms*, 678
 - 11.3.1 *ALOHA*, 678
 - 11.3.2 *Slotted ALOHA*, 682
 - 11.3.3 *Reservation-ALOHA*, 683
 - 11.3.4 *Performance Comparison of S-ALOHA and R-ALOHA*, 684
 - 11.3.5 *Polling Techniques*, 686
- 11.4 *Multiple Access Techniques Employed with INTELSAT*, 689
 - 11.4.1 *Preassigned FDM/FM/FDMA or MCPC Operation*, 690
 - 11.4.2 *MCPC Modes of Accessing an INTELSAT Satellite*, 690
 - 11.4.3 *SPADE Operation*, 693
 - 11.4.4 *TDMA in INTELSAT*, 698
 - 11.4.5 *Satellite-Switched TDMA in INTELSAT*, 704

- 11.5 Multiple Access Techniques for Local Area Networks, 708
 - 11.5.1 *Carrier-Sense Multiple Access Networks*, 708
 - 11.5.2 *Token-Ring Networks*, 710
 - 11.5.3 *Performance Comparison of CSMA/CD and Token-Ring Networks*, 711
- 11.6 Conclusion, 713

12 SPREAD-SPECTRUM TECHNIQUES

718

- 12.1 Spread-Spectrum Overview, 719
 - 12.1.1 *The Beneficial Attributes of Spread-Spectrum Systems*, 720
 - 12.1.2 *A Catalog of Spreading Techniques*, 724
 - 12.1.3 *Model for Direct-Sequence Spread-Spectrum Interference Rejection*, 726
 - 12.1.4 *Historical Background*, 727
- 12.2 Pseudonoise Sequences, 728
 - 12.2.1 *Randomness Properties*, 729
 - 12.2.2 *Shift Register Sequences*, 729
 - 12.2.3 *PN Autocorrelation Function*, 730
- 12.3 Direct-Sequence Spread-Spectrum Systems, 732
 - 12.3.1 *Example of Direct Sequencing*, 734
 - 12.3.2 *Processing Gain and Performance*, 735
- 12.4 Frequency Hopping Systems, 738
 - 12.4.1 *Frequency Hopping Example*, 740
 - 12.4.2 *Robustness*, 741
 - 12.4.3 *Frequency Hopping with Diversity*, 741
 - 12.4.4 *Fast Hopping versus Slow Hopping*, 742
 - 12.4.5 *FFH/MFSK Demodulator*, 744
 - 12.4.6 *Processing Gain*, 745
- 12.5 Synchronization, 745
 - 12.5.1 *Acquisition*, 746
 - 12.5.2 *Tracking*, 751
- 12.6 Jamming Considerations, 754
 - 12.6.1 *The Jamming Game*, 754
 - 12.6.2 *Broadband Noise Jamming*, 759
 - 12.6.3 *Partial-Band Noise Jamming*, 760
 - 12.6.4 *Multiple-Tone Jamming*, 763
 - 12.6.5 *Pulse Jamming*, 763
 - 12.6.6 *Repeat-Back Jamming*, 765
 - 12.6.7 *BLADES System*, 768
- 12.7 Commercial Applications, 769
 - 12.7.1 *Code-Division Multiple Access*, 769
 - 12.7.2 *Multipath Channels*, 771
 - 12.7.3 *The FCC Part 15 Rules for Spread-Spectrum Systems*, 772
 - 12.7.4 *Direct Sequence versus Frequency Hopping*, 773
- 12.8 Cellular Systems, 775
 - 12.8.1 *Direct Sequence CDMA*, 776

- 12.8.2 *Analog FM versus TDMA versus CDMA*, 779
- 12.8.3 *Interference-Limited versus Dimension-Limited Systems*, 781
- 12.8.4 *IS-95 CDMA Digital Cellular System*, 782
- 12.9 Conclusion, 795

13 SOURCE CODING

803

- 13.1 Sources, 804
 - 13.1.1 *Discrete Sources*, 804
 - 13.1.2 *Waveform Sources*, 809
- 13.2 Amplitude Quantizing, 811
 - 13.2.1 *Quantizing Noise*, 813
 - 13.2.2 *Uniform Quantizing*, 816
 - 13.2.3 *Saturation*, 820
 - 13.2.4 *Dithering*, 823
 - 13.2.5 *Nonuniform Quantizing*, 826
- 13.3 Differential Pulse-Code Modulation, 835
 - 13.3.1 *One-Tap Prediction*, 838
 - 13.3.2 *N-Tap Prediction*, 839
 - 13.3.3 *Delta Modulation*, 841
 - 13.3.4 *Sigma-Delta Modulation*, 842
 - 13.3.5 *Sigma-Delta A-to-D Converter (ADC)*, 847
 - 13.3.6 *Sigma-Delta D-to-A Converter (DAC)*, 848
- 13.4 Adaptive Prediction, 850
 - 13.4.1 *Forward Prediction*, 851
 - 13.4.2 *Synthesis/Analysis Coding*, 852
- 13.5 Block Coding, 853
 - 13.5.1 *Vector Quantizing*, 854
- 13.6 Transform Coding, 856
 - 13.6.1 *Quantization for Transform Coding*, 857
 - 13.6.2 *Subband Coding*, 857
- 13.7 Source Coding for Digital Data, 859
 - 13.7.1 *Properties of Codes*, 860
 - 13.7.2 *Huffman Codes*, 862
 - 13.7.3 *Run-Length Codes*, 866
- 13.8 Examples of Source Coding, 870
 - 13.8.1 *Audio Compression*, 870
 - 13.8.2 *Image Compression*, 875
- 13.9 Conclusion, 884

14 ENCRYPTION AND DECRYPTION

890

- 14.1 Models, Goals, and Early Cipher Systems, 891
 - 14.1.1 *A Model of the Encryption and Decryption Process*, 893
 - 14.1.2 *System Goals*, 893
 - 14.1.3 *Classic Threats*, 893

Contents

xv

Contents

718

works, 711

20

- 14.1.4 *Classic Ciphers*, 894
- 14.2 The Secrecy of a Cipher System, 897
 - 14.2.1 *Perfect Secrecy*, 897
 - 14.2.2 *Entropy and Equivocation*, 900
 - 14.2.3 *Rate of a Language and Redundancy*, 902
 - 14.2.4 *Unicity Distance and Ideal Secrecy*, 902
- 14.3 Practical Security, 905
 - 14.3.1 *Confusion and Diffusion*, 905
 - 14.3.2 *Substitution*, 905
 - 14.3.3 *Permutation*, 907
 - 14.3.4 *Product Cipher Systems*, 908
 - 14.3.5 *The Data Encryption Standard*, 909
- 14.4 Stream Encryption, 915
 - 14.4.1 *Example of Key Generation Using a Linear Feedback Shift Register*, 916
 - 14.4.2 *Vulnerabilities of Linear Feedback Shift Registers*, 917
 - 14.4.3 *Synchronous and Self-Synchronous Stream Encryption Systems*, 919
- 14.5 Public Key Cryptosystems, 920
 - 14.5.1 *Signature Authentication using a Public Key Cryptosystem*, 921
 - 14.5.2 *A Trapdoor One-Way Function*, 922
 - 14.5.3 *The Rivest-Shamir-Adelman Scheme*, 923
 - 14.5.4 *The Knapsack Problem*, 925
 - 14.5.5 *A Public Key Cryptosystem based on a Trapdoor Knapsack*, 927
- 14.6 Pretty Good Privacy, 929
 - 14.6.1 *Triple-DES, CAST, and IDEA*, 931
 - 14.6.2 *Diffie-Hellman (Elgamal Variation) and RSA*, 935
 - 14.6.3 *PGP Message Encryption*, 936
 - 14.6.4 *PGP Authentication and Signature*, 937
- 14.7 Conclusion, 940

15 FADING CHANNELS

944

- 15.1 The Challenge of Communicating over Fading Channels, 945
- 15.2 Characterizing Mobile-Radio Propagation, 947
 - 15.2.1 *Large-Scale Fading*, 951
 - 15.2.2 *Small-Scale Fading*, 953
- 15.3 Signal Time-Spreading, 958
 - 15.3.1 *Signal Time-Spreading Viewed in the Time-Delay Domain*, 958
 - 15.3.2 *Signal Time-Spreading Viewed in the Frequency Domain*, 960
 - 15.3.3 *Examples of Flat Fading and Frequency-Selective Fading*, 965
- 15.4 Time Variance of the Channel Caused by Motion, 966
 - 15.4.1 *Time Variance Viewed in the Time Domain*, 966
 - 15.4.2 *Time Variance Viewed in the Doppler-Shift Domain*, 969
 - 15.4.3 *Performance over a Slow- and Flat-Fading Rayleigh Channel*, 975

- 15.5 Mitigating the Degradation Effects of Fading, 978
 - 15.5.1 Mitigation to Combat Frequency-Selective Distortion, 980
 - 15.5.2 Mitigation to Combat Fast-Fading Distortion, 982
 - 15.5.3 Mitigation to Combat Loss in SNR, 983
 - 15.5.4 Diversity Techniques, 984
 - 15.5.5 Modulation Types for Fading Channels, 987
 - 15.5.6 The Role of an Interleaver, 988
- 15.6 Summary of the Key Parameters Characterizing Fading Channels, 992
 - 15.6.1 Fast Fading Distortion: Case 1, 992
 - 15.6.2 Frequency-Selective Fading Distortion: Case 2, 993
 - 15.6.3 Fast-Fading and Frequency-Selective Fading Distortion: Case 3, 993
- 15.7 Applications: Mitigating the Effects of Frequency-Selective Fading, 996
 - 15.7.1 The Viterbi Equalizer as Applied to GSM, 996
 - 15.7.2 The Rake Receiver as Applied to Direct-Sequence Spread-Spectrum (DS/SS) Systems, 999
- 15.8 Conclusion, 1001

A A REVIEW OF FOURIER TECHNIQUES

1012

- A.1 Signals, Spectra, and Linear Systems, 1012
- A.2 Fourier Techniques for Linear System Analysis, 1012
 - A.2.1 Fourier Series Transform, 1014
 - A.2.2 Spectrum of a Pulse Train, 1018
 - A.2.3 Fourier Integral Transform, 1020
- A.3 Fourier Transform Properties, 1021
 - A.3.1 Time Shifting Property, 1022
 - A.3.2 Frequency Shifting Property, 1022
- A.4 Useful Functions, 1023
 - A.4.1 Unit Impulse Function, 1023
 - A.4.2 Spectrum of a Sinusoid, 1023
- A.5 Convolution, 1025
 - A.5.1 Graphical Example of Convolution, 1027
 - A.5.2 Time Convolution Property, 1028
 - A.5.3 Frequency Convolution Property, 1030
 - A.5.4 Convolution of a Function with a Unit Impulse, 1030
 - A.5.5 Demodulation Application of Convolution, 1031
- A.6 Tables of Fourier Transforms and Operations, 1033

B FUNDAMENTALS OF STATISTICAL DECISION THEORY

1035

- B.1 Bayes' Theorem, 1035
 - B.1.1 Discrete Form of Bayes' Theorem, 1036
 - B.1.2 Mixed Form of Bayes' Theorem, 1038
- B.2 Decision Theory, 1040
 - B.2.1 Components of the Decision Theory Problem, 1040

B.2.2	<i>The Likelihood Ratio Test and the Maximum A Posteriori Criterion,</i>	1041
B.2.3	<i>The Maximum Likelihood Criterion,</i>	1042
B.3	<i>Signal Detection Example,</i>	1042
B.3.1	<i>The Maximum Likelihood Binary Decision,</i>	1042
B.3.2	<i>Probability of Bit Error,</i>	1044
C	RESPONSE OF A CORRELATOR TO WHITE NOISE	1047
D	OFTEN-USED IDENTITIES	1049
E	s-DOMAIN, z-DOMAIN AND DIGITAL FILTERING	1051
E.1	<i>The Laplace Transform,</i>	1051
E.1.1	<i>Standard Laplace Transforms,</i>	1052
E.1.2	<i>Laplace Transform Properties,</i>	1053
E.1.3	<i>Using the Laplace Transform,</i>	1054
E.1.4	<i>Transfer Function,</i>	1055
E.1.5	<i>RC Circuit Low Pass Filtering,</i>	1056
E.1.6	<i>Poles and Zeroes,</i>	1056
E.1.7	<i>Linear System Stability,</i>	1057
E.2	<i>The z-Transform,</i>	1058
E.2.1	<i>Calculating the z-Transform,</i>	1058
E.2.2	<i>The Inverse z-Transform,</i>	1059
E.3	<i>Digital Filtering,</i>	1060
E.3.1	<i>Digital Filter Transfer Function,</i>	1061
E.3.2	<i>Single Pole Filter Stability,</i>	1062
E.3.3	<i>General Digital Filter Stability,</i>	1063
E.3.4	<i>z-Plane Pole-Zero Diagram and the Unit Circle,</i>	1063
E.3.5	<i>Discrete Fourier Transform of Digital Filter Impulse Response,</i>	1064
E.4	<i>Finite Impulse Response Filter Design,</i>	1065
E.4.1	<i>FIR Filter Design,</i>	1065
E.4.2	<i>The FIR Differentiator,</i>	1067
E.5	<i>Infinite Impulse Response Filter Design,</i>	1069
E.5.1	<i>Backward Difference Operator,</i>	1069
E.5.2	<i>IIR Filter Design using the Bilinear Transform,</i>	1070
E.5.3	<i>The IIR Integrator,</i>	1071
F	LIST OF SYMBOLS	1072
	INDEX	1074

Preface

This second edition of *Digital Communications: Fundamentals and Applications* represents an update of the original publication. The key features that have been updated are:

- The error-correction coding chapters have been expanded, particularly in the areas of Reed-Solomon codes, turbo codes, and trellis-coded modulation.
- A new chapter on fading channels and how to mitigate the degrading effects of fading has been introduced.
- Explanations and descriptions of essential digital communication concepts have been amplified.
- End-of-chapter problem sets have been expanded. Also, end-of-chapter question sets (and where to find the answers), as well as end-of-chapter CD exercises have been added.
- A compact disc (CD) containing an educational version of the design software SystemView by ELANIX[®] accompanies the textbook. The CD contains a workbook with over 200 exercises, as well as a concise tutorial on digital signal processing (DSP). CD exercises in the workbook reinforce material in the textbook; concepts can be explored by viewing waveforms with a windows-based PC and by changing parameters to see the effects on the overall system. Some of the exercises provide basic training in using SystemView; others provide additional training in DSP techniques.

The teaching of a one-semester university course proceeds in a very different manner compared with that of a short-course in the same subject. At the university, one has the luxury of time—time to develop the needed skills and mathematical tools, time to practice the ideas with homework exercises. In a short-course, the treatment is almost backwards compared with the university. Because of the time factor, a short-course teacher must “jump in” early with essential concepts and applications. One of the vehicles that I found useful in structuring a short course was to start by handing out a check list. This was not merely an outline of the curriculum. It represented a collection of concepts and nomenclature that are not clearly documented, and are often misunderstood. The short-course students were thus initiated into the course by being challenged. I promised them that once they felt comfortable describing each issue, or answering each question on the list, they would be well on their way toward becoming knowledgeable in the field of digital communications. I have learned that this list of essential concepts is just as valuable for teaching full-semester courses as it is for short courses. Here then is my “check list” for digital communications.

1. What mathematical dilemma is the cause for there being several definitions of bandwidth? (See Section 1.7.2.)
2. Why is the ratio of bit energy-to-noise power spectral density, E_b/N_0 , a natural figure-of-merit for digital communication systems? (See Section 3.1.5.)
3. When representing timed events, what dilemma can easily result in confusing the most-significant bit (MSB) and the least-significant bit (LSB)? (See Section 3.2.3.2.)
4. The error performance of digital signaling suffers primarily from two degradation types. a) loss in signal-to-noise ratio, b) distortion resulting in an irreducible bit-error probability. How do they differ? (See Section 3.3.2.)
5. Often times, providing more E_b/N_0 will not mitigate the degradation due to intersymbol interference (ISI). Explain why. (See Section 3.3.2.)
6. At what location in the system is E_b/N_0 defined? (See Section 4.3.2.)
7. Digital modulation schemes fall into one of two classes with opposite behavior characteristics. a) orthogonal signaling, b) phase/amplitude signaling. Describe the behavior of each class. (See Sections 4.8.2 and 9.7.)
8. Why do binary phase shift keying (BPSK) and quaternary phase shift keying (QPSK) manifest the same bit-error-probability relationship? Does the same hold true for M -ary pulse amplitude modulation (M -PAM) and M^2 -ary quadrature amplitude modulation (M^2 -QAM) bit-error probability? (See Sections 4.8.4 and 9.8.3.1.)
9. In orthogonal signaling, why does error-performance improve with higher dimensional signaling? (See Section 4.8.5.)
10. Why is *free-space loss* a function of wavelength? (See Section 5.3.3.)
11. What is the relationship between received signal to noise (S/N) ratio and carrier to noise (C/N) ratio? (See Section 5.4.)
12. Describe four types of trade-offs that can be accomplished by using an error-correcting code. (See Section 6.3.4.)

13. Why do traditional error-correcting codes yield error-performance degradation at low values of E_b/N_0 ? (See Section 6.3.4.6)
14. Of what use is the *standard array* in understanding a block code, and in evaluating its capability? (See Section 6.6.5.)
15. Why is the Shannon limit of -1.6 dB not a useful goal in the design of real systems? (See Section 8.4.5.2.)
16. What are the consequences of the fact that the Viterbi decoding algorithm does not yield *a posteriori* probabilities? What is a more descriptive name for the Viterbi algorithm? (See Section 8.4.6.)
17. Why do binary and 4-ary orthogonal frequency shift keying (FSK) manifest the same bandwidth-efficiency relationship? (See Section 9.5.1.)
18. Describe the subtle energy and rate transformations of received signals: from data-bits to channel-bits to symbols to chips. (See Section 9.7.7.)
19. Define the following terms: Baud, State, Communications Resource, Chip, Robust Signal. (See Sections 1.1.3 and 7.2.2, Chapter 11, and Sections 12.3.2 and 12.4.2.)
20. For the case of a mobile system that experiences fading, why is signal dispersion independent of fading rapidity? (See Section 15.4.1.1.)

I hope you find it useful to be challenged in this way. Now, let us describe the purpose of the book in a more methodical way. This second edition is intended to provide a comprehensive coverage of digital communication systems for senior level undergraduates, first year graduate students, and practicing engineers. Though the emphasis is on digital communications, necessary analog fundamentals are included since analog waveforms are used for the radio transmission of digital signals. The key feature of a digital communication system is that it deals with a finite set of discrete messages, in contrast to an analog communication system in which messages are defined on a continuum. The objective at the receiver of the digital system is not to reproduce a waveform with precision; it is instead to determine from a noise-perturbed signal, which of the finite set of waveforms had been sent by the transmitter. In fulfillment of this objective, there has arisen an impressive assortment of signal processing techniques.

The book develops these techniques in the context of a unified structure. The structure, in block diagram form, appears at the beginning of each chapter; blocks in the diagram are emphasized, when appropriate, to correspond to the subject of that chapter. Major purposes of the book are to add organization and structure to a field that has grown and continues to grow rapidly, and to insure awareness of the "big picture" even while delving into the details. Signals and key processing steps are traced from the information source through the transmitter, channel, receiver, and ultimately to the information sink. Signal transformations are organized according to nine functional classes: Formatting and source coding, Baseband signaling, Bandpass signaling, Equalization, Channel coding, Multiplexing and multiple access, Spreading, Encryption, and Synchronization. Throughout the book, emphasis is placed on system goals and the need to trade off basic system parameters such as signal-to-noise ratio, probability of error, and bandwidth expenditure.

ORGANIZATION OF THE BOOK

Chapter 1 introduces the overall digital communication system and the basic signal transformations that are highlighted in subsequent chapters. Some basic ideas of random variables and the *additive white Gaussian noise* (AWGN) model are reviewed. Also, the relationship between power spectral density and autocorrelation, and the basics of signal transmission through linear systems are established. Chapter 2 covers the signal processing step, known as *formatting*, in order to render an information signal compatible with a digital system. Chapter 3 emphasizes *base-band signaling*, the detection of signals in Gaussian noise, and receiver optimization. Chapter 4 deals with *bandpass signaling* and its associated modulation and demodulation/detection techniques. Chapter 5 deals with *link analysis*, an important subject for providing overall system insight; it considers some subtleties that are often missed. Chapters 6, 7, and 8 deal with *channel coding*—a cost-effective way of providing a variety of system performance trade-offs. Chapter 6 emphasizes *linear block codes*, Chapter 7 deals with *convolutional codes*, and Chapter 8 deals with *Reed–Solomon codes* and *concatenated codes* such as *turbo codes*.

Chapter 9 considers various modulation/coding system *trade-offs* dealing with probability of bit-error performance, bandwidth efficiency, and signal-to-noise ratio. It also treats the important area of coded modulation, particularly *trellis-coded modulation*. Chapter 10 deals with *synchronization* for digital systems. It covers phase-locked loop implementation for achieving carrier synchronization. It covers bit synchronization, frame synchronization, and network synchronization, and it introduces some ways of performing synchronization using digital methods.

Chapter 11 treats *multiplexing* and *multiple access*. It explores techniques that are available for utilizing the communication resource efficiently. Chapter 12 introduces *spread spectrum* techniques and their application in such areas as multiple access, ranging, and interference rejection. This technology is important for both military and commercial applications. Chapter 13 deals with *source coding* which is a special class of data formatting. Both formatting and source coding involve digitization of data; the main difference between them is that source coding additionally involves data redundancy reduction. Rather than considering source coding immediately after formatting, it is purposely treated in a later chapter so as not to interrupt the presentation flow of the basic processing steps. Chapter 14 covers basic *encryption/decryption* ideas. It includes some classical concepts, as well as a class of systems called public key cryptosystems, and the widely used E-mail encryption software known as *Pretty Good Privacy* (PGP). Chapter 15 deals with *fading channels*. Here, we deal with applications, such as mobile radios, where characterization of the channel is much more involved than that of a nonfading one. The design of a communication system that will withstand the degradation effects of fading can be much more challenging than the design of its nonfading counterpart. In this chapter, we describe a variety of techniques that can mitigate the effects of fading, and we show some successful designs that have been implemented.

It is assumed that the reader is familiar with Fourier methods and convolution. Appendix A reviews these techniques, emphasizing those properties that are

particularly useful in the study of **communication** theory. It also assumed that the reader has a knowledge of basic **probability** and has some familiarity with random variables. Appendix B builds on these disciplines for a short treatment on statistical decision theory with emphasis on hypothesis testing—so important in the understanding of detection theory. A new section, **Appendix E**, has been added to serve as a short tutorial on *s*-domain, *z*-domain, and **digital** filtering. A concise DSP tutorial also appears on the CD that accompanies the book.

If the book is used for a two-term course, a simple partitioning is suggested; the first seven chapters can be taught in the first term, and the last eight chapters in the second term. If the book is used for a one-term introductory course, it is suggested that the course material be selected from the following chapters: 1, 2, 3, 4, 5, 6, 7, 9, 10, 12.

ACKNOWLEDGMENTS

It is difficult to write a technical book without contributions from others. I have received an abundance of such assistance, for which I am deeply grateful. For their generous help, I want to thank Dr. Andrew Viterbi, Dr. Chuck Wheatley, Dr. Ed Tiedeman, Dr. Joe Odenwalder, and Serge Willinegger of Qualcomm. I also want to thank Dr. Dariush Divsalar of Jet Propulsion Laboratory (JPL), Dr. Bob Bogusch of Mission Research, Dr. Tom Stanley of the Federal Communications Commission, Professor Larry Milstein of the University of California, San Diego, Professor Ray Pickholtz of George Washington University, Professor Daniel Costello of Notre Dame University, Professor Ted Rappaport of Virginia Polytechnic Institute, Phil Kossin of Lincom, Les Brown of Motorola, as well as Dr. Bob Price and Frank Amoroso.

I also want to acknowledge those people who played a big part in helping me with the first edition of the book. They are: Dr. Maurice King, Don Martin and Ned Feldman of The Aerospace Corporation, Dr. Marv Simon of JPL, Dr. Bill Lindsey of Lincom, Professor Wayne Stark of the University of Michigan, as well as Dr. Jim Omura, Dr. Adam Lender, and Dr. Todd Citron.

I want to thank Dr. Maurice King for contributing Chapter 10 on Synchronization, and Professor Fred Harris of San Diego State University for contributing Chapter 13 on Source Coding. Also, thanks to Michelle Landry for writing the sections on Pretty Good Privacy in Chapter 14, and to Andrew Guidi for contributing end-of-chapter problems in Chapter 15.

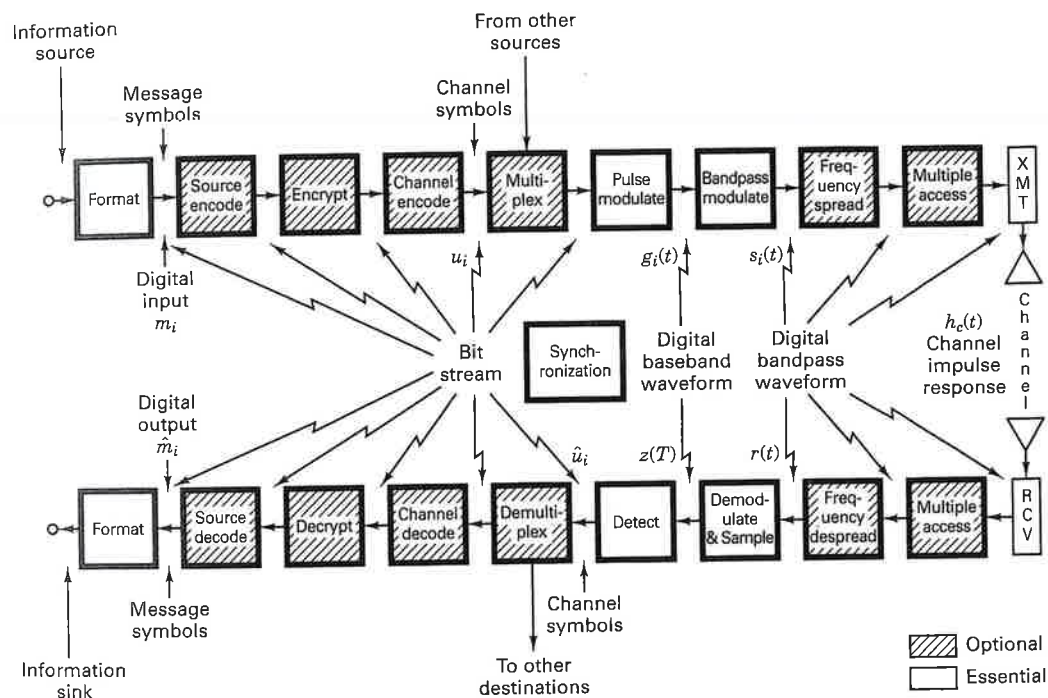
I am particularly indebted to my friends and colleagues Fred Harris, Professor Dan Bukofzer of California State University at Fresno, and Dr. Maury Schiff of Elanix, who put up with my incessant argumentative discussions anytime that I called on them. I also want to thank my very best teachers—they are my students at the University of California, Los Angeles, as well as those students all over the world who attended my short courses. Their questions motivated me and provoked me to write this second edition. I hope that I have answered all their questions with clarity.

I offer special thanks for technical clarifications that my son, Dean Sklar, suggested; he took on the difficult role of being his father's chief critic and "devil's advocate." I am particularly indebted to Professor Bob Stewart of the University of Strathclyde, Glasgow, who contributed countless hours of work in writing and preparing the CD and in authoring Appendix E. I thank Rose Kernan, my editor, for watching over me and this project, and I thank Bernard Goodwin, Publisher at Prentice Hall, for indulging me and believing in me. His recommendations were invaluable. Finally, I am extremely grateful to my wife, Gwen, for her encouragement, devotion, and valuable advice. She protected me from the "slings and arrows" of everyday life, making it possible for me to complete this second edition.

BERNARD SKLAR

Tarzana, California

Signals and Spectra



This book presents the ideas and techniques fundamental to digital communication systems. Emphasis is placed on system design goals and on the need for trade-offs among basic system parameters such as signal-to-noise ratio (SNR), probability of error, and bandwidth expenditure. We shall deal with the transmission of information (voice, video, or data) over a path (channel) that may consist of wires, waveguides, or space.

Digital communication systems are becoming increasingly attractive because of the ever-growing demand for data communication and because digital transmission offers data processing options and flexibilities not available with analog transmission. In this book, a digital system is often treated in the context of a satellite communications link. Sometimes the treatment is in the context of a mobile radio system, in which case signal transmission typically suffers from a phenomenon called *fading*. In general, the task of characterizing and mitigating the degradation effects of a fading channel is more challenging than performing similar tasks for a nonfading channel.

The principal feature of a digital communication system (DCS) is that during a finite interval of time, it sends a waveform from a finite set of possible waveforms, in contrast to an analog communication system, which sends a waveform from an infinite variety of waveform shapes with theoretically infinite resolution. In a DCS, the objective at the receiver is *not* to reproduce a transmitted waveform with precision; instead, the objective is to determine from a noise-perturbed signal which waveform from the finite set of waveforms was sent by the transmitter. An important measure of system performance in a DCS is the probability of error (P_E).

1.1 DIGITAL COMMUNICATION SIGNAL PROCESSING

1.1.1 Why Digital?

Why are communication systems, military and commercial alike, “going digital”? There are many reasons. The primary advantage is the ease with which digital signals, compared with analog signals, are regenerated. Figure 1.1 illustrates an ideal binary digital pulse propagating along a transmission line. The shape of the waveform is affected by two basic mechanisms: (1) as all transmission lines and circuits have some nonideal frequency transfer function, there is a distorting effect on the ideal pulse; and (2) unwanted electrical noise or other interference further distorts the pulse waveform. Both of these mechanisms cause the pulse shape to degrade as a function of line length, as shown in Figure 1.1. During the time that the transmitted pulse can still be reliably identified (before it is degraded to an ambiguous state), the pulse is amplified by a digital amplifier that recovers its original ideal shape. The pulse is thus “reborn” or regenerated. Circuits that perform this function at regular intervals along a transmission system are called *regenerative repeaters*.

Digital circuits are less subject to distortion and interference than are analog circuits. Because binary digital circuits operate in one of two states—fully on or fully off—to be meaningful, a disturbance must be large enough to change the circuit operating point from one state to the other. Such two-state operation facilitates signal regeneration and thus prevents noise and other disturbances from accumulating in transmission. Analog signals, however, are *not* two-state signals; they can take an *infinite variety* of shapes. With analog circuits, even a small disturbance can render the reproduced waveform unacceptably distorted. Once the analog signal is distorted, the distortion cannot be removed by amplification. Because accumulated noise is irrevocably bound to analog signals, they cannot be perfectly regenerated. With digital techniques, extremely low error rates producing

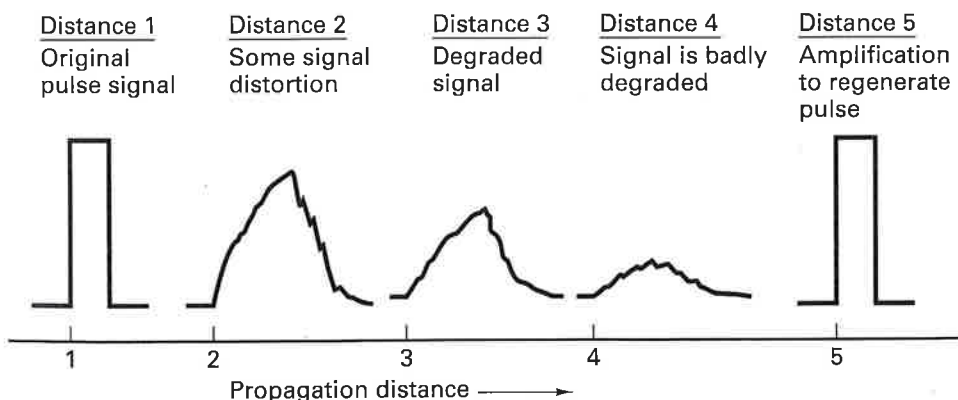


Figure 1.1 Pulse degradation and regeneration.

high signal fidelity are possible through error detection and correction but similar procedures are not available with analog.

There are other important advantages to digital communications. Digital circuits are *more reliable* and can be produced at a lower cost than analog circuits. Also, digital hardware lends itself to *more flexible* implementation than analog hardware [e.g., microprocessors, digital switching, and large-scale integrated (LSI) circuits]. The combining of digital signals using time-division multiplexing (TDM) is *simpler* than the combining of analog signals using frequency-division multiplexing (FDM). Different types of digital signals (data, telegraph, telephone, television) can be treated as identical signals in transmission and switching—a *bit is a bit*. Also, for convenient switching, digital messages can be handled in autonomous groups called *packets*. Digital techniques lend themselves naturally to signal processing functions that protect against interference and jamming, or that provide encryption and privacy. (Such techniques are discussed in Chapters 12 and 14, respectively.) Also, much data communication is from computer to computer, or from digital instruments or terminal to computer. Such digital terminations are naturally best served by digital communication links.

What are the costs associated with the beneficial attributes of digital communication systems? Digital systems tend to be very signal-processing intensive compared with analog. Also, digital systems need to allocate a significant share of their resources to the task of synchronization at various levels. (See Chapter 10.) With analog systems, on the other hand, synchronization often is accomplished more easily. One disadvantage of a digital communication system is *nongraceful degradation*. When the signal-to-noise ratio drops below a certain threshold, the quality of service can change suddenly from very good to very poor. In contrast, most analog communication systems degrade more gracefully.

1.1.2 Typical Block Diagram and Transformations

The functional block diagram shown in Figure 1.2 illustrates the signal flow and the signal-processing steps through a typical digital communication system (DCS). This figure can serve as a kind of road map, guiding the reader through the chapters of this book. The upper blocks—format, source encode, encrypt, channel encode, multiplex, pulse modulate, bandpass modulate, frequency spread, and multiple access—denote signal transformations from the source to the transmitter (XMT). The lower blocks denote signal transformations from the receiver (RCV) to the sink, essentially reversing the signal processing steps performed by the upper blocks. The *modulate* and *demodulate/detect* blocks together are called a *modem*. The term “modem” often encompasses several of the signal processing steps shown in Figure 1.2; when this is the case, the modem can be thought of as the “brains” of the system. The transmitter and receiver can be thought of as the “muscles” of the system. For wireless applications, the transmitter consists of a frequency up-conversion stage to a radio frequency (RF), a high-power amplifier, and an antenna. The receiver portion consists of an antenna and a low-noise amplifier (LNA). Frequency down-conversion is performed in the front end of the receiver and/or the demodulator.

n but similar

s. Digital circuits. than analog circuits. Digital circuits (LSI) are more efficient than analog circuits (TDM) in multiplexing, encryption, and decryption. Also, digital processing is more efficient than analog processing. Digital processing is more efficient than analog processing. Digital processing is more efficient than analog processing.

igital communication systems. Digital communication systems are more efficient than analog communication systems. Digital communication systems are more efficient than analog communication systems. Digital communication systems are more efficient than analog communication systems.

flow and the (DCS). This is a key feature of this code, multiple access—). The lower part of the system, essentially the "modulator and demodulator" often when this is applied to a transmitter, is less applicable to frequency. The system consists of an information source and a receiver.

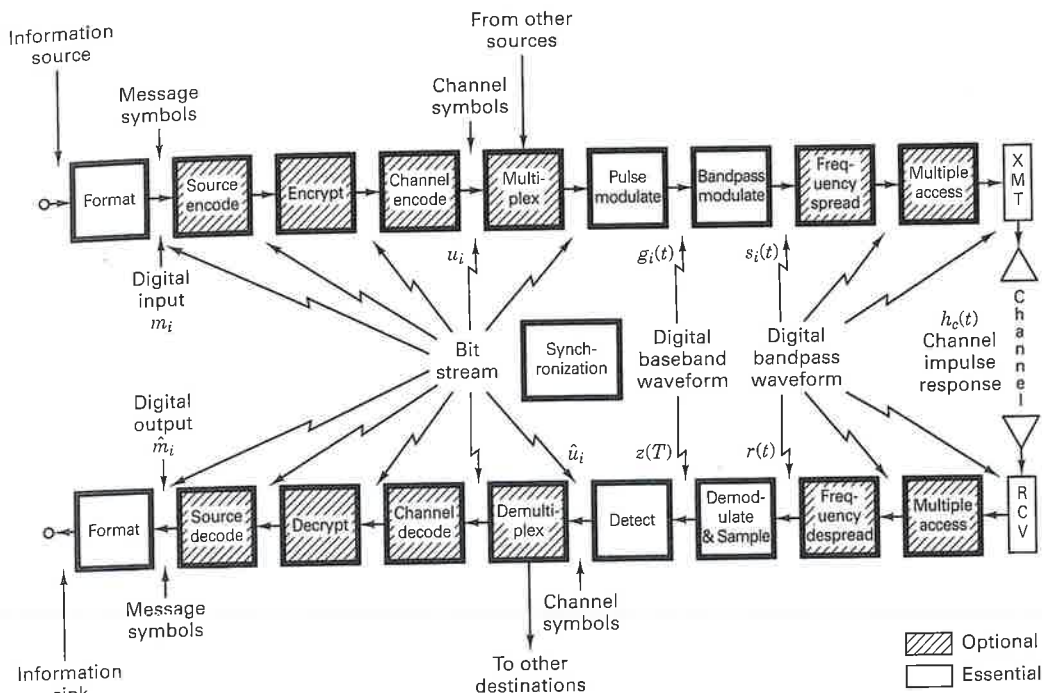


Figure 1.2 Block diagram of a typical digital communication system.

Figure 1.2 illustrates a kind of reciprocity between the blocks in the upper transmitter part of the figure and those in the lower receiver part. The signal processing steps that take place in the transmitter are, for the most part, reversed in the receiver. In Figure 1.2, the input information source is converted to binary digits (*bits*); the bits are then grouped to form *digital messages* or *message symbols*. Each such symbol (m_i , where $i = 1, \dots, M$) can be regarded as a member of a *finite alphabet* set containing M members. Thus, for $M = 2$, the message symbol m_i is binary (meaning that it constitutes just a single bit). Even though binary symbols fall within the general definition of M -ary, nevertheless the name M -ary is usually applied to those cases where $M > 2$; hence, such symbols are each made up of a sequence of two or more bits. (Compare such a *finite alphabet* in a DCS with an analog system, where the message waveform is typically a member of an infinite set of possible waveforms.) For systems that use *channel coding* (error correction coding), a sequence of message symbols becomes transformed to a sequence of *channel symbols* (code symbols), where each channel symbol is denoted u_i . Because a message symbol or a channel symbol can consist of a single bit or a grouping of bits, a sequence of such symbols is also described as a *bit stream*, as shown in Figure 1.2.

Consider the key signal processing blocks shown in Figure 1.2; only formatting, modulation, demodulation/detection, and synchronization are essential for a DCS. *Formatting* transforms the source information into bits, thus assuring com-

patibility between the information and the signal processing within the DCS. From this point in the figure up to the pulse-modulation block, the information remains in the form of a *bit stream*. Modulation is the process by which message symbols or channel symbols (when channel coding is used) are converted to *waveforms* that are compatible with the requirements imposed by the transmission channel. *Pulse modulation* is an essential step because each symbol to be transmitted must first be transformed from a binary representation (voltage levels representing binary ones and zeros) to a *baseband* waveform. The term baseband refers to a signal whose spectrum extends from (or near) dc up to some finite value, usually less than a few megahertz. The pulse-modulation block usually includes filtering for minimizing the transmission bandwidth. When pulse modulation is applied to binary symbols, the resulting binary waveform is called a pulse-code-modulation (PCM) waveform. There are several types of PCM waveforms (described in Chapter 2); in telephone applications, these waveforms are often called *line codes*. When pulse modulation is applied to nonbinary symbols, the resulting waveform is called an *M*-ary pulse-modulation waveform. There are several types of such waveforms, and they too are described in Chapter 2, where the one called *pulse-amplitude modulation* (PAM) is emphasized. After pulse modulation, each message symbol or channel symbol takes the form of a baseband waveform $g_i(t)$, where $i = 1, \dots, M$. In any electronic implementation, the bit stream, prior to pulse-modulation, is represented with voltage levels. One might wonder why there is a separate block for pulse modulation when in fact different voltage levels for binary ones and zeros can be viewed as impulses or as ideal rectangular pulses, each pulse occupying one bit time. There are two important differences between such voltage levels and the baseband waveforms used for modulation. First, the pulse-modulation block allows for a variety of binary and *M*-ary pulse-waveform types. Section 2.8.2 describes the different useful attributes of these types of waveforms. Second, the filtering within the pulse-modulation block yields pulses that occupy more than just one-bit time. Filtering yields pulses that are spread in time, thus the pulses are "smeared" into neighboring bit-times. This filtering is sometimes referred to as pulse shaping; it is used to contain the transmission bandwidth within some desired spectral region.

For an application involving RF transmission, the next important step is *bandpass modulation*; it is required whenever the transmission medium will not support the propagation of pulse-like waveforms. For such cases, the medium requires a bandpass waveform $s_i(t)$, where $i = 1, \dots, M$. The term *bandpass* is used to indicate that the baseband waveform $g_i(t)$ is frequency translated by a carrier wave to a frequency that is much larger than the spectral content of $g_i(t)$. As $s_i(t)$ propagates over the channel, it is impacted by the channel characteristics, which can be described in terms of the channel's *impulse response* $h_c(t)$ (see Section 1.6.1). Also, at various points along the signal route, additive random noise distorts the received signal $r(t)$, so that its reception must be termed a corrupted version of the signal $s_i(t)$ that was launched at the transmitter. The received signal $r(t)$ can be expressed as

$$r(t) = s_i(t) * h_c(t) + n(t) \quad i = 1, \dots, M \quad (1.1)$$

where $*$ represents a convolution operation (see Appendix A), and $n(t)$ represents a noise process (see Section 1.5.5).

In the reverse direction, the receiver front end and/or the demodulator provides frequency down-conversion for each bandpass waveform $r(t)$. The demodulator restores $r(t)$ to an optimally shaped baseband pulse $z(t)$ in preparation for detection. Typically, there can be several filters associated with the receiver and demodulator—filtering to remove unwanted high frequency terms (in the frequency down-conversion of bandpass waveforms), and filtering for pulse shaping. Equalization can be described as a filtering option that is used in or after the demodulator to reverse any degrading effects on the signal that were caused by the channel. Equalization becomes essential whenever the impulse response of the channel, $h_c(t)$, is so poor that the received signal is badly distorted. An equalizer is implemented to compensate for (i.e., remove or diminish) any signal distortion caused by a nonideal $h_c(t)$. Finally, the sampling step transforms the shaped pulse $z(t)$ to a sample $z(T)$, and the detection step transforms $z(T)$ to an estimate of the channel symbol $\hat{u}_i$ or an estimate of the message symbol $\hat{m}_i$ (if there is no channel coding). Some authors use the terms “demodulation” and “detection” interchangeably. However, in this book, *demodulation* is defined as recovery of a waveform (baseband pulse), and *detection* is defined as decision-making regarding the digital meaning of that waveform.

The other signal processing steps within the modem are design options for specific system needs. *Source coding* produces analog-to-digital (A/D) conversion (for analog sources) and removes redundant (unneeded) information. Note that a typical DCS would either use the *source coding* option (for both digitizing and compressing the source information), or it would use the simpler *formatting* transformation (for digitizing alone). A system would not use both source coding and formatting, because the former already includes the essential step of digitizing the information. Encryption, which is used to provide communication privacy, prevents unauthorized users from understanding messages and from injecting false messages into the system. *Channel coding*, for a given data rate, can reduce the probability of error, P_E , or reduce the required signal-to-noise ratio to achieve a desired P_E at the expense of transmission bandwidth or decoder complexity. *Multiplexing* and *multiple-access procedures* combine signals that might have different characteristics or might originate from different sources, so that they can share a portion of the communications resource (e.g., spectrum, time). Frequency spreading can produce a signal that is relatively invulnerable to interference (both natural and intentional) and can be used to enhance the privacy of the communicators. It is also a valuable technique used for multiple access.

The signal processing blocks shown in Figure 1.2 represent a typical arrangement; however, these blocks are sometimes implemented in a different order. For example, multiplexing can take place prior to channel encoding, or prior to modulation, or—with a two-step modulation process (subcarrier and carrier)—it can be performed between the two modulation steps. Similarly, frequency spreading can take place at various locations along the upper portion of Figure 1.2; its precise location depends on the particular technique used. Synchronization and its key element, a clock signal, is involved in the control of all signal processing within the

(1.1)

DCS. For simplicity, the synchronization block in Figure 1.2 is drawn without any connecting lines, when in fact it actually plays a role in regulating the operation of almost every block shown in the figure.

Figure 1.3 shows the basic signal processing functions, which may be viewed as transformations, classified into the following nine groups:

1. Formatting and source coding
2. Baseband signaling
3. Bandpass signaling
4. Equalization
5. Channel coding
6. Multiplexing and multiple access
7. Spreading
8. Encryption
9. Synchronization

Although this organization has some inherent overlap, it provides a useful structure for the book. Beginning with Chapter 2, the nine basic transformations are considered individually. In Chapter 2, the basic formatting techniques for transforming the source information into message symbols are discussed, as well as the selection of baseband pulse waveforms and pulse filtering for making the message symbols compatible with baseband transmission. The reverse steps of demodulation, equalization, sampling, and detection are described in Chapter 3. Formatting and source coding are similar processes, in that they both involve data digitization. However, the term "source coding" has taken on the connotation of data compression in addition to digitization; it is treated later (in Chapter 13), as a special case of formatting.

In Figure 1.3, the *Baseband Signaling* block contains a list of binary choices under the heading of PCM waveforms or line codes. In this block, a nonbinary category of waveforms called *M*-ary pulse modulation is also listed. Another transformation in Figure 1.3, labeled *Bandpass Signaling* is partitioned into two basic blocks, coherent and noncoherent. Demodulation is typically accomplished with the aid of *reference* waveforms. When the references used are a measure of all the signal attributes (particularly phase), the process is termed *coherent*; when phase information is not used, the process is termed *noncoherent*. Both techniques are detailed in Chapter 4.

Chapter 5 is devoted to *link analysis*. Of the many specifications, analyses, and tabulations that support a developing communication system, link analysis stands out in its ability to provide overall system insight. In Chapter 5 we bring together all the link fundamentals that are essential for the analysis of most communication systems.

Channel coding deals with the techniques used to enhance digital signals so that they are less vulnerable to such channel impairments as noise, fading, and jamming. In Figure 1.3 channel coding is partitioned into two blocks, waveform coding

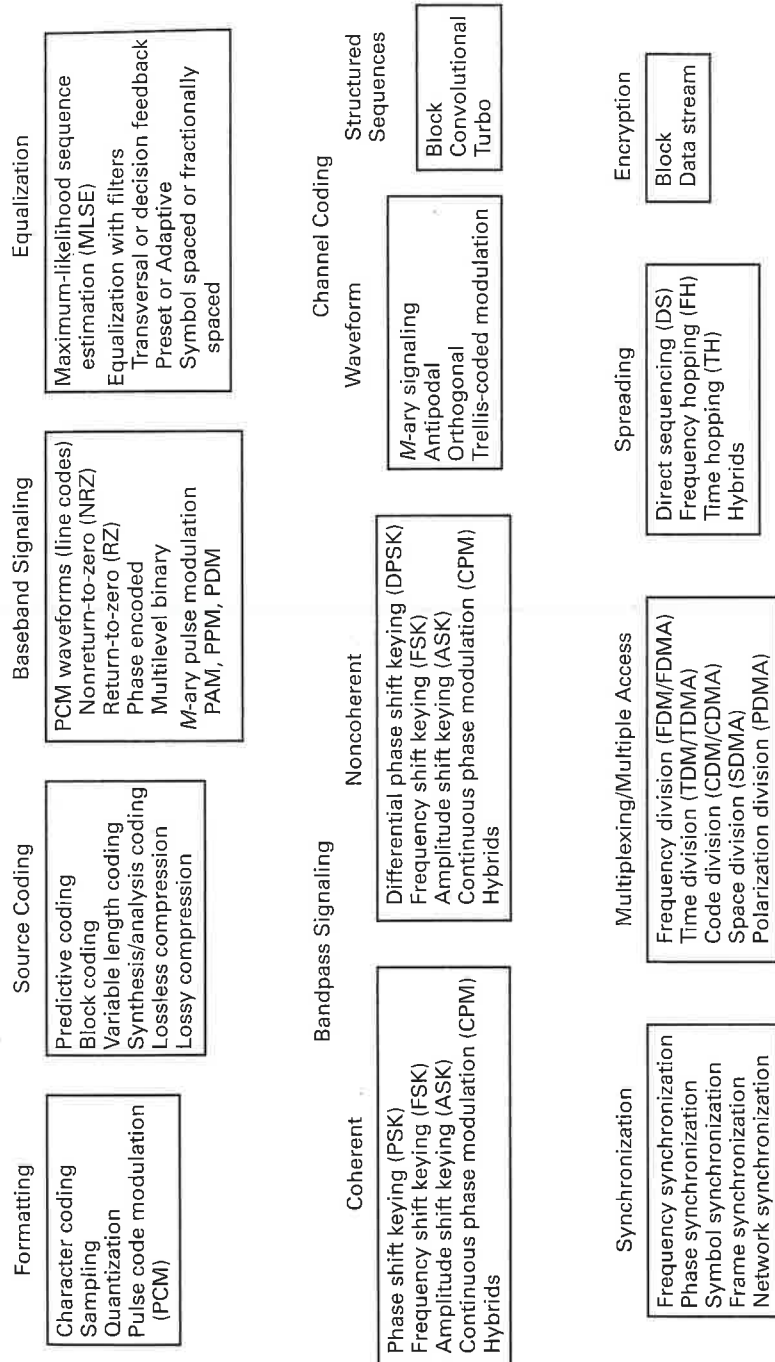


Figure 1.3 Basic digital communication transformations.

and structured sequences. *Waveform coding* involves the use of new waveforms, yielding improved detection performance over that of the original waveforms. *Structured sequences* involve the use of redundant bits to determine whether or not an error has occurred due to noise on the channel. One of these techniques, known as *automatic repeat request* (ARQ), simply recognizes the occurrence of an error and requests that the sender retransmit the message; other techniques, known as *forward error correction* (FEC), are capable of automatically correcting the errors (within specified limitations). Under the heading of structured sequences, we shall discuss three prevalent techniques—block, convolutional, and turbo coding. In Chapter 6, we primarily consider *linear block coding*. In Chapter 7 we consider *convolutional coding*, Viterbi decoding (and other decoding algorithms), and hard versus soft decoding procedures. Chapter 8 treats concatenated coding, which has led to the class of codes known as *turbo codes*, and it also examines the details of *Reed-Solomon* codes.

In Chapter 9 we summarize the design goals for a communication system and present various modulation and coding trade-offs that need to be considered in the design of a system. Theoretical limitations, such as the Nyquist criterion and the Shannon limit, are discussed. Also, *bandwidth-efficient* modulation schemes, such as trellis-coded modulation, are examined.

Chapter 10 deals with *synchronization*. In digital communications, synchronization involves the estimation of both time and frequency. The subject is divided into five subcategories as shown in Figure 1.3. Coherent systems need to synchronize their frequency reference with the carrier (and possibly subcarrier) in both frequency and phase. For noncoherent systems, phase synchronization is not needed. The fundamental time-synchronization process is symbol synchronization (or bit synchronization for binary symbols). The demodulator and detector need to know when to start and end the process of symbol detection and bit detection; a timing error will degrade detection performance. The next time-synchronization level, frame synchronization, allows the reconstruction of the message. Finally, network synchronization allows coordination with other users so resources may be used efficiently. In Chapter 10, we are concerned with the alignment of the timing of spatially separated periodic processes.

Chapter 11 deals with *multiplexing* and *multiple access*. The two terms mean very similar things. Both involve the idea of resource sharing. The main difference between the two is that multiplexing takes place locally (e.g., on a printed circuit board, within an assembly, or even within a facility), and multiple access takes place remotely (e.g., multiple users need to share the use of a satellite transponder). Multiplexing involves an algorithm that is known a priori; usually, it is hardwired into the system. Multiple access, on the other hand, is generally adaptive, and may require some overhead to enable the algorithm to operate. In Chapter 11, we discuss the classical ways of sharing a communications resource: frequency division, time division, and code division. Also, some of the multiple-access techniques that have emerged as a result of satellite communications are considered.

Chapter 12 introduces a transformation originally developed for military communications called *spreading*. The chapter deals with the spread spectrum techniques that are important for achieving interference protection and privacy.

Signals can be spread in frequency, in time, or in both frequency and time. This chapter primarily deals with frequency spreading. The chapter also illustrates how frequency-spreading techniques are used to share the bandwidth-limited resource in commercial cellular telephony.

Chapter 13 treats *source coding*, which involves the efficient description of source information. It deals with the process of compactly describing a signal to within a specified fidelity criterion. Source coding can be applied to digital or analog signals; by reducing data redundancy, source codes can reduce a system's data rate. Thus, the main advantage of source coding is to decrease the amount of required system resources (e.g., bandwidth).

Chapter 14 deals with *encryption* and *decryption*, the basic goals of which are communication privacy and authentication. Maintaining privacy means preventing unauthorized persons from extracting information (eavesdropping) from the channel. Establishing authentication means preventing unauthorized persons from injecting spurious signals (spoofing) into the channel. In this chapter we highlight the data encryption standard (DES) and the basic ideas regarding a class of encryption systems called *public key cryptosystems*. We also examine the novel scheme of Pretty Good Privacy (PGP) which is an important file-encryption method for sending data via electronic mail.

The final chapter of the book, Chapter 15, deals with fading channels. In it, we address fading that affects mobile systems such as cellular and personal communication systems (PCS). The chapter itemizes the fundamental fading manifestations, types of degradation, and methods to mitigate the degradation. Two particular mitigation techniques are examined: the Viterbi equalizer implemented in the Global System for Mobile Communication (GSM), and the Rake receiver used in CDMA systems.

1.1.3 Basic Digital Communication Nomenclature

The following are some of the basic digital signal nomenclature that frequently appears in digital communication literature:

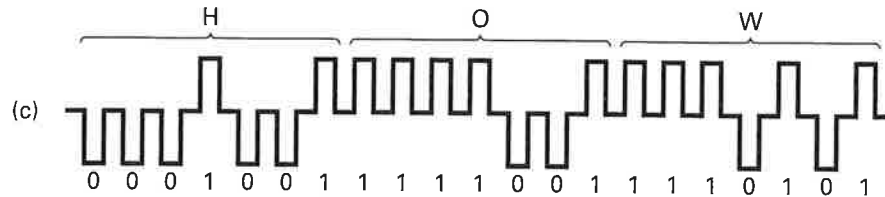
Information source. This is the device producing information to be communicated by means of the DCS. Information sources can be *analog* or *discrete*. The output of an analog source can have any value in a continuous range of amplitudes, whereas the output of a discrete information source takes its value from a finite set. Analog information sources can be transformed into digital sources through the use of *sampling* and *quantization*. Sampling and quantization techniques called formatting and source coding (see Figure 1.3) are described in Chapters 2 and 13.

Textual message. This is a sequence of characters. (See Figure 1.4a.) For digital transmission, the message will be a sequence of digits or symbols from a finite symbol set or alphabet.

Character. A character is a member of an alphabet or set of symbols. (See Figure 1.4b.) Characters may be mapped into a sequence of binary digits.

(a) HOW ARE YOU?
OK
\$9, 567, 216.73

(b) A
9
&



(d) 1 Binary symbol ($k = 1, M = 2$)
10 Quaternary symbol ($k = 2, M = 4$)
011 8-ary symbol ($k = 3, M = 8$)

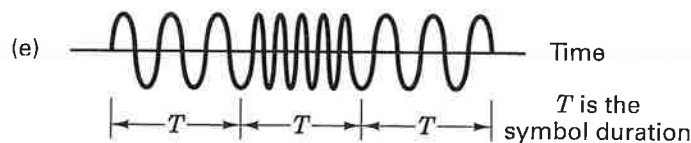
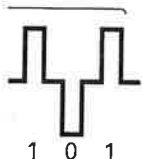


Figure 1.4 Nomenclature examples. (a) Textual messages. (b) Characters. (c) Bit stream (7-bit ASCII). (d) Symbols m_i , $i = 1, \dots, M$, $M = 2^k$. (e) Bandpass digital waveform $s_i(t)$, $i = 1, \dots, M$.

There are several standardized codes used for character encoding, including the American Standard Code for Information Interchange (ASCII), Extended Binary Coded Decimal Interchange Code (EBCDIC), Hollerith, Baudot, Murray, and Morse.

Binary digit (bit). This is the fundamental information unit for all digital systems. The term *bit* also is used as a unit of information content, as described in Chapter 9.

Bit stream. This is a sequence of binary digits (ones and zeros). A bit stream is often termed a *baseband* signal, which implies that its spectral content extends from (or near) dc up to some finite value, usually less than a few megahertz. In Figure 1.4c, the message, "HOW," is represented with the 7-bit ASCII character code, where the bit stream is shown by using a convenient picture of 2-level pulses. The sequence of pulses is drawn using very stylized (ideal-rectangular) shapes with spaces between successive pulses. In a real system, the pulses would never appear as they are depicted here, because such spaces would serve no useful purpose. For a given bit rate, the spaces would increase the bandwidth needed for transmission; or, for a



Char-
 $M = 2^k$

ling, including
(ASCII), Ex-
C), Hollerith,

all digital sys-
s, as described

. A bit stream
spectral content
less than a
nted with the
using a con-
drawn using
on successive
are depicted
given bit rate,
sion; or, for a

ra Chap. 1

given bandwidth, they would increase the time delay needed to receive the message.

Symbol (digital message). A symbol is a group of k bits considered as a unit. We refer to this unit as a *message symbol* m_i ($i = 1, \dots, M$) from a finite symbol set or alphabet. (See Figure 1.4d.) The size of the alphabet, M , is $M = 2^k$, where k is the number of bits in the symbol. For *baseband* transmission, each m_i symbol will be represented by one of a set of baseband pulse waveforms $g_1(t), g_2(t), \dots, g_M(t)$. When transmitting a sequence of such pulses, the unit *Baud* is sometimes used to express pulse rate (symbol rate). For typical *bandpass* transmission, each $g_i(t)$ pulse will then be represented by one of a set of bandpass waveforms $s_1(t), s_2(t), \dots, s_M(t)$. Thus, for wireless systems, the symbol m_i is sent by transmitting the digital waveform $s_i(t)$ for T seconds, the symbol-time duration. The next symbol is sent during the next time interval, T . The fact that the symbol set transmitted by the DCS is finite is a primary difference between a DCS and an analog system. The DCS receiver need only decide which of the M waveforms was transmitted; however, an analog receiver must be capable of accurately estimating a continuous range of waveforms.

Digital waveform. This is a voltage or current waveform (a pulse for baseband transmission, or a sinusoid for bandpass transmission) that represents a digital symbol. The waveform characteristics (amplitude, width, and position for pulses or amplitude, frequency, and phase for sinusoids) allow its identification as one of the symbols in the finite symbol alphabet. Figure 1.4e shows an example of a bandpass digital waveform. Even though the waveform is sinusoidal and consequently has an analog appearance, it is called a *digital waveform* because it is encoded with digital information. In the figure, during each time interval, T , a preassigned frequency indicates the value of a digit.

Data rate. This quantity in bits per second (bits/s) is given by $R = k/T = (1/T) \log_2 M$ bits/s, where k bits identify a symbol from an $M = 2^k$ -symbol alphabet, and T is the k -bit symbol duration.

1.1.4 Digital versus Analog Performance Criteria

A principal difference between analog and digital communication systems has to do with the way in which we evaluate their performance. Analog systems draw their waveforms from a continuum, which therefore forms an infinite set—that is, a receiver must deal with an infinite number of possible waveshapes. The figure of merit for the performance of analog communication systems is a fidelity criterion, such as signal-to-noise ratio, percent distortion, or expected mean-square error between the transmitted and received waveforms.

By contrast, a digital communication system transmits signals that represent digits. These digits form a finite set or alphabet, and the set is known a priori to the receiver. A figure of merit for digital communication systems is the probability of incorrectly detecting a digit, or the probability of error (P_E).

1.2 CLASSIFICATION OF SIGNALS

1.2.1 Deterministic and Random Signals

A signal can be classified as *deterministic*, meaning that there is no uncertainty with respect to its value at any time, or as *random*, meaning that there is some degree of uncertainty before the signal actually occurs. Deterministic signals or waveforms are modeled by explicit mathematical expressions, such as $x(t) = 5 \cos 10t$. For a random waveform it is *not* possible to write such an explicit expression. However, when examined over a long period, a random waveform, also referred to as a *random process*, may exhibit certain regularities that can be described in terms of probabilities and statistical averages. Such a model, in the form of a probabilistic description of the random process, is particularly useful for characterizing signals and noise in communication systems.

1.2.2 Periodic and Nonperiodic Signals

A signal $x(t)$ is called *periodic in time* if there exists a constant $T_0 > 0$ such that

$$x(t) = x(t + T_0) \quad \text{for } -\infty < t < \infty \quad (1.2)$$

where t denotes time. The smallest value of T_0 that satisfies this condition is called the *period* of $x(t)$. The period T_0 defines the duration of one complete cycle of $x(t)$. A signal for which there is no value of T_0 that satisfies Equation (1.2) is called a *nonperiodic signal*.

1.2.3 Analog and Discrete Signals

An *analog signal* $x(t)$ is a continuous function of time; that is, $x(t)$ is uniquely defined for all t . An electrical analog signal arises when a physical waveform (e.g., speech) is converted into an electrical signal by means of a transducer. By comparison, a *discrete signal* $x(kT)$ is one that exists only at discrete times; it is characterized by a sequence of numbers defined for each time, kT , where k is an integer and T is a fixed time interval.

1.2.4 Energy and Power Signals

An electrical signal can be represented as a voltage $v(t)$ or a current $i(t)$ with instantaneous power $p(t)$ across a resistor $\mathcal{R}$ defined by

$$p(t) = \frac{v^2(t)}{\mathcal{R}} \quad (1.3a)$$

or

$$p(t) = i^2(t)\mathcal{R} \quad (1.3b)$$

In communication systems, power is often normalized by assuming $\mathcal{R}$ to be 1Ω , although $\mathcal{R}$ may be another value in the actual circuit. If the actual value of the power is needed, it is obtained by "denormalization" of the normalized value. For the normalized case, Equations 1.3a and 1.3b have the same form. Therefore, regardless of whether the signal is a voltage or current waveform, the normalization convention allows us to express the instantaneous power as

$$p(t) = x^2(t) \quad (1.4)$$

where $x(t)$ is either a voltage or a current signal. The energy dissipated during the time interval $(-T/2, T/2)$ by a real signal with instantaneous power expressed by Equation (1.4) can then be written as

$$E_x^T = \int_{-T/2}^{T/2} x^2(t) dt \quad (1.5)$$

and the average power dissipated by the signal during the interval is

$$P_x^T = \frac{1}{T} E_x^T = \frac{1}{T} \int_{-T/2}^{T/2} x^2(t) dt \quad (1.6)$$

The performance of a communication system depends on the received signal *energy*; higher energy signals are detected more reliably (with fewer errors) than are lower energy signals—the received *energy does the work*. On the other hand, *power* is the *rate* at which energy is delivered. It is important for different reasons. The power determines the voltages that must be applied to a transmitter and the intensities of the electromagnetic fields that one must contend with in radio systems (i.e., fields in waveguides that connect the transmitter to the antenna, and fields around the radiating elements of the antenna).

In analyzing communication signals, it is often desirable to deal with the *waveform energy*. We classify $x(t)$ as an *energy signal* if, and only if, it has nonzero but finite energy ($0 < E_x < \infty$) for all time, where

$$\begin{aligned} E_x &= \lim_{T \rightarrow \infty} \int_{-T/2}^{T/2} x^2(t) dt \\ &= \int_{-\infty}^{\infty} x^2(t) dt \end{aligned} \quad (1.7)$$

In the real world, we always transmit signals having finite energy ($0 < E_x < \infty$). However, in order to describe *periodic signals*, which by definition [Equation (1.2)] exist for all time and thus have infinite energy, and in order to deal with random signals that have infinite energy, it is convenient to define a class of signals called *power signals*. A signal is defined as a power signal if, and only if, it has finite but nonzero power ($0 < P_x < \infty$) for all time, where

$$P_x = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t) dt \quad (1.8)$$

The energy and power classifications are mutually exclusive. An energy signal has finite energy but *zero average power*, whereas a power signal has finite average power but *infinite energy*. A waveform in a system may be constrained in either its power or energy values. As a general rule, periodic signals and random signals are classified as power signals, while signals that are both deterministic and nonperiodic are classified as energy signals [1, 2].

Signal energy and power are both important parameters in specifying a communication system. The classification of a signal as either an energy signal or a power signal is a convenient model to facilitate the mathematical treatment of various signals and noise. In Section 3.1.5, these ideas are developed further, in the context of a digital communication system.

1.2.5 The Unit Impulse Function

A useful function in communication theory is the unit impulse or *Dirac delta function* $\delta(t)$. The impulse function is an abstraction—an infinitely large amplitude pulse, with zero pulse width, and unity weight (area under the pulse), concentrated at the point where its argument is zero. The unit impulse is characterized by the following relationships:

$$\int_{-\infty}^{\infty} \delta(t) dt = 1 \quad (1.9)$$

$$\delta(t) = 0 \quad \text{for } t \neq 0 \quad (1.10)$$

$$\delta(t) \text{ is unbounded at } t = 0 \quad (1.11)$$

$$\int_{-\infty}^{\infty} x(t) \delta(t - t_0) dt = x(t_0) \quad (1.12)$$

The unit impulse function $\delta(t)$ is not a function in the usual sense. When operations involve $\delta(t)$, the convention is to interpret $\delta(t)$ as a unit-area pulse of finite amplitude and nonzero duration, after which the limit is considered as the pulse duration approaches zero. $\delta(t - t_0)$ can be depicted graphically as a spike located at $t = t_0$ with height equal to its integral or area. Thus $A\delta(t - t_0)$ with A constant represents an impulse function whose area or weight is equal to A , that is zero everywhere except at $t = t_0$.

Equation (1.12) is known as the *sifting* or *sampling property* of the unit impulse function; the unit impulse multiplier selects a sample of the function $x(t)$ evaluated at $t = t_0$.

1.3 SPECTRAL DENSITY

The *spectral density* of a signal characterizes the distribution of the signal's energy or power in the frequency domain. This concept is particularly important when considering filtering in communication systems. We need to be able to evaluate the

signal and noise at the filter output. The energy spectral density (ESD) or the power spectral density (PSD) is used in the evaluation.

1.3.1 Energy Spectral Density

The total energy of a real-valued energy signal $x(t)$, defined over the interval, $(-\infty, \infty)$, is described by Equation (1.7). Using Parseval's theorem [1], we can relate the energy of such a signal expressed in the time domain to the energy expressed in the frequency domain, as

$$E_x = \int_{-\infty}^{\infty} x^2(t) dt = \int_{-\infty}^{\infty} |X(f)|^2 df \quad (1.13)$$

where $X(f)$ is the Fourier transform of the nonperiodic signal $x(t)$. (For a review of Fourier techniques, see Appendix A.) Let $\psi_x(f)$ denote the squared magnitude spectrum, defined as

$$\psi_x(f) = |X(f)|^2 \quad (1.14)$$

The quantity $\psi_x(f)$ is the waveform *energy spectral density* (ESD) of the signal $x(t)$. Therefore, from Equation (1.13), we can express the total energy of $x(t)$ by integrating the spectral density with respect to frequency:

$$E_x = \int_{-\infty}^{\infty} \psi_x(f) df \quad (1.15)$$

This equation states that the energy of a signal is equal to the area under the $\psi_x(f)$ versus frequency curve. Energy spectral density describes the signal energy per unit bandwidth measured in joules/hertz. There are equal energy contributions from both **positive and negative frequency components**, since for a real signal, $x(t)$, $|X(f)|$ is an **even function of frequency**. Therefore, the energy spectral density is **symmetrical in frequency about the origin**, and thus the total energy of the signal $x(t)$ can be expressed as

$$E_x = 2 \int_0^{\infty} \psi_x(f) df \quad (1.16)$$

1.3.2 Power Spectral Density

The average power P_x of a real-valued power signal $x(t)$ is defined in Equation (1.8). If $x(t)$ is a *periodic signal* with period T_0 , it is classified as a power signal. The expression for the average power of a periodic signal takes the form of Equation (1.6), where the time average is taken over the signal period T_0 , as follows:

$$P_x = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t) dt \quad (1.17a)$$

Parseval's theorem for a real-valued periodic signal [1] takes the form

$$P_x = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t) dt = \sum_{n=-\infty}^{\infty} |c_n|^2 \quad (1.17b)$$

where the c_n terms are the complex Fourier series coefficients of the periodic signal. (See Appendix A.)

To apply Equation (1.17b), we need only know the magnitude of the coefficients, $|c_n|$. The *power spectral density* (PSD) function $G_x(f)$ of the periodic signal $x(t)$ is a real, even, and nonnegative function of frequency that gives the distribution of the power of $x(t)$ in the frequency domain, defined as

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2 \delta(f - nf_0) \quad (1.18)$$

Equation (1.18) defines the power spectral density of a periodic signal $x(t)$ as a succession of the weighted delta functions. Therefore, the PSD of a periodic signal is a discrete function of frequency. Using the PSD defined in Equation (1.18), we can now write the average normalized power of a real-valued signal as

$$P_x = \int_{-\infty}^{\infty} G_x(f) df = 2 \int_0^{\infty} G_x(f) df \quad (1.19)$$

Equation (1.18) describes the PSD of periodic (power) signals only. If $x(t)$ is a nonperiodic signal it *cannot* be expressed by a Fourier series, and if it is a nonperiodic power signal (having infinite energy) it *may not* have a Fourier transform. However, we may still express the power spectral density of such signals in the *limiting sense*. If we form a *truncated version* $x_T(t)$ of the nonperiodic power signal $x(t)$ by observing it only in the interval $(-T/2, T/2)$, then $x_T(t)$ has finite energy and has a proper Fourier transform $X_T(f)$. It can be shown [2] that the power spectral density of the nonperiodic $x(t)$ can then be defined in the limit as

$$G_x(f) = \lim_{T \rightarrow \infty} \frac{1}{T} |X_T(f)|^2 \quad (1.20)$$

Example 1.1 Average Normalized Power

- Find the average normalized power in the waveform, $x(t) = A \cos 2\pi f_0 t$, using time averaging.
- Repeat part (a) using the summation of spectral coefficients.

Solution

- Using Equation (1.17a), we have

$$\begin{aligned} P_x &= \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} A^2 \cos^2 2\pi f_0 t dt \\ &= \frac{A^2}{2T_0} \int_{-T_0/2}^{T_0/2} (1 + \cos 4\pi f_0 t) dt \\ &= \frac{A^2}{2T_0} (T_0) = \frac{A^2}{2} \end{aligned}$$

(b) Using Equations (1.18) and (1.19) gives us

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2 \delta(f - nf_0)$$

$$\left. \begin{aligned} c_1 &= c_{-1} = \frac{A}{2} \\ c_n &= 0 \quad \text{for } n = 0, \pm 2, \pm 3, \dots \end{aligned} \right\} \quad (\text{see Appendix A})$$

$$G_x(f) = \left(\frac{A}{2}\right)^2 \delta(f - f_0) + \left(\frac{A}{2}\right)^2 \delta(f + f_0)$$

$$P_x = \int_{-\infty}^{\infty} G_x(f) df = \frac{A^2}{2}$$

1.4 AUTOCORRELATION

1.4.1 Autocorrelation of an Energy Signal

Correlation is a matching process; *autocorrelation* refers to the matching of a signal with a delayed version of itself. The autocorrelation function of a real-valued energy signal $x(t)$ is defined as

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t + \tau) dt \quad \text{for } -\infty < \tau < \infty \quad (1.21)$$

The autocorrelation function $R_x(\tau)$ provides a measure of how closely the signal matches a copy of itself as the copy is shifted τ units in time. The variable τ plays the role of a **scanning or searching** parameter. $R_x(\tau)$ is not a function of time; it is only a function of the time difference τ between the waveform and its shifted copy.

The autocorrelative function of a real-valued *energy* signal has the following properties:

1. $R_x(\tau) = R_x(-\tau)$
2. $|R_x(\tau)| \leq R_x(0)$ for all τ
3. $R_x(\tau) \leftrightarrow \Psi_x(f)$

symmetrical in τ about zero
maximum value occurs at the origin
autocorrelation and ESD form a Fourier transform pair, as designated by the double-headed arrows
value at the origin is equal to the energy of the signal

$$4. R_x(0) = \int_{-\infty}^{\infty} x^2(t) dt$$

If items 1 through 3 are satisfied, $R_x(\tau)$ satisfies the properties of an autocorrelation function. Property 4 can be derived from property 3 and thus need not be included as a basic test.

1.4.2 Autocorrelation of a Periodic (Power) Signal

The autocorrelation function of a real-valued power signal $x(t)$ is defined as

$$R_x(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t + \tau) dt \quad \text{for } -\infty < \tau < \infty \quad (1.22)$$

When the power signal $x(t)$ is periodic with period T_0 , the time average in Equation (1.22) may be taken over a *single period* T_0 , and the autocorrelation function can be expressed as

$$R_x(\tau) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t)x(t + \tau) dt \quad \text{for } -\infty < \tau < \infty \quad (1.23)$$

The autocorrelation function of a real-valued *periodic* signal has properties similar to those of an energy signal:

1. $R_x(\tau) = R_x(-\tau)$ symmetrical in τ about zero
2. $|R_x(\tau)| \leq R_x(0)$ for all τ maximum value occurs at the origin
3. $R_x(\tau) \leftrightarrow G_x(f)$ autocorrelation and PSD form a Fourier transform pair
4. $R_x(0) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t) dt$ value at the origin is equal to the average power of the signal

1.5 RANDOM SIGNALS

The main objective of a communication system is the transfer of information over a channel. All useful message signals appear random; that is, the receiver does not know, a priori, which of the possible message waveforms will be transmitted. Also, the noise that accompanies the message signals is due to random electrical signals. Therefore, we need to be able to form efficient descriptions of random signals.

1.5.1 Random Variables

Let a *random variable* $X(A)$ represent the functional relationship between a random event A and a real number. For notational convenience, we shall designate the random variable by X , and let the functional dependence upon A be implicit. The random variable may be discrete or continuous. The *distribution function* $F_X(x)$ of the random variable X is given by

$$F_X(x) = P(X \leq x) \quad (1.24)$$

where $P(X \leq x)$ is the probability that the value taken by the random variable X is less than or equal to a real number x . The distribution function $F_X(x)$ has the following properties:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$
3. $F_X(-\infty) = 0$
4. $F_X(+\infty) = 1$

Another useful function relating to the random variable X is the *probability density function* (pdf), denoted

$$p_X(x) = \frac{dF_X(x)}{dx} \quad (1.25a)$$

As in the case of the distribution function, the pdf is a function of a real number x . The name "density function" arises from the fact that the probability of the event $x_1 \leq X \leq x_2$ equals

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= P(X \leq x_2) - P(X \leq x_1) \\ &= F_X(x_2) - F_X(x_1) \\ &= \int_{x_1}^{x_2} p_X(x) dx \end{aligned} \quad (1.25b)$$

From Equation (1.25b), the probability that a random variable X has a value in some very narrow range between x and $x + \Delta x$ can be approximated as

$$P(x \leq X \leq x + \Delta x) \approx p_X(x) \Delta x \quad (1.25c)$$

Thus, in the limit as Δx approaches zero, we can write

$$P(X = x) = p_X(x) dx \quad (1.25d)$$

The probability density function has the following properties:

1. $p_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} p_X(x) dx = F_X(+\infty) - F_X(-\infty) = 1$.

Thus, a probability density function is always a nonnegative function with a total area of one. Throughout the book we use the designation $p_X(x)$ for the probability density function of a *continuous* random variable. For ease of notation, we will often omit the subscript X and write simply $p(x)$. We will use the designation $p(X = x_i)$ for the probability of a random variable X , where X can take on *discrete* values only.

1.5.1.1 Ensemble Averages

The *mean value* m_X or *expected value* of a random variable X , is defined by

$$m_X = E\{X\} = \int_{-\infty}^{\infty} xp_X(x) dx \quad (1.26)$$

where $\mathbf{E}\{\cdot\}$ is called the *expected value operator*. The n th moment of a probability distribution of a random variable X is defined by

$$\mathbf{E}\{X^n\} = \int_{-\infty}^{\infty} x^n p_X(x) dx \quad (1.27)$$

For the purposes of communication system analysis, the most important moments of X are the first two moments. Thus, $n = 1$ in Equation (1.27) gives m_X as discussed above, whereas $n = 2$ gives the mean-square value of X , as follows:

$$\mathbf{E}\{X^2\} = \int_{-\infty}^{\infty} x^2 p_X(x) dx \quad (1.28)$$

We can also define *central moments*, which are the moments of the difference between X and m_X . The second central moment, called the *variance* of X , is defined as

$$\text{var}(X) = \mathbf{E}\{X - m_X\}^2 = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x) dx \quad (1.29)$$

The variance of X is also denoted as σ_X^2 , and its square root, σ_X , is called the *standard deviation* of X . Variance is a measure of the “randomness” of the random variable X . By specifying the variance of a random variable, we are constraining the width of its probability density function. The variance and the mean-square value are related by

$$\begin{aligned} \sigma_X^2 &= \mathbf{E}\{X^2 - 2m_X X + m_X^2\} \\ &= \mathbf{E}\{X^2\} - 2m_X \mathbf{E}\{X\} + m_X^2 \\ &= \mathbf{E}\{X^2\} - m_X^2 \end{aligned}$$

Thus, the variance is equal to the difference between the mean-square value and the square of the mean.

1.5.2 Random Processes

A random process $X(A, t)$ can be viewed as a function of two variables: *an event* A and *time*. Figure 1.5 illustrates a random process. In the figure there are N *sample functions* of time, $\{X_j(t)\}$. Each of the sample functions can be regarded as the output of a different noise generator. For a specific event A_j , we have a single time function $X(A_j, t) = X_j(t)$ (i.e., a sample function). The totality of all sample functions is called an *ensemble*. For a specific time t_k , $X(A, t_k)$ is a *random variable* $X(t_k)$ whose value depends on the event. Finally, for a specific event, $A = A_j$ and a specific time $t = t_k$, $X(A_j, t_k)$ is simply a *number*. For notational convenience we shall designate the random process by $X(t)$, and let the functional dependence upon A be implicit.

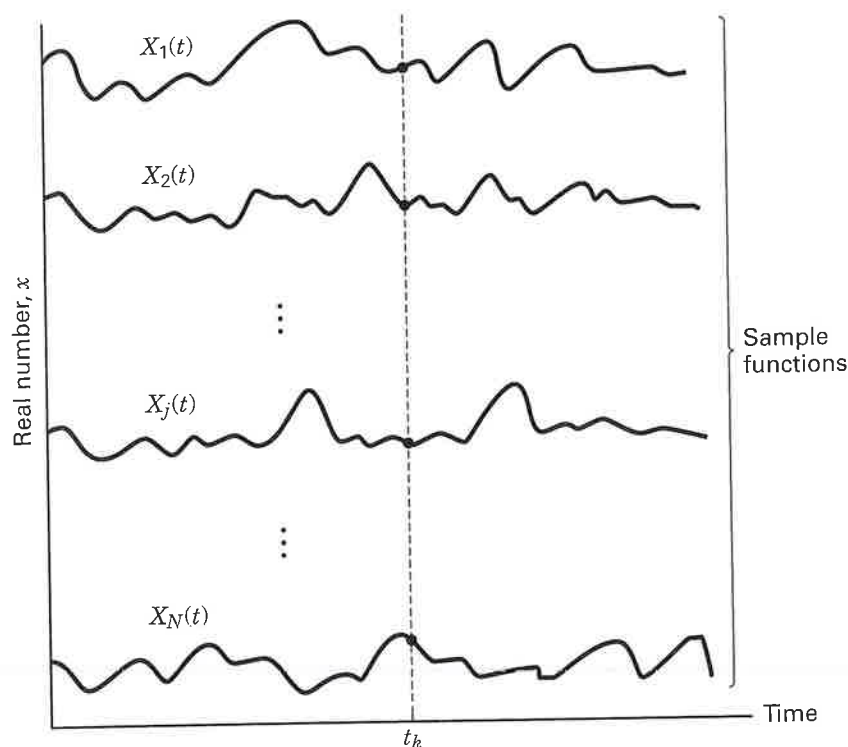


Figure 1.5 Random noise process.

1.5.2.1 Statistical Averages of a Random Process

Because the value of a random process at any future time is unknown (since the identity of the event A is unknown), a random process whose distribution functions are continuous can be described statistically with a probability density function (pdf). In general, the form of the pdf of a random process will be different for different times. In most situations it is not practical to determine empirically the probability distribution of a random process. However, a partial description consisting of the mean and autocorrelation function are often adequate for the needs of communication systems. We define the mean of the random process $X(t)$ as

$$\mathbf{E}\{X(t_k)\} = \int_{-\infty}^{\infty} x p_{X_k}(x) dx = m_X(t_k) \quad (1.30)$$

where $X(t_k)$ is the random variable obtained by observing the random process at time t_k and the pdf of $X(t_k)$, the density over the ensemble of events at time t_k , is designated $p_{X_k}(x)$.

We define the autocorrelation function of the random process $X(t)$ to be a function of two variables, t_1 and t_2 , given by

$$R_X(t_1, t_2) = \mathbf{E}\{X(t_1)X(t_2)\} \quad (1.31)$$

where $X(t_1)$ and $X(t_2)$ are random variables obtained by observing $X(t)$ at times t_1 and t_2 , respectively. The autocorrelation function is a measure of the degree to which two time samples of the same random process are related.

1.5.2.2 Stationarity

A random process $X(t)$ is said to be *stationary* in the *strict sense* if none of its statistics are affected by a shift in the time origin. A random process is said to be *wide-sense stationary* (WSS) if two of its statistics, its mean and autocorrelation function, do not vary with a shift in the time origin. Thus, a process is WSS if

$$\mathbf{E}\{X(t)\} = m_X = \text{a constant} \quad (1.32)$$

and

$$R_X(t_1, t_2) = R_X(t_1 - t_2) \quad (1.33)$$

Strict-sense stationary implies wide-sense stationary, but not vice versa. Most of the useful results in communication theory are predicated on random information signals and noise being wide-sense stationary. From a practical point of view, it is not necessary for a random process to be stationary for all time but only for some observation interval of interest.

For stationary processes, the autocorrelation function in Equation (1.33) does not depend on time but only on the difference between t_1 and t_2 . That is, all pairs of values of $X(t)$ at points in time separated by $\tau = t_1 - t_2$ have the same correlation value. Thus, for stationary systems, we can denote $R_X(t_1, t_2)$ simply as $R_X(\tau)$.

1.5.2.3 Autocorrelation of a Wide-Sense Stationary Random Process

Just as the variance provides a measure of randomness for random variables, the autocorrelation function provides a similar measure for random processes. For a wide-sense stationary process, the autocorrelation function is only a function of the *time difference* $\tau = t_1 - t_2$; that is,

$$R_X(\tau) = \mathbf{E}\{X(t)X(t + \tau)\} \quad \text{for } -\infty < \tau < \infty \quad (1.34)$$

For a zero mean WSS process, $R_X(\tau)$ indicates the extent to which the random values of the process separated by τ seconds in time are statistically correlated. In other words, $R_X(\tau)$ gives us an idea of the frequency response that is associated with a random process. If $R_X(\tau)$ changes slowly as τ increases from zero to some value, it indicates that, on average, sample values of $X(t)$ taken at $t = t_1$ and $t = t_1 + \tau$ are nearly the same. Thus, we would expect a frequency domain representation of $X(t)$ to contain a preponderance of low frequencies. On the other hand, if $R_X(\tau)$ decreases rapidly as τ is increased, we would expect $X(t)$ to change rapidly with time and thereby contain mostly high frequencies.

Properties of the autocorrelation function of a real-valued wide-sense stationary process are as follows:

$X(t)$ at times t_1
the degree to

se if none of its
ss is said to be
autocorrelation
s WSS if

$$(1.32)$$

$$(1.33)$$

versa. Most of
m information
it of view, it is
only for some

ion (1.33) does
t is, all pairs of
ne correlation
: $R_X(\tau)$.

rocess

dom variables,
processes. For
a function of

$$(1.34)$$

which the ran-
stically corre-
sponse that is
ises from zero
en at $t = t_1$ and
rain represen-
other hand, if
hange rapidly

-sense station-

ra Chap. 1

1. $R_X(\tau) = R_X(-\tau)$
2. $|R_X(\tau)| \leq R_X(0)$ for all τ
3. $R_X(\tau) \leftrightarrow G_X(f)$
4. $R_X(0) = \mathbf{E}\{X^2(t)\}$

symmetrical in τ about zero
maximum value occurs at the origin
autocorrelation and power spectral density form a
Fourier transform pair
value at the origin is equal to the average power
of the signal

1.5.3 Time Averaging and Ergodicity

To compute m_X and $R_X(\tau)$ by ensemble averaging, we would have to average across all the sample functions of the process and would need to have complete knowledge of the first- and second-order joint probability density functions. Such knowledge is generally not available.

When a random process belongs to a special class, known as an *ergodic process*, its time averages equal its ensemble averages, and the statistical properties of the process can be determined by *time averaging over a single sample function* of the process. For a random process to be ergodic, it must be stationary in the strict sense. (The converse is not necessary.) However, for communication systems, where we are satisfied to meet the conditions of wide-sense stationarity, we are interested only in the mean and autocorrelation functions.

We can say that a random process is *ergodic in the mean* if

$$m_X = \lim_{T \rightarrow \infty} 1/T \int_{-T/2}^{T/2} X(t) dt \quad (1.35)$$

and it is *ergodic in the autocorrelation function* if

$$R_X(\tau) = \lim_{T \rightarrow \infty} 1/T \int_{-T/2}^{T/2} X(t)X(t + \tau) dt \quad (1.36)$$

Testing for the ergodicity of a random process is usually very difficult. In practice one makes an intuitive judgment as to whether it is reasonable to interchange the time and ensemble averages. A reasonable assumption in the analysis of most communication signals (in the absence of transient effects) is that the random waveforms are ergodic in the mean and the autocorrelation function. Since time averages equal ensemble averages for ergodic processes, fundamental electrical engineering parameters, such as dc value, rms value, and average power can be related to the moments of an ergodic random process. Following is a summary of these relationships:

1. The quantity $m_X = \mathbf{E}\{X(t)\}$ is equal to the dc level of the signal.
2. The quantity m_X^2 is equal to the normalized power in the dc component.
3. The second moment of $X(t)$, $\mathbf{E}\{X^2(t)\}$, is equal to the total average normalized power.

4. The quantity $\sqrt{\mathbf{E}\{X^2(t)\}}$ is equal to the root-mean-square (rms) value of the voltage or current signal.
5. The variance σ_X^2 is equal to the average normalized power in the time-varying or ac component of the signal.
6. If the process has zero mean (i.e., $m_X = m_X^2 = 0$), then $\sigma_X^2 = \mathbf{E}\{X^2\}$ and the variance is the same as the mean-square value, or the variance represents the total power in the normalized load.
7. The standard deviation σ_X is the rms value of the ac component of the signal.
8. If $m_X = 0$, then σ_X is the rms value of the signal.

1.5.4 Power Spectral Density and Autocorrelation of a Random Process

A random process $X(t)$ can generally be classified as a power signal having a power spectral density (PSD) $G_X(f)$ of the form shown in Equation (1.20). $G_X(f)$ is particularly useful in communication systems, because it describes the distribution of a signal's power in the frequency domain. The PSD enables us to evaluate the signal power that will pass through a network having known frequency characteristics. We summarize the principal features of PSD functions as follows:

1. $G_X(f) \geq 0$ and is always real valued
2. $G_X(f) = G_X(-f)$ for $X(t)$ real-valued
3. $G_X(f) \leftrightarrow R_X(\tau)$ PSD and autocorrelation form a Fourier transform pair
4. $P_X = \int_{-\infty}^{\infty} G_X(f) df$ relationship between average normalized power and PSD

In Figure 1.6, we present a visualization of autocorrelation and power spectral density functions. What does the term *correlation* mean? When we inquire about the correlation between two phenomena, we are asking how closely do they correspond in behavior or appearance, how well do they match one another. In mathematics, an autocorrelation function of a signal (in the time domain) describes the correspondence of the signal to itself in the following way. An exact copy of the signal is made and located in time at minus infinity. Then we move the copy an increment in the direction of positive time and ask the question, "How well do these two (the original versus the copy) match"? We move the copy another step in the positive direction and ask, "How well do they match now?" And so forth. The correlation between the two is plotted as a function of time, denoted τ , which can be thought of as a scanning parameter.

Figure 1.6a–d highlights some of these steps. Figure 1.6a illustrates a single sample waveform from a WSS random process, $X(t)$. The waveform is a binary random sequence with unit-amplitude positive and negative (bipolar) pulses. The positive and negative pulses occur with equal probability. The duration of each binary digit is T seconds, and the average or dc value of the random sequence is zero. Figure 1.6b shows the same sequence displaced τ_1 seconds in time; this sequence is therefore denoted $X(t - \tau_1)$. Let us assume that $X(t)$ is ergodic in the autocorrela-

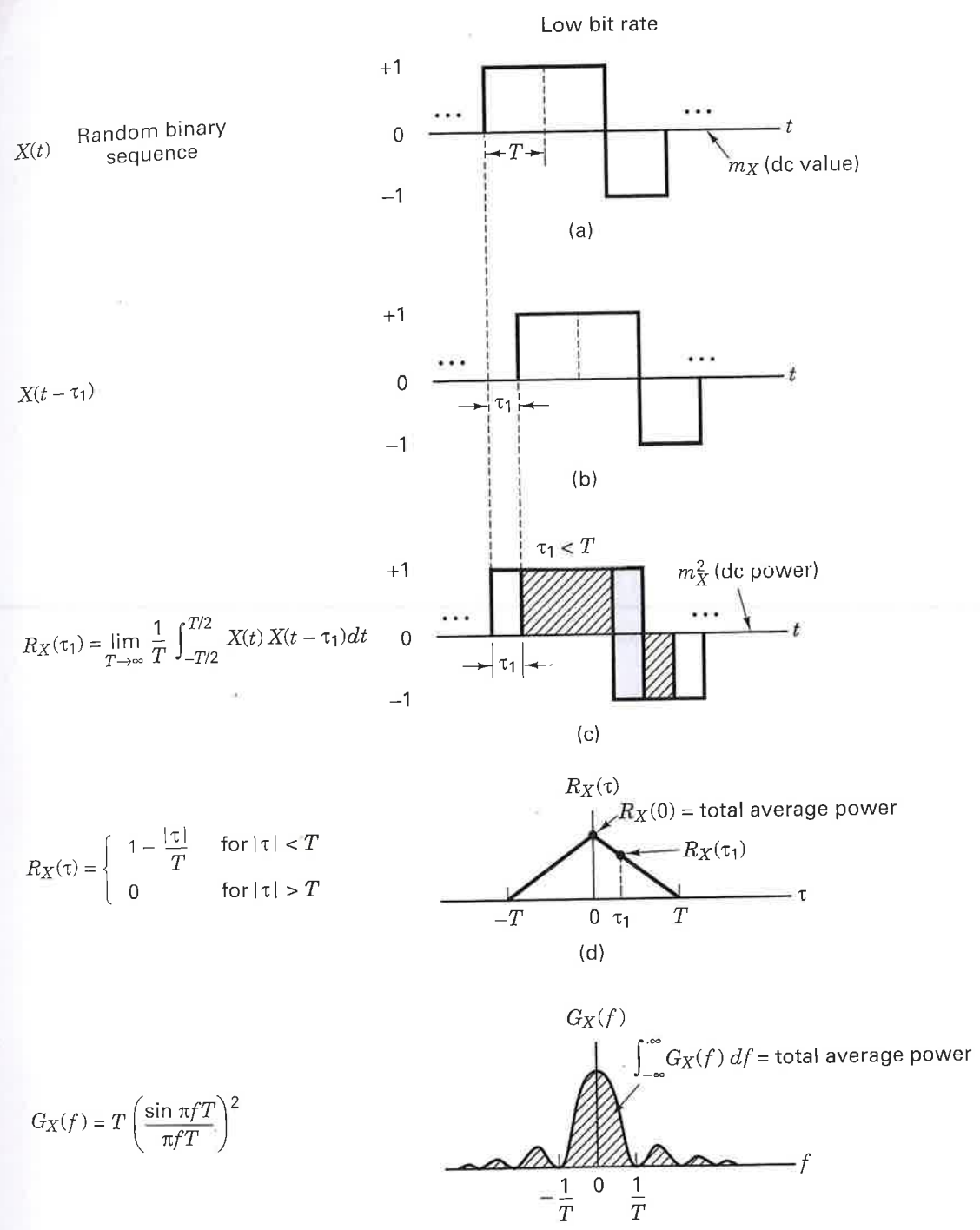
s) value of the
 e time-varying
 $E\{X^2\}$ and the
 represents the
 of the signal.

Process
 aving a power
 $S_X(f)$ is partic-
 tribution of a
 ate the signal
 characteristics.

ansform pair
 ower

power spec-
 n we inquire
 osely do they
 e another. In
 in) describes
 t copy of the
 the copy an
 How well do
 other step in
 so forth. The
 τ , which can

ates a single
 a binary ran-
 es. The posi-
 each binary
 is zero. Fig-
 sequence is
 autocorrela-



$X(t)$ Random binary sequence

$X(t - \tau_1)$

$$R_X(\tau_1) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} X(t) X(t - \tau_1) dt$$

$$R_X(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & \text{for } |\tau| < T \\ 0 & \text{for } |\tau| > T \end{cases}$$

$$G_X(f) = T \left(\frac{\sin \pi f T}{\pi f T} \right)^2$$

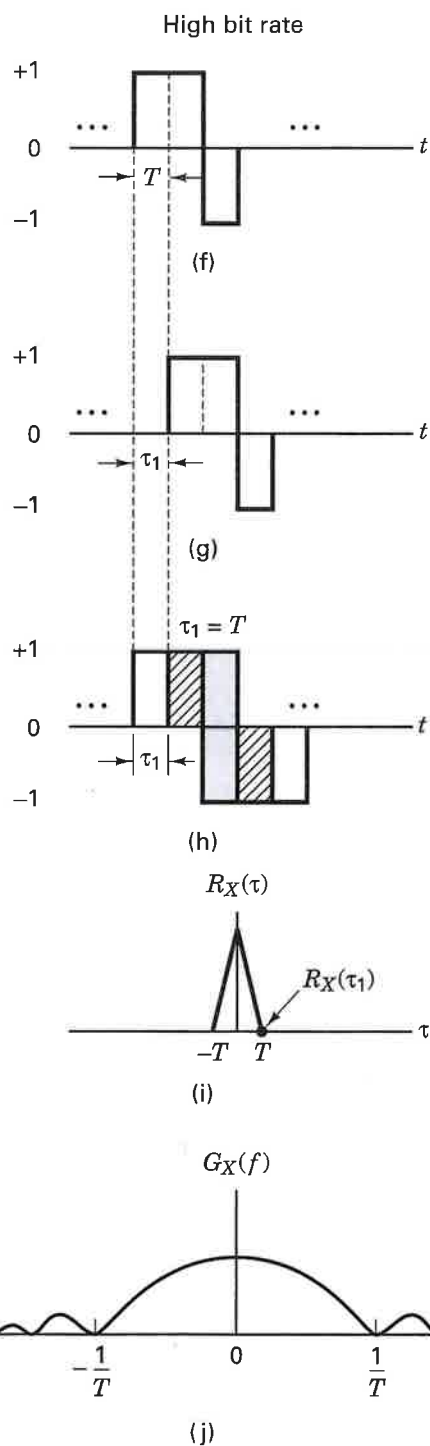


Figure 1.6 continued

tion function so that we can use time averaging instead of ensemble averaging to find $R_X(\tau)$. The value of $R_X(\tau_1)$ is obtained by taking the product of the two sequences $X(t)$ and $X(t - \tau_1)$ and finding the average value using Equation (1.36). Equation (1.36) is accurate for ergodic processes *only in the limit*. However, integration over an integer number of periods can provide us with an estimate of $R_X(\tau)$. Notice that $R_X(\tau_1)$ can be obtained by a positive or negative shift of $X(t)$. Figure 1.6c illustrates such a calculation, using the single sample sequence (Figure 1.6a) and its shifted replica (Figure 1.6b). The cross-hatched areas under the product curve $X(t)X(t - \tau_1)$ contribute to positive values of the product, and the grey areas contribute to negative values. The integration of $X(t)X(t - \tau_1)$ over several pulse times yields a net value of area which is one point, the $R_X(\tau_1)$ point of the $R_X(\tau)$ curve. The sequences can be further shifted by $\tau_2, \tau_3, \dots$, each shift yielding a point on the overall autocorrelation function $R_X(\tau)$ shown in Figure 1.6d. Every random sequence of bipolar pulses has an autocorrelation plot of the general shape shown in Figure 1.6d. The plot peaks at $R_X(0)$ [the best match occurs when τ equals zero, since $R(\tau) \leq R(0)$ for all τ], and it declines as τ increases. Figure 1.6d shows points corresponding to $R_X(0)$ and $R_X(\tau_1)$.

The analytical expression for the autocorrelation function $R_X(\tau)$ shown in Figure 1.6d, is [1]

$$R_X(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & \text{for } |\tau| \leq T \\ 0 & \text{for } |\tau| > T \end{cases} \quad (1.37)$$

Notice that the autocorrelation function gives us frequency information; it tells us something about the bandwidth of the signal. Autocorrelation is a time-domain function; there are no frequency-related terms in the relationship shown in Equation (1.37). How does it give us bandwidth information about the signal? Consider that the signal is a very slowly moving (low bandwidth) signal. As we step the copy along the τ axis, at each step asking the question, "How good is the match between the original and the copy?" the match will be quite good for a while. In other words, the triangular-shaped autocorrelation function in Figure 1.6d and Equation (1.37) will ramp down gradually with τ . But if we have a very rapidly moving (high bandwidth) signal, perhaps a very small shift in τ will result in there being zero correlation. In this case, the autocorrelation function will have a very steep appearance. Therefore, the relative shape of the autocorrelation function tells us something about the bandwidth of the underlying signal. Does it ramp down gently? If so, then we are dealing with a low bandwidth signal. Is the function steep? If so, then we are dealing with a high bandwidth signal.

The autocorrelation function allows us to express a random signal's power spectral density directly. Since the PSD and the autocorrelation function are Fourier transforms of each other, the PSD, $G_X(f)$, of the random bipolar-pulse sequence can be found using Table A.1 as the transform of $R_X(\tau)$ in Equation (1.37). Observe that

$$G_X(f) = T \left(\frac{\sin \pi f T}{\pi f T} \right)^2 = T \operatorname{sinc}^2 fT \quad (1.38)$$

where

$$\operatorname{sinc} y = \frac{\sin \pi y}{\pi y} \quad (1.39)$$

The general shape of $G_X(f)$ is illustrated in Figure 1.6e.

Notice that the area under the PSD curve represents the average power in the signal. One convenient measure of *bandwidth* is the width of the main spectral lobe. (See Section 1.7.2.) Figure 1.6e illustrates that the bandwidth of a signal is inversely related to the symbol duration or pulse width, Figures 1.6f–j repeat the steps shown in Figures 1.6a–e, except that the pulse duration is shorter. Notice that the shape of the shorter pulse duration $R_X(\tau)$ is narrower, shown in Figure 1.6i, than it is for the longer pulse duration $R_X(\tau)$, shown in Figure 1.6d. In Figure 1.6i, $R_X(\tau_1) = 0$; in other words, a shift of τ_1 in the case of the shorter pulse duration example is enough to produce a zero match, or a complete decorrelation between the shifted sequences. Since the pulse duration T is shorter (pulse rate is higher) in Figure 1.6f, than in Figure 1.6a, the bandwidth occupancy in Figure 1.6j is greater than the bandwidth occupancy shown in Figure 1.6e for the lower pulse rate.

1.5.5 Noise in Communication Systems

The term *noise* refers to *unwanted* electrical signals that are always present in electrical systems. The presence of noise superimposed on a signal tends to obscure or mask the signal; it limits the receiver's ability to make correct symbol decisions, and thereby limits the rate of information transmission. Noise arises from a variety of sources, both man made and natural. The *man-made noise* includes such sources as spark-plug ignition noise, switching transients, and other radiating electromagnetic signals. *Natural noise* includes such elements as the atmosphere, the sun, and other galactic sources.

Good engineering design can eliminate much of the noise or its undesirable effect through filtering, shielding, the choice of modulation, and the selection of an optimum receiver site. For example, sensitive radio astronomy measurements are typically located at remote desert locations, far from man-made noise sources. However, there is one natural source of noise, called *thermal* or *Johnson noise*, that cannot be eliminated. Thermal noise [4, 5] is caused by the thermal motion of electrons in all dissipative components—resistors, wires, and so on. The same electrons that are responsible for electrical conduction are also responsible for thermal noise.

We can describe thermal noise as a zero-mean *Gaussian* random process. A Gaussian process $n(t)$ is a random function whose value n at any arbitrary time t is statistically characterized by the Gaussian probability density function

$$p(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{n}{\sigma} \right)^2 \right] \quad (1.40)$$

(1.38)

where σ^2 is the variance of n . The *normalized or standardized Gaussian density function* of a zero-mean process is obtained by assuming that $\sigma = 1$. This normalized pdf is shown sketched in Figure 1.7.

We will often represent a random signal as the sum of a Gaussian noise random variable and a dc signal. That is,

(1.39)

$$z = a + n$$

where z is the random signal, a is the dc component, and n is the Gaussian noise random variable. The pdf $p(z)$ is then expressed as

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{z - a}{\sigma} \right)^2 \right] \quad (1.41)$$

where, as before, σ^2 is the variance of n . The Gaussian distribution is often used as the system noise model because of a theorem, called the *central limit theorem* [3], which states that under very general conditions the probability distribution of the sum of j statistically independent random variables approaches the Gaussian distribution as $j \rightarrow \infty$, no matter what the individual distribution functions may be. Therefore, even though individual noise mechanisms might have other than Gaussian distributions, the aggregate of many such mechanisms will tend toward the Gaussian distribution.

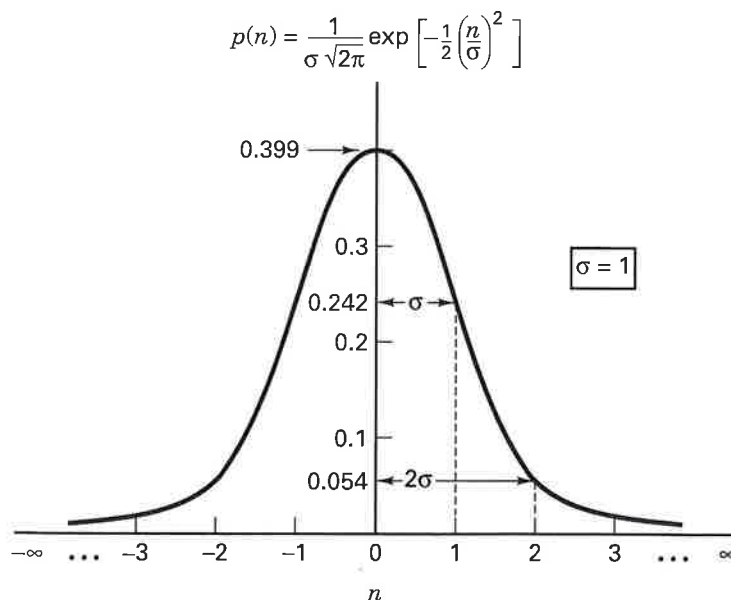


Figure 1.7 Normalized ($\sigma = 1$) Gaussian probability density function.

(1.40)

1.5.5.1 White Noise

The primary spectral characteristic of thermal noise is that its power spectral density is *the same* for all frequencies of interest in most communication systems; in other words, a thermal noise source emanates an equal amount of noise power per unit bandwidth at all frequencies—from dc to about 10^{12} Hz. Therefore, a simple model for thermal noise assumes that its power spectral density $G_n(f)$ is flat for all frequencies, as shown in Figure 1.8a, and is denoted as

$$G_n(f) = \frac{N_0}{2} \quad \text{watts/hertz} \quad (1.42)$$

where the factor of 2 is included to indicate that $G_n(f)$ is a *two-sided* power spectral density. When the noise power has such a uniform spectral density we refer to it as *white noise*. The adjective “white” is used in the same sense as it is with white light, which contains equal amounts of all frequencies within the visible band of electromagnetic radiation.

The autocorrelation function of white noise is given by the inverse Fourier transform of the noise power spectral density (see Table A.1), denoted as follows:

$$R_n(\tau) = \mathcal{F}^{-1}\{G_n(f)\} = \frac{N_0}{2} \delta(\tau) \quad (1.43)$$

Thus the autocorrelation of white noise is a delta function weighted by the factor $N_0/2$ and occurring at $\tau = 0$, as seen in Figure 1.8b. Note that $R_n(\tau)$ is zero for $\tau \neq 0$; that is, any two different samples of white noise, no matter how close together in time they are taken, are uncorrelated.

The average power P_n of white noise is *infinite* because its bandwidth is infinite. This can be seen by combining Equations (1.19) and (1.42) to yield

$$P_n = \int_{-\infty}^{\infty} \frac{N_0}{2} df = \infty \quad (1.44)$$

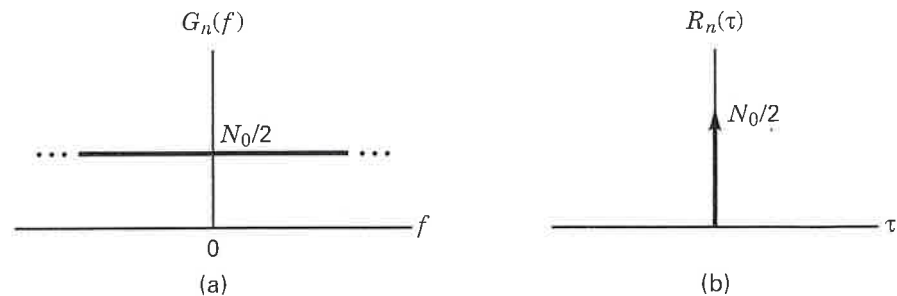


Figure 1.8 (a) Power spectral density of white noise. (b) Autocorrelation function of white noise.

Although white noise is a useful abstraction, no noise process can truly be white; however, the noise encountered in many real systems can be assumed to be approximately white. We can only observe such noise after it has passed through a real system which will have a finite bandwidth. Thus, as long as the bandwidth of the noise is appreciably larger than that of the system, the noise can be considered to have an infinite bandwidth.

The delta function in Equation (1.43) means that the noise signal $n(t)$ is totally decorrelated from its time-shifted version, for any $\tau > 0$. Equation (1.43) indicates that any two different samples of a white noise process are uncorrelated. Since thermal noise is a Gaussian process and the samples are uncorrelated, the noise samples are also independent [3]. Therefore, the effect on the detection process of a channel with *additive white Gaussian noise* (AWGN) is that the noise affects each transmitted symbol *independently*. Such a channel is called a *memoryless channel*. The term “additive” means that the noise is simply superimposed or added to the signal—that there are no multiplicative mechanisms at work.

Since thermal noise is present in all communication systems and is the prominent noise source for most systems, the thermal noise characteristics—additive, white, and Gaussian—are most often used to model the noise in communication systems. Since zero-mean Gaussian noise is completely characterized by its *variance*, this model is particularly simple to use in the detection of signals and in the design of optimum receivers. In this book we shall assume, unless otherwise stated, that the system is corrupted by *additive zero-mean white Gaussian noise*, even though this is sometimes an oversimplification.

1.6 SIGNAL TRANSMISSION THROUGH LINEAR SYSTEMS

Having developed a set of models for signals and noise, we now consider the characterization of systems and their effects on such signals and noise. Since a system can be characterized equally well in the time domain or the frequency domain, techniques will be developed in both domains to analyze the response of a linear system to an arbitrary input signal. The signal, applied to the input of the system, as shown in Figure 1.9, can be described either as a time-domain signal, $x(t)$, or by its Fourier transform, $X(f)$. The use of time-domain analysis yields the time-domain output $y(t)$, and in the process, $h(t)$, the characteristic or *impulse response* of the network will be defined. When the input is considered in the frequency domain, we shall define a *frequency transfer function* $H(f)$ for the system, which will determine the frequency-domain output $Y(f)$. The system is assumed to be linear and time invariant. It is also assumed that there is no stored energy in the system at the time the input is applied.

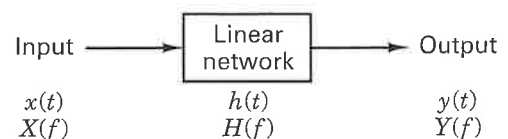


Figure 1.9 Linear system and its key parameters.

1.6.1 Impulse Response

The linear time invariant system or network illustrated in Figure 1.9 is characterized in the time domain by an impulse response $h(t)$, which is the response when the input is equal to a unit impulse $\delta(t)$; that is,

$$h(t) = y(t) \quad \text{when } x(t) = \delta(t) \quad (1.45)$$

Consider the name *impulse response*. That is a very appropriate name for this event. Characterizing a linear system in terms of its impulse response has a straightforward physical interpretation. At the system input, we apply a unit impulse (a nonrealizable signal, having infinite amplitude, zero width, and unit area), as illustrated in Figure 1.10a. Applying such an impulse to the system can be thought of as giving the system "a whack." How does the system respond to such a force (impulse) at the input? The output response $h(t)$ is the system's impulse response. (A possible shape is depicted in Figure 1.10b.)

The response of the network to an arbitrary input signal $x(t)$ is found by the convolution of $x(t)$ with $h(t)$, expressed as

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau) h(t - \tau) d\tau \quad (1.46)$$

where $*$ denotes the convolution operation. (See Section A.5.) The system is assumed to be *causal*, which means that there can be *no* output prior to the time, $t = 0$, when the input is applied. Therefore, the lower limit of integration can be changed to zero, and we can express the output $y(t)$ in either the form

$$y(t) = \int_0^{\infty} x(\tau) h(t - \tau) d\tau \quad (1.47a)$$

or the form

$$y(t) = \int_0^{\infty} x(t - \tau) h(\tau) d\tau \quad (1.47b)$$

Each of the expressions in Equations (1.46) and (1.47) is called the *convolution integral*. Convolution is a basic mathematical tool that plays an important role in understanding all communication systems. Thus, the reader is urged to review

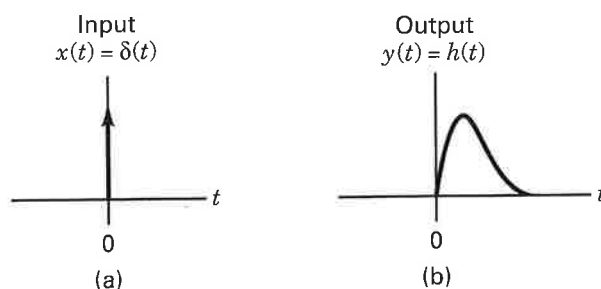


Figure 1.10 (a) Input signal $x(t)$ is a unit impulse function. (b) Output signal $y(t)$ is the system's impulse response $h(t)$.

section A.5, where one can see that Equations (1.46) and (1.47) are the results of a straightforward process.

1.6.2 Frequency Transfer Function

The frequency-domain output signal $Y(f)$ is obtained by taking the Fourier transform of both sides of Equation (1.46). Since convolution in the time domain transforms to multiplication in the frequency domain (and vice versa), Equation (1.46) yields

$$Y(f) = X(f)H(f) \quad (1.48)$$

or

$$H(f) = \frac{Y(f)}{X(f)} \quad (1.49)$$

provided, of course, that $X(f) \neq 0$ for all f . Here $H(f) = \mathcal{F}\{h(t)\}$, the Fourier transform of the impulse response function, is called the *frequency transfer function* or the *frequency response* of the network. In general, $H(f)$ is complex and can be written as

$$H(f) = |H(f)| e^{j\theta(f)} \quad (1.50)$$

where $|H(f)|$ is the magnitude response. The phase response is defined as

$$\theta(f) = \tan^{-1} \frac{\text{Im} \{H(f)\}}{\text{Re} \{H(f)\}} \quad (1.51)$$

where the terms "Re" and "Im" denote "the real part of" and "the imaginary part of," respectively.

The frequency transfer function of a linear time-invariant network can easily be measured in the laboratory with a sinusoidal generator at the input of the network and an oscilloscope at the output. When the input waveform $x(t)$ is expressed as

$$x(t) = A \cos 2\pi f_0 t$$

the output of the network will be

$$y(t) = A |H(f_0)| \cos [2\pi f_0 t + \theta(f_0)] \quad (1.52)$$

The input frequency f_0 is stepped through the values of interest; at each step, the amplitude and phase at the output are measured.

1.6.2.1 Random Processes and Linear Systems

If a random process forms the input to a time-invariant linear system, the output will also be a random process. That is, each sample function of the input process yields a sample function of the output process. The input power spectral density $G_X(f)$ and the output power spectral density $G_Y(f)$ are related as follows:

$$G_Y(f) = G_X(f) |H(f)|^2 \quad (1.53)$$

Equation (1.53) provides a simple way of finding the power spectral density out of a time-invariant linear system when the input is a random process.

In Chapters 3 and 4 we consider the detection of signals in Gaussian noise. We will utilize a fundamental property of a Gaussian process applied to a linear system, as follows. It can be shown that if a Gaussian process $X(t)$ is applied to a time-invariant linear filter, the random process $Y(t)$ developed at the output of the filter is also Gaussian [6].

1.6.3 Distortionless Transmission

What is required of a network for it to behave like an *ideal* transmission line? The output signal from an ideal transmission line may have some time delay compared with the input, and it may have a different amplitude than the input (just a scale change), but otherwise it must have no distortion—it must have the same shape as the input. Therefore, for ideal distortionless transmission, we can describe the output signal as

$$y(t) = Kx(t - t_0) \quad (1.54)$$

where K and t_0 are constants. Taking the Fourier transform of both sides (see Section A.3.1), we write

$$Y(f) = KX(f)e^{-j2\pi ft_0} \quad (1.55)$$

Substituting the expression (1.55) for $Y(f)$ into Equation (1.49), we see that the required system transfer function for distortionless transmission is

$$H(f) = Ke^{-j2\pi ft_0} \quad (1.56)$$

Therefore, to achieve *ideal distortionless transmission*, the overall system response must have a constant magnitude response and its phase shift must be linear with frequency. It is not enough that the system amplify or attenuate all frequency components equally. All of the signal's frequency components must also arrive with identical time delay in order to add up correctly. Since the time delay t_0 is related to the phase shift θ and the radian frequency $\omega = 2\pi f$ by

$$t_0 \text{ (seconds)} = \frac{\theta \text{ (radians)}}{2\pi f \text{ (radians/second)}} \quad (1.57a)$$

it is clear that phase shift must be proportional to frequency in order for the time delay of all components to be identical. A characteristic often used to measure delay distortion of a signal is called *envelope delay* or *group delay*, which is defined as

$$\tau(f) = -\frac{1}{2\pi} \frac{d\theta(f)}{df} \quad (1.57b)$$

(1.53)

density out of

Gaussian noise. lied to a linear is applied to a e output of the

(1.54)

oth sides (see

(1.55)

e see that the

(1.56)

stem response oe linear with equency com- so arrive with t_0 is related to

(1.57a)

r for the time d to measure which is de-

(1.57b)

a Chap. 1

Therefore, for distortionless transmission, an equivalent way of characterizing phase to be a linear function of frequency is to characterize the **envelope delay** $\tau(f)$ as a constant. In practice, a signal will be distorted in passing through some parts of a system. Phase or amplitude correction (*equalization*) networks may be introduced elsewhere in the system to correct for this distortion. It is the overall input-output characteristic of the system that determines its performance.

1.6.3.1 Ideal Filter

One cannot build the ideal network described in Equation (1.56). The problem is that Equation (1.56) implies an infinite bandwidth capability, where the bandwidth of a system is defined as the interval of positive frequencies over which the magnitude $|H(f)|$ remains within a specified value. In Section 1.7 various measures of bandwidth are enumerated. As an approximation to the ideal infinite-bandwidth network, let us choose a truncated network that passes, without distortion, all frequency components between f_ℓ and f_u , where f_ℓ is the lower cutoff frequency and f_u is the upper cutoff frequency, as shown in Figure 1.11. Each of these networks is called an *ideal filter*. Outside the range $f_\ell < f < f_u$, which is called the *passband*, the ideal filter is assumed to have a response of zero magnitude. The effective width of the passband is specified by the filter bandwidth $W_f = (f_u - f_\ell)$ hertz.

When $f_\ell \neq 0$ and $f_u \neq \infty$, the filter is called a *bandpass filter* (BPF), shown in Figure 1.11a. When $f_\ell = 0$ and f_u has a finite value, the filter is called a *low-pass filter* (LPF), shown in Figure 1.11b. When f_ℓ has a nonzero value and when $f_u \rightarrow \infty$, the filter is called a *high-pass filter* (HPF), shown in Figure 1.11c.

Following Equation (1.56) and letting $K = 1$, for the ideal low-pass filter transfer function with bandwidth $W_f = f_u$ hertz, shown in Figure 1.11b, we can write the transfer function as

$$H(f) = |H(f)| e^{-j\theta(f)} \quad (1.58)$$

where

$$|H(f)| = \begin{cases} 1 & \text{for } |f| < f_u \\ 0 & \text{for } |f| \geq f_u \end{cases} \quad (1.59)$$

and

$$e^{-j\theta(f)} = e^{-j2\pi f t_0} \quad (1.60)$$

The impulse response of the ideal low-pass filter, illustrated in Figure 1.12, is

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f) e^{j2\pi f t} df \\ &= \int_{-f_u}^{f_u} e^{-j2\pi f t_0} e^{j2\pi f t} df \end{aligned} \quad (1.61)$$

or

$$\begin{aligned}
 &= \int_{-f_u}^{f_u} e^{j2\pi f(t-t_0)} df \\
 &= 2f_u \frac{\sin 2\pi f_u(t-t_0)}{2\pi f_u(t-t_0)} \\
 &= 2f_u \operatorname{sinc} 2f_u(t-t_0)
 \end{aligned} \tag{1.62}$$

where $\operatorname{sinc} x$ is as defined in Equation (1.39). The impulse response shown in Figure 1.12 is noncausal, which means that it has a nonzero output prior to the

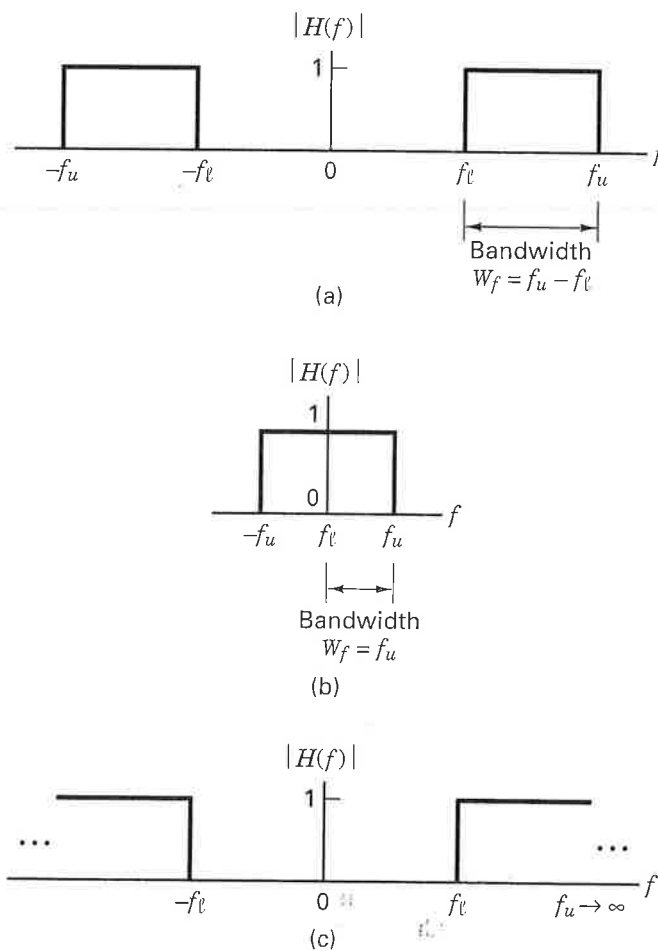


Figure 1.11 Ideal filter transfer function. (a) Ideal bandpass filter. (b) Ideal low-pass filter. (c) Ideal high-pass filter.

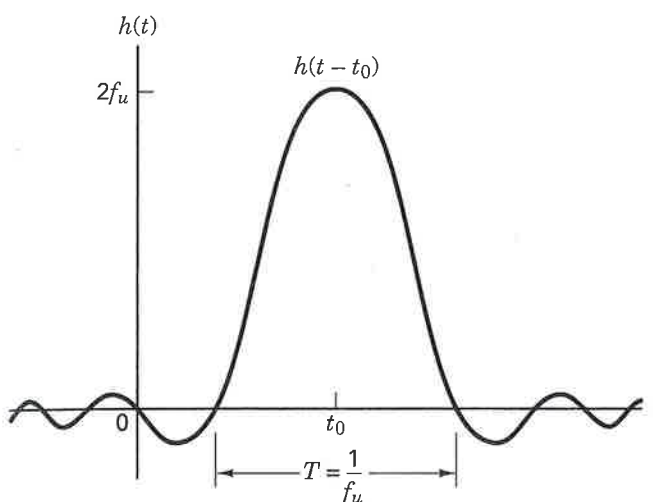


Figure 1.12 Impulse response of the ideal low-pass filter.

application of an input at time $t = 0$. Therefore, it should be clear that the ideal filter described in Equation (1.58) is not realizable.

Example 1.2 Effect of an Ideal Filter on White Noise

White noise with power spectral density $G_n(f) = N_0/2$, shown in Figure 1.8a, forms the input to the ideal low-pass filter shown in Figure 1.11b. Find the power spectral density $G_Y(f)$ and the autocorrelation function $R_Y(\tau)$ of the output signal.

Solution

$$\begin{aligned} G_Y(f) &= G_n(f) |H(f)|^2 \\ &= \begin{cases} \frac{N_0}{2} & \text{for } |f| < f_u \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The autocorrelation is the inverse Fourier transform of the power spectral density and is given by (see Table A.1)

$$\begin{aligned} R_Y(\tau) &= N_0 f_u \frac{\sin 2\pi f_u \tau}{2\pi f_u \tau} \\ &= N_0 f_u \text{sinc } 2f_u \tau \end{aligned}$$

Comparing this result with Equation (1.62), we see that $R_Y(\tau)$ has the same shape as the impulse response of the ideal low-pass filter shown in Figure 1.12. In this example the ideal low-pass filter transforms the autocorrelation function of white noise (defined by the delta function) into a sinc function. After filtering, we no longer have white noise. The output noise signal will have zero correlation with shifted copies of itself, only at shifts of $\tau = n/2f_u$, where n is any integer other than zero.

1.6.3.2 Realizable Filters

The very simplest example of a realizable low-pass filter is made up of resistance ($\mathcal{R}$) and capacitance (C), as shown in Figure 1.13a; it is called an $\mathcal{RC}$ filter, and its transfer function can be expressed as [7]

$$H(f) = \frac{1}{1 + j2\pi f\mathcal{R}C} = \frac{1}{\sqrt{1 + (2\pi f\mathcal{R}C)^2}} e^{-j\theta(f)} \quad (1.63)$$

where $\theta(f) = \tan^{-1} 2\pi f\mathcal{R}C$. The magnitude characteristic $|H(f)|$ and the phase characteristic $\theta(f)$ are plotted in Figures 1.13b and c, respectively. The low-pass filter bandwidth is defined to be its half-power point; this point is the frequency at which the output signal power has fallen to one-half of its peak value, or the frequency at which the magnitude of the output voltage has fallen to $1/\sqrt{2}$ of its peak value.

The half-power point is generally expressed in decibel (dB) units as the -3 -dB point, or the point that is 3 dB down from the peak, where the decibel is defined as the ratio of two amounts of power, P_1 and P_2 , existing at two points. By definition,

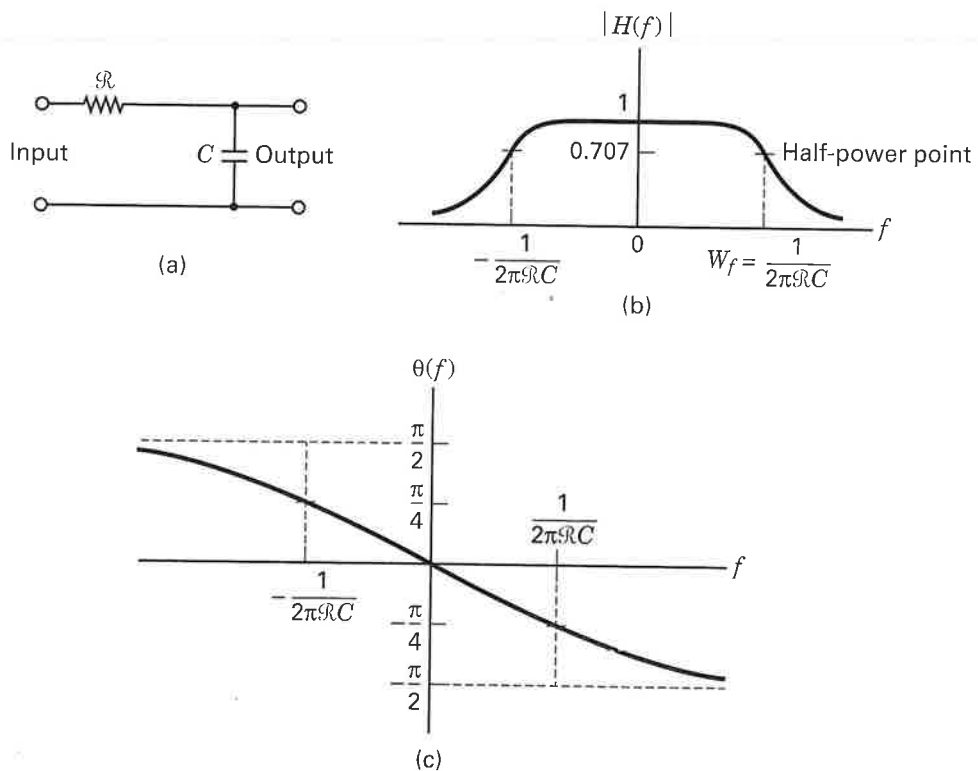


Figure 1.13 $\mathcal{RC}$ filter and its transfer function. (a) $\mathcal{RC}$ filter. (b) Magnitude characteristic of the $\mathcal{RC}$ filter. (c) Phase characteristic of the $\mathcal{RC}$ filter.

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} = 10 \log_{10} \frac{V_2^2/\mathcal{R}_2}{V_1^2/\mathcal{R}_1} \quad (1.64a)$$

where V_1 and V_2 are voltages and $\mathcal{R}_1$ and $\mathcal{R}_2$ are resistances. For communication systems, *normalized power* is generally used for analysis; in this case, $\mathcal{R}_1$ and $\mathcal{R}_2$ are set equal to 1 Ω , so that

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} = 10 \log_{10} \frac{V_2^2}{V_1^2} \quad (1.64b)$$

The amplitude response can be expressed in decibels by

$$|H(f)|_{\text{dB}} = 20 \log_{10} \frac{V_2}{V_1} = 20 \log_{10} |H(f)| \quad (1.64c)$$

where V_1 and V_2 are the input and output voltages, respectively, and where the input and output resistances have been assumed equal.

From Equation (1.63) it is easy to verify that the half-power point of the low-pass $\mathcal{RC}$ filter corresponds to $\omega = 1/\mathcal{RC}$ radians per second or $f = 1/(2\pi\mathcal{RC})$ hertz. Thus the bandwidth W_f in hertz is $1/(2\pi\mathcal{RC})$. The filter *shape factor* is a measure of how well a realizable filter approximates the ideal filter. It is typically defined as the ratio of the filter bandwidths at the -60 -dB and -6 -dB amplitude response points. A sharp-cutoff bandpass filter can be made with a shape factor as low as about 2. By comparison, the shape factor of the simple $\mathcal{RC}$ low-pass filter is almost 600.

There are several useful approximations to the ideal low-pass filter characteristic. One of these, the *Butterworth filter*, approximates the ideal low-pass filter with the function

$$|H_n(f)| = \frac{1}{\sqrt{1 + (f/f_u)^{2n}}} \quad n \geq 1 \quad (1.65)$$

where f_u is the upper -3 -dB cutoff frequency and n is referred to as the *order* of the filter. The higher the order, the greater will be the complexity and the cost to implement the filter. The magnitude function, $|H(f)|$, is sketched (single sided) for several values of n in Figure 1.14. Note that as n gets larger, the magnitude characteristics approach that of the ideal filter. Butterworth filters are popular because they are the best approximation to the ideal, in the sense of *maximal flatness* in the filter passband.

Example 1.3 Effect of an $\mathcal{RC}$ Filter on White Noise

White noise with spectral density $G_n(f) = N_0/2$, shown in Figure 1.8a, forms the input to the $\mathcal{RC}$ filter shown in Figure 1.13a. Find the power spectral density $G_Y(f)$ and the autocorrelation function $R_Y(\tau)$ of the output signal.

Solution

$$\begin{aligned} G_Y(f) &= G_n(f) |H(f)|^2 \\ &= \frac{N_0}{2} \frac{1}{1 + (2\pi f \mathcal{RC})^2} \\ R_Y(\tau) &= \mathcal{F}^{-1}\{G_Y(f)\} \end{aligned}$$

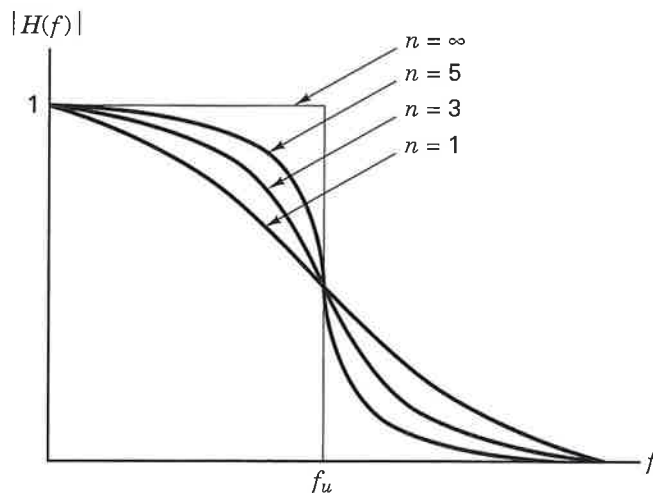


Figure 1.14 Butterworth filter magnitude response.

Using Table A.1, we find that the inverse Fourier transform of $G_Y(f)$ is

$$R_Y(\tau) = \frac{N_0}{4RC} \exp\left(-\frac{|\tau|}{RC}\right)$$

As might have been predicted, we no longer have white noise after filtering. The RC filter transforms the input autocorrelation function of white noise (defined by the delta function) into an exponential function. For a narrowband filter (a large RC product), the output noise will exhibit higher correlation between noise samples of a fixed time shift than will the output noise from a wideband filter.

1.6.4 Signals, Circuits, and Spectra

Signals have been described in terms of their spectra. Similarly, networks or circuits have been described in terms of their spectral characteristics or frequency transfer functions. How is a signal's bandwidth affected as a result of the signal passing through a filter circuit? Figure 1.15 illustrates two cases of interest. In Figure 1.15a (case 1), the input signal has a narrowband spectrum, and the filter transfer function is a wideband function. From Equation (1.48), we see that the output signal spectrum is simply the product of these two spectra. In Figure 1.15a we can verify that multiplication of the two spectral functions will result in a spectrum with a bandwidth approximately equal to the smaller of the two bandwidths (when one of the two spectral functions goes to zero, the multiplication yields zero). Therefore, for case 1, the output signal spectrum is constrained by the input signal spectrum alone. Similarly, we see that for case 2, in Figure 1.15b, where the input signal is a wideband signal but the filter has a narrowband transfer function, the bandwidth of the output signal is constrained by the filter bandwidth; the output signal will be a filtered (distorted) rendition of the input signal.

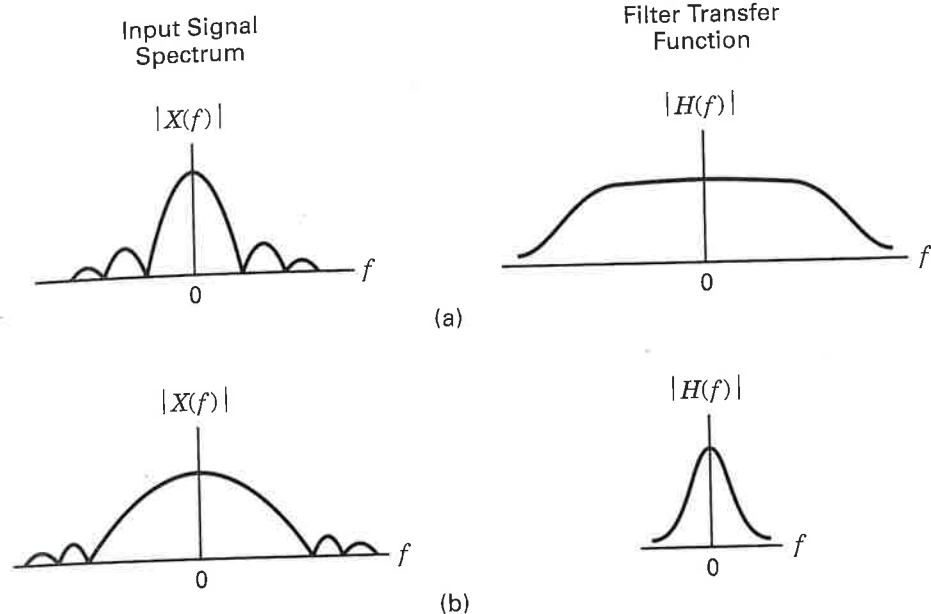


Figure 1.15 Spectral characteristics of the input signal and the circuit contribute to the spectral characteristics of the output signal. (a) Case 1: Output bandwidth is constrained by input signal bandwidth. (b) Case 2: Output bandwidth is constrained by filter bandwidth.

The effect of a filter on a waveform can also be viewed in the time domain. The output $y(t)$ resulting from convolving an ideal input pulse $x(t)$ (having amplitude V_m and pulse width T) with the impulse response of a low-pass RC filter can be written as [8]

$$y(t) = \begin{cases} V_m(1 - e^{-t/RC}) & \text{for } 0 \leq t \leq T \\ V'_m e^{-(t-T)/RC} & \text{for } t > T \end{cases} \quad (1.66)$$

where

$$V'_m = V_m(1 - e^{-T/RC}) \quad (1.67)$$

Let us define the pulse bandwidth as

$$W_p = \frac{1}{T} \quad (1.68)$$

and the RC filter bandwidth, as

$$W_f = \frac{1}{2\pi RC} \quad (1.69)$$

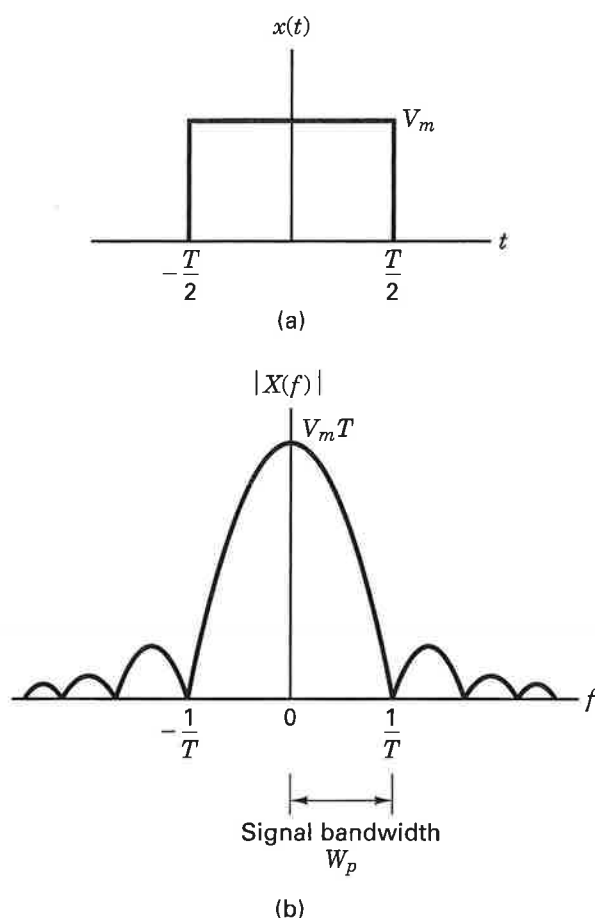


Figure 1.16 (a) Ideal pulse. (b) Magnitude spectrum of the ideal pulse.

The ideal input pulse $x(t)$ and its magnitude spectrum $|X(f)|$ are shown in Figure 1.16. The $\mathcal{RC}$ filter and its magnitude characteristic $|H(f)|$ are shown in Figures 1.13a and b, respectively. Following Equations (1.66) to (1.69), three cases are illustrated in Figure 1.17. Example 1 illustrates the case where $W_p \ll W_f$. Notice that the output response $y(t)$ is a reasonably good approximation of the input pulse $x(t)$, shown in dashed lines. This represents an example of *good fidelity*. In example 2, where $W_p \approx W_f$, we can still recognize that a pulse had been transmitted from the output $y(t)$. Finally, example 3 illustrates the case in which $W_p \gg W_f$. Here the presence of the pulse is barely perceptible from $y(t)$. Can you think of an application where the large filter bandwidth or good fidelity of example 1 is called for? A *precise ranging application*, perhaps, where the pulse time of arrival translates into distance, necessitates a pulse with a steep rise time. Which example characterizes the binary digital communications application? *It is example 2*. As we pointed out earlier regarding Figure 1.1, one of the principal features of binary digital communications is that each received pulse need only be accurately *perceived* as being in one of its two states; a high-fidelity signal need not be maintained. Example 3 has

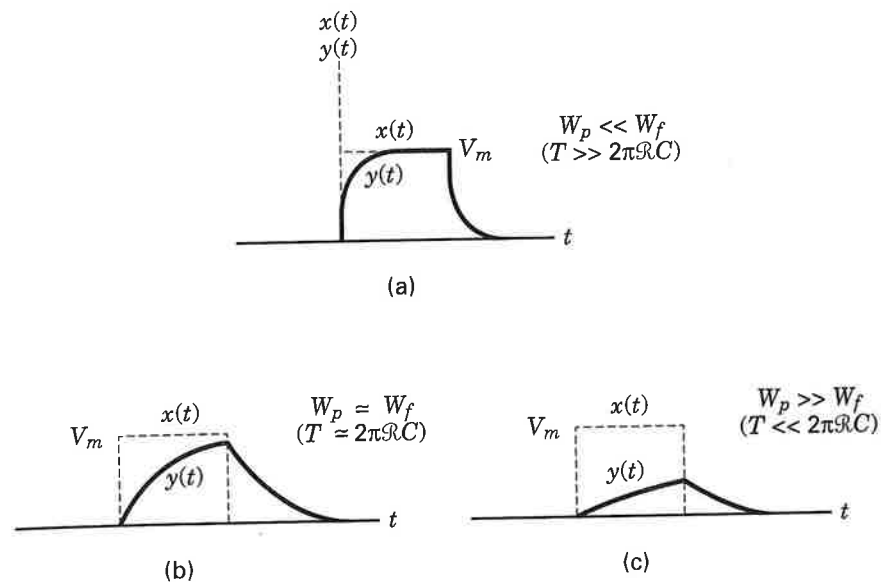


Figure 1.17 Three examples of filtering an ideal pulse. (a) Example 1: Good-fidelity output. (b) Example 2: Good-recognition output. (c) Example 3: Poor-recognition output.

been included for completeness; it would not be used as a design criterion for a practical system.

1.7 BANDWIDTH OF DIGITAL DATA

1.7.1 Baseband versus Bandpass

An easy way to translate the spectrum of a low-pass or baseband signal $x(t)$ to a higher frequency is to multiply or *heterodyne* the baseband signal with a carrier wave $\cos 2\pi f_c t$, as shown in Figure 1.18. The resulting waveform, $x_c(t)$, is called a *double-sideband (DSB) modulated signal* and is expressed as

$$x_c(t) = x(t) \cos 2\pi f_c t \quad (1.70)$$

From the frequency shifting theorem (see Section A.3.2), the spectrum of the DSB signal $x_c(t)$ is given by

$$X_c(f) = \frac{1}{2} [X(f - f_c) + X(f + f_c)] \quad (1.71)$$

The magnitude spectrum $|X(f)|$ of the baseband signal $x(t)$ having a bandwidth f_m and the magnitude spectrum $|X_c(f)|$ of the DSB signal $x_c(t)$ having a bandwidth W_{DSB} are shown in Figure 1.18b and c, respectively. In the plot of $|X_c(f)|$, spectral

pulse. (b) Magni-eal pulse.

shown in Figure own in Figures e cases are illus- W_f . Notice that input pulse $x(t)$, y . In example 2, mitted from the $\gg W_f$. Here the k of an applica- is called for? A il translates into le characterizes we pointed out digital commu- ived as being in . Example 3 has

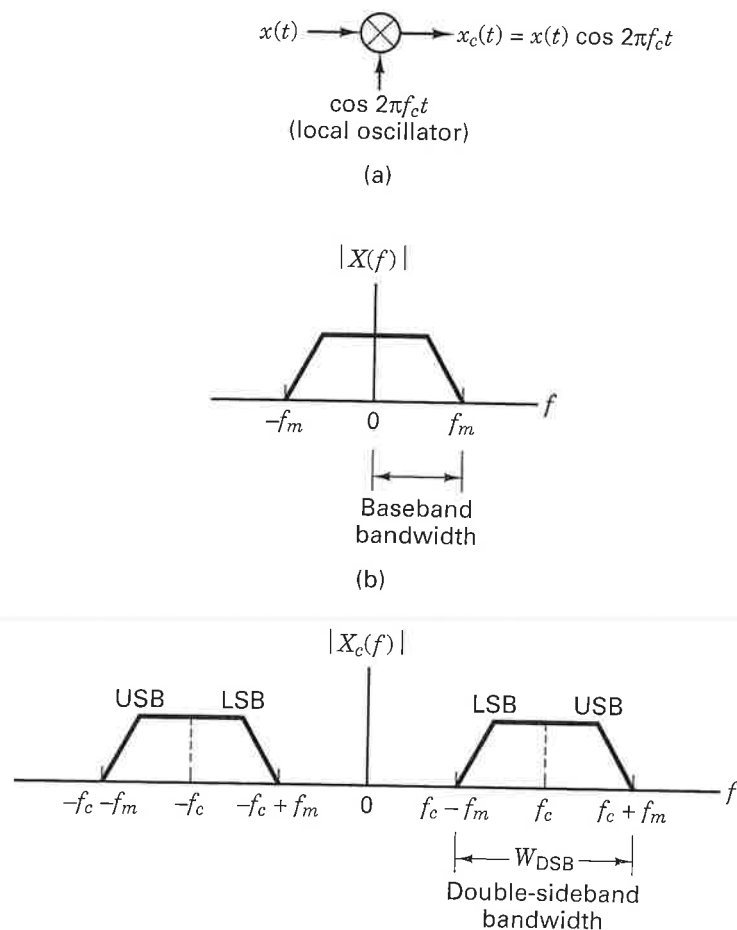


Figure 1.18 Comparison of baseband and double-sideband spectra. (a) Heterodyning. (b) Baseband spectrum. (c) Double-sideband spectrum.

components corresponding to positive baseband frequencies appear in the range f_c to $(f_c + f_m)$. This part of the DSB spectrum is called the *upper sideband* (USB). Spectral components corresponding to negative baseband frequencies appear in the range $(f_c - f_m)$ to f_c . This part of the DSB spectrum is called the *lower sideband* (LSB). Mirror images of the USB and LSB spectra appear in the negative-frequency half of the plot. The *carrier wave* is sometimes referred to as a *local oscillator* (LO) signal, a *mixing signal*, or a *heterodyne signal*. Generally, the carrier wave frequency is much higher than the bandwidth of the baseband signal; that is,

$$f_c \gg f_m$$

From Figure 1.18, we can readily compare the bandwidth f_m required to transmit the baseband signal with the bandwidth W_{DSB} required to transmit the DSB signal; we see that

$$W_{\text{DSB}} = 2f_m \quad (1.72)$$

That is, we need twice as much transmission bandwidth to transmit a DSB version of the signal than we do to transmit its baseband counterpart.

1.7.2 The Bandwidth Dilemma

Many important theorems of communication and information theory are based on the assumption of *strictly bandlimited* channels, which means that no signal power whatever is allowed outside the defined band. We are faced with the dilemma that strictly bandlimited signals, as depicted by the spectrum $|X_1(f)|$ in Figure 1.19b, are not realizable, because they imply signals with infinite duration, as seen by $x_1(t)$ in Figure 1.19a (the inverse Fourier transform of $X_1(f)$). Duration-limited signals, as seen by $x_2(t)$ in Figure 1.19c, can clearly be realized. However, such signals are just as unreasonable, since their Fourier transforms contain energy at arbitrarily high frequencies as depicted by the spectrum $|X_2(f)|$ in Figure 1.19d. In summary, for all bandlimited spectra, the waveforms are not realizable, and for all realizable waveforms, the absolute bandwidth is infinite. The mathematical description of a real signal does not permit the signal to be strictly duration limited and strictly bandlimited. Hence, the mathematical models are abstractions; it is no wonder that there is no single universal definition of bandwidth.

All bandwidth criteria have in common the attempt to specify a measure of the width, W , of a nonnegative real-valued spectral density defined for all frequen-

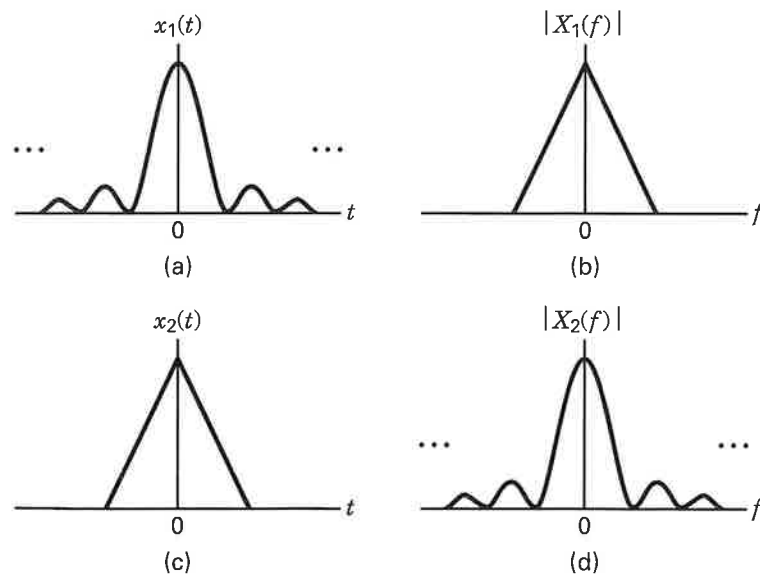


Figure 1.19 (a) Strictly bandlimited signal in the time domain. (b) In the frequency domain. (c) Strictly time limited signal in the time domain. (d) In the frequency domain.

cies $|f| < \infty$. Figure 1.20 illustrates some of the most common definitions of bandwidth; in general, the various criteria are not interchangeable. The single-sided power spectral density for a single heterodyned pulse $x_c(t)$ takes the analytical form

$$G_x(f) = T \left[\frac{\sin \pi(f - f_c)T}{\pi(f - f_c)T} \right]^2 \quad (1.73)$$

where f_c is the carrier wave frequency and T is the pulse duration. This power spectral density, whose general appearance is sketched in Figure 1.20, also characterizes a *random pulse sequence*, assuming that the averaging time is long relative to the pulse duration. The plot consists of a main lobe and smaller symmetrical sidelobes. The general shape of the plot is valid for most digital modulation formats; some formats, however, do not have well-defined lobes. The bandwidth criteria depicted in Figure 1.20 are as follows:

- (a) *Half-power bandwidth*. This is the interval between frequencies at which $G_x(f)$ has dropped to half-power, or 3 dB below the peak value.
- (b) *Equivalent rectangular or noise equivalent bandwidth*, W_N . This bandwidth is defined by the relationship $W_N = P_x / G_x(f_c)$, where P_x is the total signal power over all frequencies and $G_x(f_c)$ is the maximum value (assumed at the band center) of $G_x(f)$. W_N is the bandwidth of a fictitious (ideal rectangular) filter,

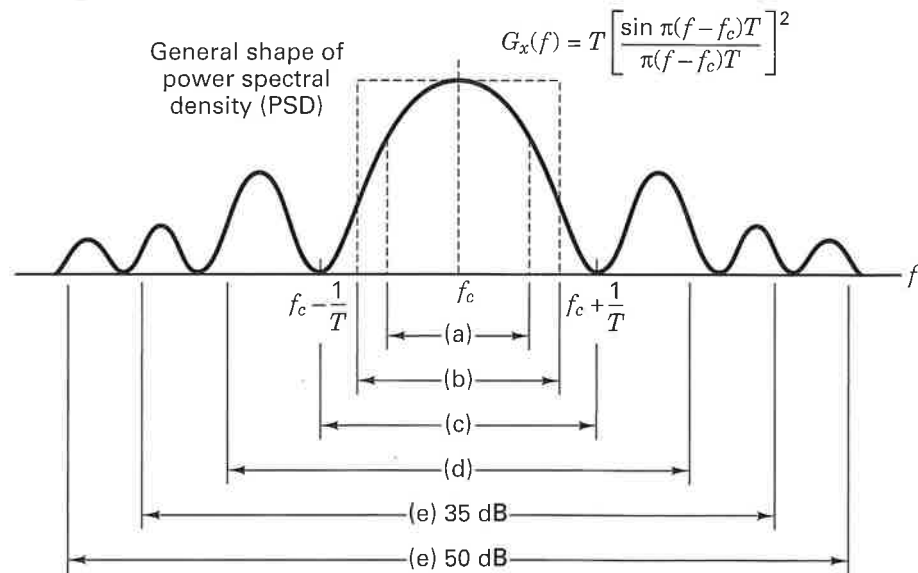


Figure 1.20 Bandwidth of digital data. (a) Half-power. (b) Noise equivalent. (c) Null to null. (d) 99% of power. (e) Bounded PSD (defines attenuation outside bandwidth) at 35 and 50 dB.

definitions of band-
The single-sided
he analytical form

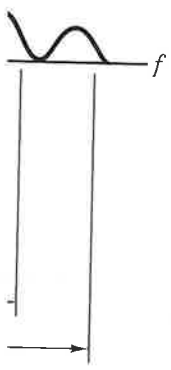
(1.73)

This power spec-
also characterizes
ng relative to the
netrical sidelobes.
on formats; some
criteria depicted

uencies at which
ue.

This bandwidth is
otal signal power
med at the band
ectangular) filter,

$\left. \begin{matrix} \\ \end{matrix} \right]^2$



e equiv-
s atten-

with the same band-center gain as an actual system, that would pass as much white-noise power as the actual system. The concept of W_N facilitates describing or comparing practical linear systems by using idealized equivalents.

- (c) *Null-to-null bandwidth.* The most popular measure of bandwidth for digital communications is the width of the main spectral lobe, where most of the signal power is contained. This criterion lacks complete generality since some modulation formats lack well-defined lobes.
- (d) *Fractional power containment bandwidth.* This bandwidth criterion has been adopted by the Federal Communications Commission (FCC Rules and Regulations Section 2.202) and states that the occupied bandwidth is the band that leaves exactly 0.5% of the signal power above the upper band limit and exactly 0.5% of the signal power below the lower band limit. Thus 99% of the signal power is inside the occupied band.
- (e) *Bounded power spectral density.* A popular method of specifying bandwidth is to state that everywhere outside the specified band, $G_x(f)$ must have fallen at least to a certain stated level below that found at the band center. Typical attenuation levels might be 35 or 50 dB.
- (f) *Absolute bandwidth.* This is the interval between frequencies, outside of which the spectrum is zero. This is a useful abstraction. However, for all realizable waveforms, the absolute bandwidth is infinite.

Example 1.4 Strictly Bandlimited Signals

The concept of a signal that is strictly limited to a band of frequencies is not realizable. Prove this by showing that a *strictly bandlimited* signal must also be a signal of *infinite time duration*.

Solution

Let $x(t)$ be a signal with Fourier transform $X(f)$ that is strictly limited to the band of frequencies centered at $\pm f_c$ and of width $2W$. We may express $X(f)$ in terms of an ideal filter transfer function $H(f)$, illustrated in Figure 1.21a, as

$$X(f) = X'(f)H(f) \tag{1.74}$$

where $X'(f)$ is the Fourier transform of a signal $x'(t)$, not necessarily bandlimited, and

$$H(f) = \text{rect}\left(\frac{f - f_c}{2W}\right) + \text{rect}\left(\frac{f + f_c}{2W}\right) \tag{1.75}$$

in which

$$\text{rect}\left(\frac{f}{2W}\right) = \begin{cases} 1 & \text{for } -W < f < W \\ 0 & \text{for } |f| > W \end{cases}$$

We can express $X(f)$ in terms of $X'(f)$ as

$$X(f) = \begin{cases} X'(f) & \text{for } (f_c - W) \leq |f| \leq (f_c + W) \\ 0 & \text{otherwise} \end{cases}$$

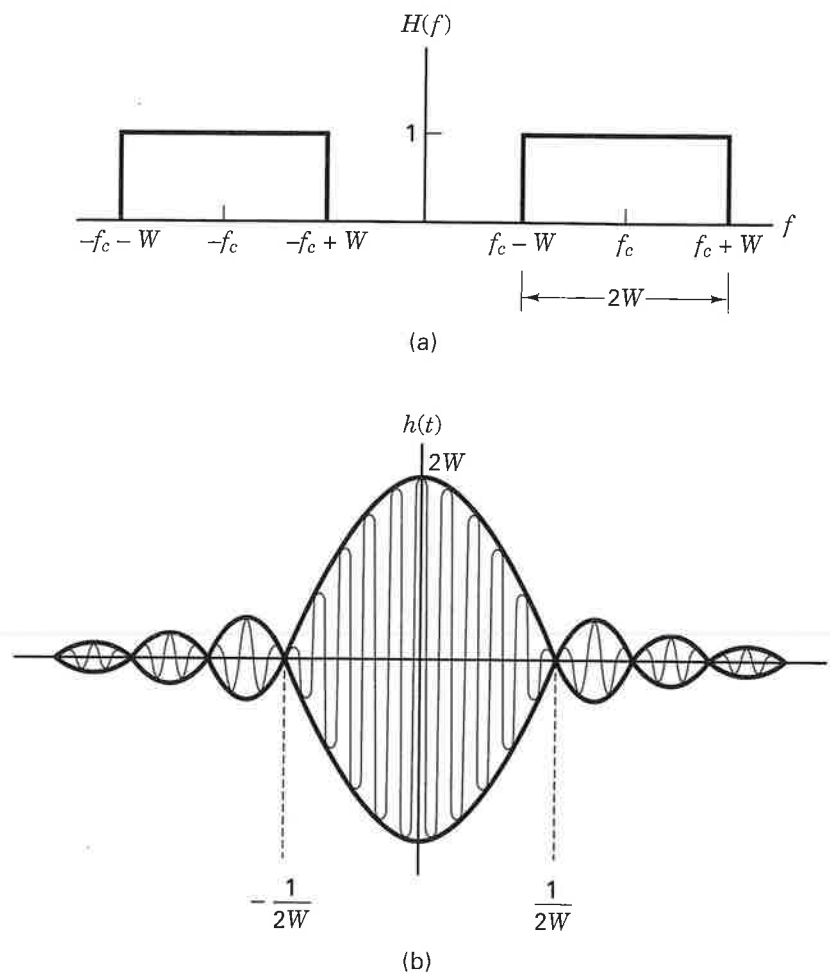


Figure 1.21 Transfer function and impulse response for a strictly bandlimited signal. (a) Ideal bandpass filter. (b) Ideal bandpass impulse response.

Multiplication in the frequency domain, as seen in Equation (1.74), transforms to convolution in the time domain as

$$x(t) = x'(t) * h(t) \quad (1.76)$$

where $h(t)$, the inverse Fourier transform of $H(f)$, can be written as (see Tables A.1 and A.2)

$$h(t) = 2W (\text{sinc } 2Wt) \cos 2\pi f_c t$$

and is illustrated in Figure 1.21b. We note that $h(t)$ is of *infinite duration*. It follows, therefore, that $x(t)$ obtained in Equation (1.76) by convolving $x'(t)$ with $h(t)$ is also of infinite duration and therefore is *not realizable*.

1.8 CONCLUSION

In this chapter, the goals of the book have been outlined and the basic nomenclature has been defined. The fundamental concepts of time-varying signals, such as classification, spectral density, and autocorrelation, have been reviewed. Also, random signals have been considered, and white Gaussian noise, the primary noise model in most communication systems, has been characterized, statistically and spectrally. Finally, we have treated the important area of signal transmission through linear systems and have examined some of the realizable approximations to the ideal case. We have also established that the concept of an absolute bandwidth is an abstraction, and that in the real world we are faced with the need to choose a definition of bandwidth that is useful for our particular application. In the remainder of the book, each of the signal processing steps introduced in this chapter will be explored in the context of the typical system block diagram appearing at the beginning of each chapter.

REFERENCES

1. Haykin, S., *Communication Systems*, John Wiley & Sons, Inc., New York, 1983.
2. Shanmugam, K. S., *Digital and Analog Communication Systems*, John Wiley & Sons, Inc., New York, 1979.
3. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Company, New York, 1965.
4. Johnson, J. B., "Thermal Agitation of Electricity in Conductors," *Phys. Rev.*, vol. 32, July 1928, pp. 97-109.
5. Nyquist, H., "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, vol. 32, July 1928, pp. 110-113.
6. Van Trees, H. L., *Detection, Estimation, and Modulation Theory*, Part 1, John Wiley & Sons, New York, 1968.
7. Schwartz, M., *Information Transmission, Modulation, and Noise*, McGraw-Hill Book Company, New York, 1970.
8. Millman, J., and Taub, H., *Pulse, Digital, and Switching Waveforms*, McGraw-Hill Book Company, New York, 1965.

PROBLEMS

- 1.1. Classify the following signals as energy signals or power signals. Find the normalized energy or normalized power of each.

(a) $x(t) = A \cos 2\pi f_0 t$ for $-\infty < t < \infty$

(b) $x(t) = \begin{cases} A \cos 2\pi f_0 t & \text{for } -T_0/2 \leq t \leq T_0/2, \text{ where } T_0 = 1/f_0 \\ 0 & \text{elsewhere} \end{cases}$

(c) $x(t) = \begin{cases} A \exp(-at) & \text{for } t > 0, a > 0 \\ 0 & \text{elsewhere} \end{cases}$

(d) $x(t) = \cos t + 5 \cos 2t$ for $-\infty < t < \infty$

- 1.2. Determine the energy spectral density of a square pulse $x(t) = \text{rect}(t/T)$, where $\text{rect}(t/T)$ equals 1, for $-T/2 \leq t \leq T/2$, and equals 0, elsewhere. Calculate the normalized energy E_x in the pulse.
- 1.3. Find an expression for the average normalized power in a periodic signal in terms of its complex Fourier series coefficients.
- 1.4. Using time averaging, find the average normalized power in the waveform $x(t) = 10 \cos 10t + 20 \cos 20t$.
- 1.5. Repeat Problem 1.4 using the summation of spectral coefficients.
- 1.6. Determine which, if any, of the following functions have the properties of autocorrelation functions. Justify your determination. [Note: $\mathcal{F}\{R(\tau)\}$ must be a nonnegative function. Why?]
- $x(\tau) = \begin{cases} 1 & \text{for } -1 \leq \tau \leq 1 \\ 0 & \text{otherwise} \end{cases}$
 - $x(\tau) = \delta(\tau) + \sin 2\pi f_0 \tau$
 - $x(\tau) = \exp(|\tau|)$
 - $x(\tau) = 1 - |\tau|$ for $-1 \leq \tau \leq 1$, 0 elsewhere
- 1.7. Determine which, if any, of the following functions have the properties of power spectral density functions. Justify your determination.
- $X(f) = \delta(f) + \cos^2 2\pi f$
 - $X(f) = 10 + \delta(f - 10)$
 - $X(f) = \exp(-2\pi |f - 10|)$
 - $X(f) = \exp[-2\pi(f^2 - 10)]$
- 1.8. Find the autocorrelation function of $x(t) = A \cos(2\pi f_0 t + \phi)$ in terms of its period, $T_0 = 1/f_0$. Find the average normalized power of $x(t)$, using $P_x = R(0)$.
- 1.9. (a) Use the results of Problem 1.8 to find the autocorrelation function $R(\tau)$ of waveform $x(t) = 10 \cos 10t + 20 \cos 20t$.
 (b) Use the relationship $P_x R(0)$ to find the average normalized power in $x(t)$. Compare the answer with the answers to Problems 1.4 and 1.5.
- 1.10. For the function $x(t) = 1 + \cos 2\pi f_0 t$, calculate (a) the average value of $x(t)$; (b) the average power of $x(t)$; (c) the rms value of $x(t)$.
- 1.11. Consider a random process given by $X(t) = A \cos(2\pi f_0 t + \phi)$, where A and f_0 are constants and ϕ is a random variable that is uniformly distributed over $(0, 2\pi)$. If $X(t)$ is an ergodic process, the time averages of $X(t)$ in the limit as $t \rightarrow \infty$ are equal to the corresponding ensemble averages of $X(t)$.
 (a) Use time averaging over an integer number of periods to calculate the approximations to the first and second moments of $X(t)$.
 (b) Use Equations (1.26) and (1.28) to calculate the ensemble-average approximations to the first and second moments of $X(t)$. Compare the results with your answers in part (a).
- 1.12. The Fourier transform of a signal $x(t)$ is defined by $X(f) = \text{sinc } f$, where the sinc function is as defined in Equation (1.39). Find the autocorrelation function, $R_x(\tau)$, of the signal $x(t)$.
- 1.13. Use the sampling property of the unit impulse function to evaluate the following integrals.

- (a) $\int_{-\infty}^{\infty} \cos 6t \delta(t - 3) dt$
- (b) $\int_{-\infty}^{\infty} 10\delta(t)(1 + t)^{-1} dt$
- (c) $\int_{-\infty}^{\infty} \delta(t + 4)(t^2 + 6t + 1) dt$
- (d) $\int_{-\infty}^{\infty} \exp(-t^2)\delta(t - 2) dt$

1.14. Find $X_1(f) * X_2(f)$ for the spectra shown in Figure P1.1.

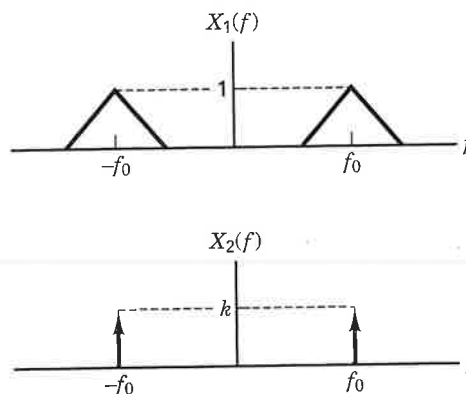


Figure P1.1

1.15. The two-sided power spectral density, $G_x(f) = 10^{-6} f^2$, of a waveform $x(t)$ is shown in Figure P1.2.

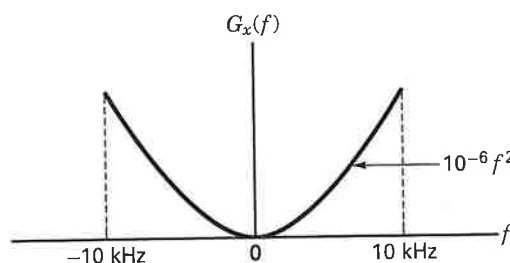


Figure P1.2

- (a) Find the normalized average power in $x(t)$ over the frequency band from 0 to 10 kHz.
- (b) Find the normalized average power contained in the frequency band from 5 to 6 kHz.

- 1.16. Decibels are logarithmic measures of *power ratios*, as described in Equation (1.64a). Sometimes, a similar formulation is used to express nonpower measurements in decibels (referenced to some designated unit). As an example, calculate how many decibels of hamburger meat you would buy to feed 2 hamburgers each to a group of 100 people. Assume that you and the butcher have agreed on the unit of “ $\frac{1}{2}$ pound of meat” (the amount in one hamburger) as a reference unit.
- 1.17. Consider the Butterworth low-pass amplitude response given in Equation (1.65).
 (a) Find the value of n so that $|H(f)|^2$ is constant to within ± 1 dB over the range $|f| \leq 0.9f_u$.
 (b) Show that as n approaches infinity, the amplitude response approaches that of an ideal low-pass filter.
- 1.18. Consider the network in Figure 1.9, whose frequency transfer function is $H(f)$. An impulse $\delta(t)$ is applied at the input. Show that the response $y(t)$ at the output is the inverse Fourier transform of $H(f)$.
- 1.19. An example of a *holding circuit*, commonly used in pulse systems, is shown in Figure P1.3. Determine the impulse response of this circuit.
- 1.20. Given the spectrum

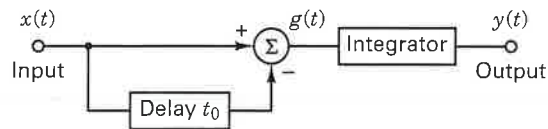


Figure P1.3

$$G_x(f) = 10^{-4} \left\{ \frac{\sin [\pi(f - 10^6)10^{-4}]}{\pi(f - 10^6)10^{-4}} \right\}^2$$

find the value of the signal bandwidth using the following bandwidth definitions:

- (a) Half-power bandwidth.
 (b) Noise equivalent bandwidth.
 (c) Null-to-null bandwidth.
 (d) 99% of power bandwidth. (Hint: Use numerical methods.)
 (e) Bandwidth beyond which the attenuation is 35 dB.
 (f) Absolute bandwidth.

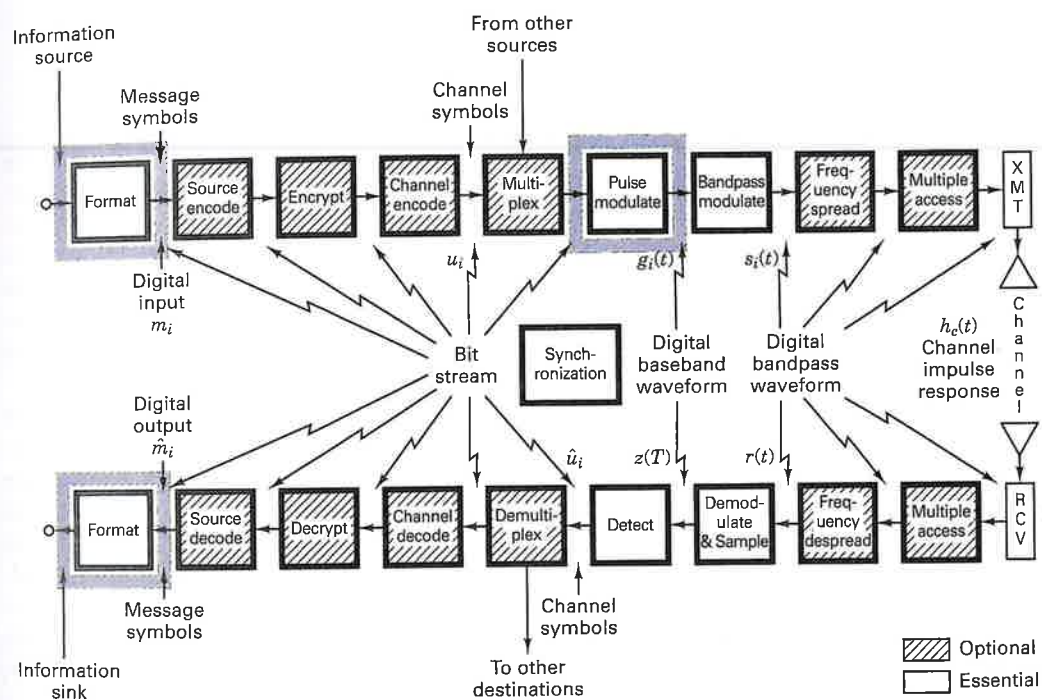
QUESTIONS

- 1.1. How does the plot of a signal's autocorrelation function reveal its bandwidth occupancy? (See Section 1.5.4.)
- 1.2. What two requirements must be fulfilled in order to insure distortionless transmission through a linear system? (See Section 1.6.3.)
- 1.3. Define the parameter *envelope delay* or *group delay*. (See Section 1.6.3.)
- 1.4. What mathematical dilemma is the cause for there being several different definitions of bandwidth? (See Section 1.7.2.)

EXERCISES

Using the Companion CD, run the exercises associated with Chapter 1.

Formatting and Baseband Modulation



The goal of the first essential signal processing step, *formatting*, is to insure that the message (or source signal) is compatible with digital processing. *Transmit formatting* is a transformation from source information to digital symbols. (It is the reverse transformation in the receive chain.) When data compression in addition to formatting is employed, the process is termed *source coding*. Some authors consider formatting a special case of source coding. We treat formatting (and baseband modulation) in this chapter, and treat source coding as a special case of the *efficient description* of source information in Chapter 13.

In Figure 2.1, the highlighted block labeled “formatting” contains a list of topics that deal with transforming information to digital messages. The digital messages are considered to be in the logical format of binary ones and zeros until they are transformed by the next essential step, called pulse modulation, into *baseband* (pulse) waveforms. Such waveforms can then be transmitted over a cable. In Figure 2.1, the highlighted block labeled “baseband signaling” contains a list of pulse modulating waveforms that are described in this chapter. The term baseband refers to a signal whose spectrum extends from (or near) dc up to some finite value, usually less than a few megahertz. In Chapter 3, the subject of baseband signaling is continued with emphasis on demodulation and detection.

2.1 BASEBAND SYSTEMS

In Figure 1.2 we presented a block diagram of a typical digital communication system. A version of this functional diagram, focusing primarily on the formatting and transmission of *baseband* signals, is shown in Figure 2.2. Data already in a digital

to insure that the *Transmit format*ols. (It is the reon in addition to : Some authors atting (and base-ecial case of the

contains a list of The digital mesl zeros until they n, into *baseband* cable. In Figure ist of pulse mod-eband refers to a te value, usually gnaling is contin-

mmunication sys-e formatting and eady in a digital

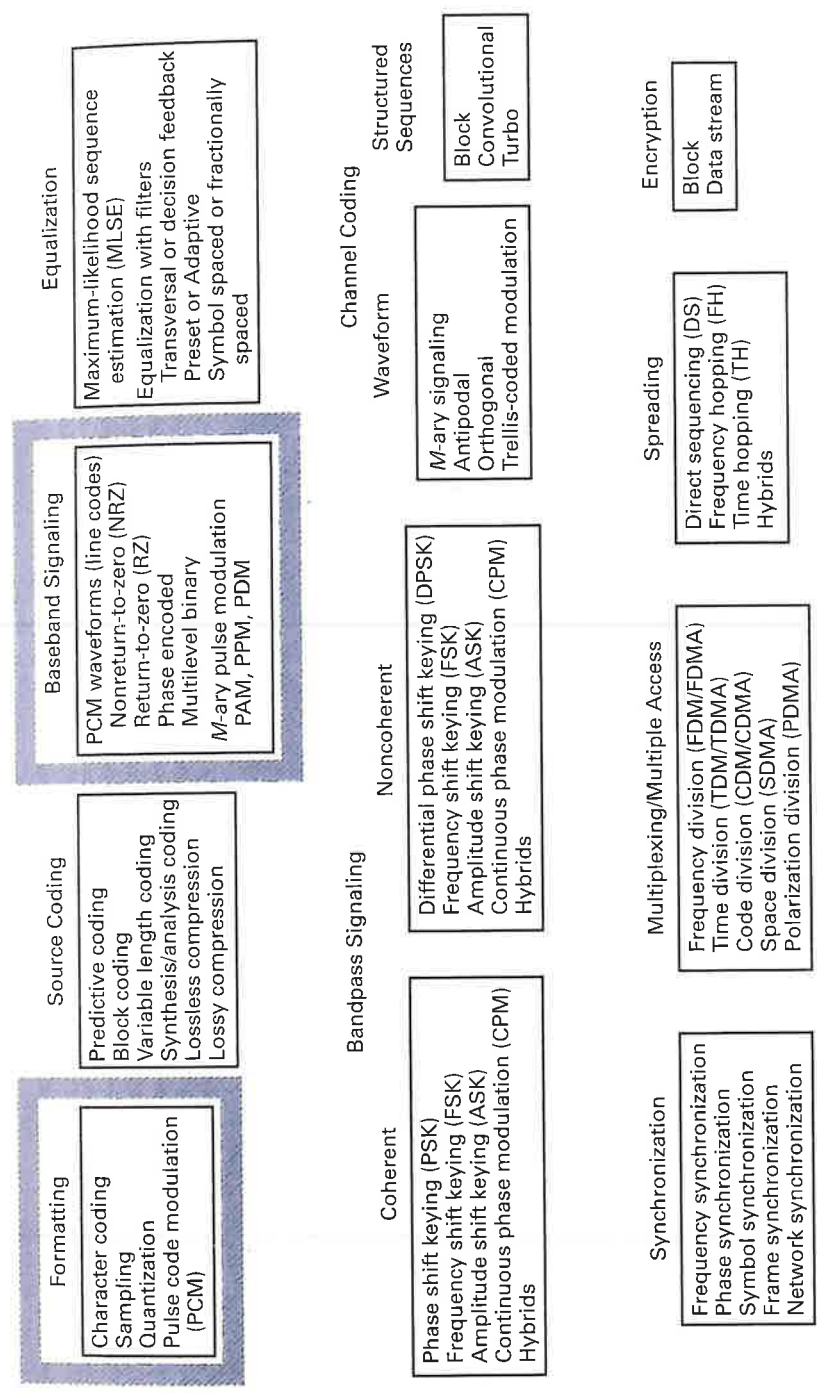


Figure 2.1 Basic digital communication transformations

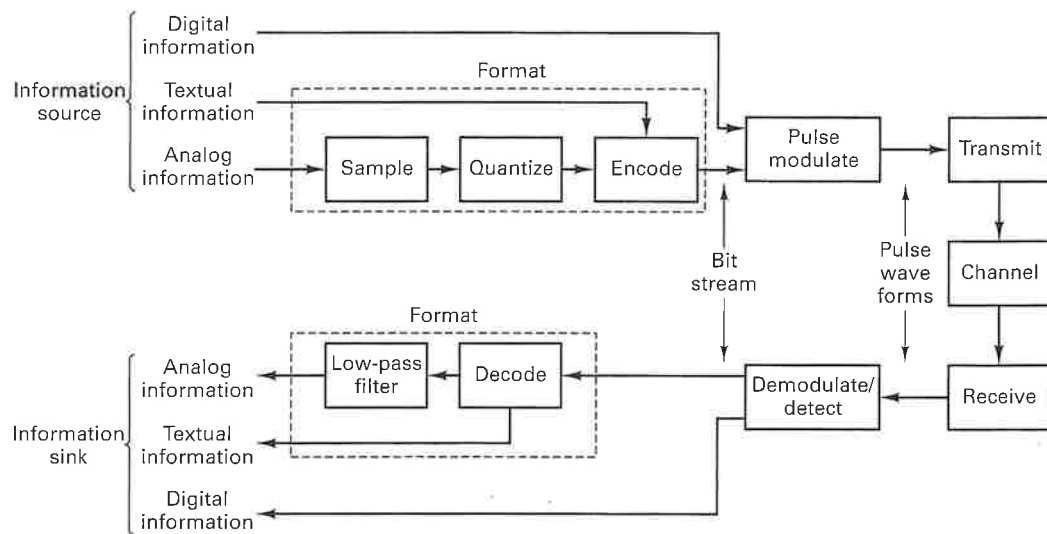


Figure 2.2 Formatting and transmission of baseband signals.

format would bypass the formatting function. Textual information is transformed into binary digits by use of a coder. Analog information is formatted using three separate processes: sampling, quantization, and coding. In all cases, the formatting step results in a sequence of binary digits.

These digits are to be transmitted through a *baseband channel*, such as a pair of wires or a coaxial cable. However, no channel can be used for the transmission of binary digits without first transforming the digits to *waveforms* that are compatible with the channel. For baseband channels, compatible waveforms are pulses.

In Figure 2.2, the conversion from a bit stream to a sequence of pulse waveforms takes place in the block labeled pulse modulate. The output of the modulator is typically a sequence of pulses with characteristics that correspond to the digits being sent. After transmission through the channel, the pulse waveforms are recovered (demodulated) and detected to produce an estimate of the transmitted digits; the final step, (reverse) formatting, recovers an estimate of the source information.

2.2 FORMATTING TEXTUAL DATA (CHARACTER CODING)

The original form of most communicated data (except for computer-to-computer transmissions) is either textual or analog. If the data consist of alphanumeric text, they will be character encoded with one of several standard formats; examples include the American Standard Code for Information Interchange (ASCII), the Extended Binary Coded Decimal Interchange Code (EBCDIC), Baudot, and Hollerith. The textual material is thereby transformed into a digital format. The ASCII format is shown in Figure 2.3; the EBCDIC format is shown in Figure 2.4.

Figure 2.4 EBCDIC character code set.

Figure 2.4 EBCDIC character code set.

The bit numbers signify the order of serial transmission, where bit number 1 is the first signaling element. Character coding, then, is the step that transforms text into binary digits (bits). Sometimes existing character codes are modified to meet specialized needs. For example, the 7-bit ASCII code (Figure 2.3) can be modified to include an added bit for error detection purposes. (See Chapter 6.) On the other hand, sometimes the code is truncated to a 6-bit ASCII version, which provides capability for only 64 characters instead of the 128 characters allowed by 7-bit ASCII.

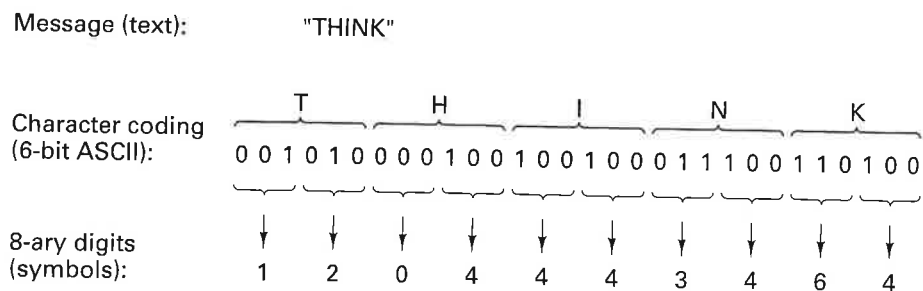
2.3 MESSAGES, CHARACTERS, AND SYMBOLS

Textual messages comprise a sequence of alphanumeric characters. When digitally transmitted, the characters are first encoded into a sequence of bits, called a *bit stream* or *baseband signal*. Groups of k bits can then be combined to form new digits, or *symbols*, from a finite symbol set or alphabet of $M = 2^k$ such symbols. A system using a symbol set size of M is referred to as an M -ary system. The value of k or M represents an important initial choice in the design of any digital communication system. For $k = 1$, the system is termed *binary*, the size of the symbol set is $M = 2$, and the modulator uses one of the two different waveforms to represent the binary “one” and the other to represent the binary “zero.” For this special case, the symbol and the bit are the same. For $k = 2$, the system is termed *quaternary* or *4-ary* ($M = 4$). At each symbol time, the modulator uses one of the four different waveforms that represents the symbol. The partitioning of the sequence of message bits is determined by the specification of the symbol set size, M . The following example should help clarify the relationship between the following terms: “message,” “character,” “symbol,” “bit,” and “digital waveform.”

2.3.1 Example of Messages, Characters, and Symbols

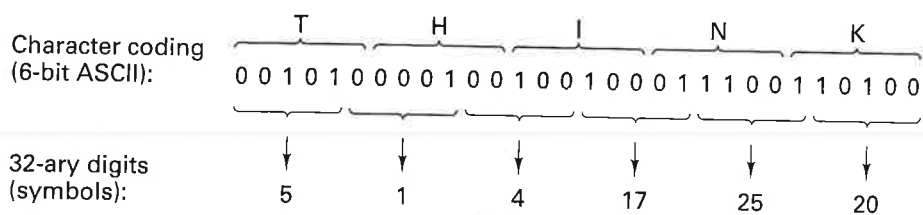
Figure 2.5 shows examples of bit stream partitioning, based on the system specification for the values of k and M . The textual message in the figure is the word “THINK.” Using 6-bit ASCII character coding (bit numbers 1 to 6 from Figure 2.3) yields a bit stream comprising 30 bits. In Figure 2.5a, the symbol set size, M , has been chosen to be 8 (each symbol represents an 8-ary digit). The bits are therefore partitioned into groups of three ($k = \log_2 8$); the resulting 10 numbers represent the 10 octal symbols to be transmitted. The transmitter must have a repertoire of eight waveforms $s_i(t)$, where $i = 1, \dots, 8$, to represent the possible symbols, any one of which may be transmitted during a symbol time. The final row of Figure 2.5a lists the 10 waveforms that an 8-ary modulating system transmits to represent the textual message “THINK.”

In Figure 2.5b, the symbol set size, M , has been chosen to be 32 (each symbol represents a 32-ary digit). The bits are therefore taken five at a time, and the resulting group of six numbers represent the six 32-ary symbols to be transmitted. Notice that there is no need for the symbol boundaries and the character boundaries to coincide. The first symbol represents $\frac{5}{6}$ of the first character, “T.” The second symbol



8-ary waveforms: $s_1(t)$ $s_2(t)$ $s_0(t)$ $s_4(t)$ $s_4(t)$ $s_4(t)$ $s_3(t)$ $s_4(t)$ $s_6(t)$ $s_4(t)$

(a)



32-ary waveforms: $s_5(t)$ $s_1(t)$ $s_4(t)$ $s_{17}(t)$ $s_{25}(t)$ $s_{20}(t)$

(b)

Figure 2.5 Messages, characters, and symbols. (a) 8-ary example. (b) 32-ary example.

represents the remaining $\frac{1}{6}$ of the character "T" and $\frac{4}{6}$ of the next character, "H," and so on. It is not necessary that the characters be partitioned more aesthetically. The system sees the characters as a string of digits to be transmitted; only the end user (or the user's teleprinter machine) ascribes textual meaning to the final delivered sequence of bits. In this 32-ary case, a transmitter needs a repertoire of 32 waveforms $s_i(t)$, where $i = 1, \dots, 32$, one for each possible symbol that may be transmitted. The final row of the figure lists the six waveforms that a 32-ary modulating system transmits to represent the textual message "THINK."

2.4 FORMATTING ANALOG INFORMATION

If the information is analog, it cannot be character encoded as in the case of textual data; the information must first be transformed into a digital format. The process of transforming an analog waveform into a form that is compatible with a digital com-

munication system starts with sampling the waveform to produce a discrete pulse-amplitude-modulated waveform, as described below.

2.4.1 The Sampling Theorem

The link between an analog waveform and its sampled version is provided by what is known as the *sampling process*. This process can be implemented in several ways, the most popular being the *sample-and-hold* operation. In this operation, a switch and storage mechanism (such as a transistor and a capacitor, or a shutter and a filmstrip) form a sequence of samples of the continuous input waveform. The output of the sampling process is called *pulse amplitude modulation* (PAM) because the successive output intervals can be described as a sequence of pulses with amplitudes derived from the input waveform samples. The analog waveform can be approximately retrieved from a PAM waveform by simple low-pass filtering. An important question: how closely can a filtered PAM waveform approximate the original input waveform? This question can be answered by reviewing the *sampling theorem*, which states the following [1]: A bandlimited signal having no spectral components above f_m hertz can be determined uniquely by values sampled at uniform intervals of

$$T_s \leq \frac{1}{2f_m} \text{ sec} \quad (2.1)$$

This particular statement is also known as the *uniform sampling theorem*. Stated another way, the upper limit on T_s can be expressed in terms of the sampling rate, denoted $f_s = 1/T_s$. The restriction, stated in terms of the sampling rate, is known as the *Nyquist criterion*. The statement is

$$f_s \geq 2f_m \quad (2.2)$$

The sampling rate $f_s = 2f_m$ is also called the *Nyquist rate*. The Nyquist criterion is a theoretically sufficient condition to allow an analog signal to be *reconstructed completely* from a set of uniformly spaced discrete-time samples. In the sections that follow, the validity of the sampling theorem is demonstrated using different sampling approaches.

2.4.1.1 Impulse Sampling

Here we demonstrate the validity of the sampling theorem using the frequency convolution property of the Fourier transform. Let us first examine the case of *ideal sampling* with a sequence of unit impulse functions. Assume an analog waveform, $x(t)$, as shown in Figure 2.6a, with a Fourier transform, $X(f)$, which is zero outside the interval $(-f_m < f < f_m)$, as shown in Figure 2.6b. The sampling of $x(t)$ can be viewed as the product of $x(t)$ with a periodic train of unit impulse functions $x_\delta(t)$, shown in Figure 2.6c and defined as

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \quad (2.3)$$

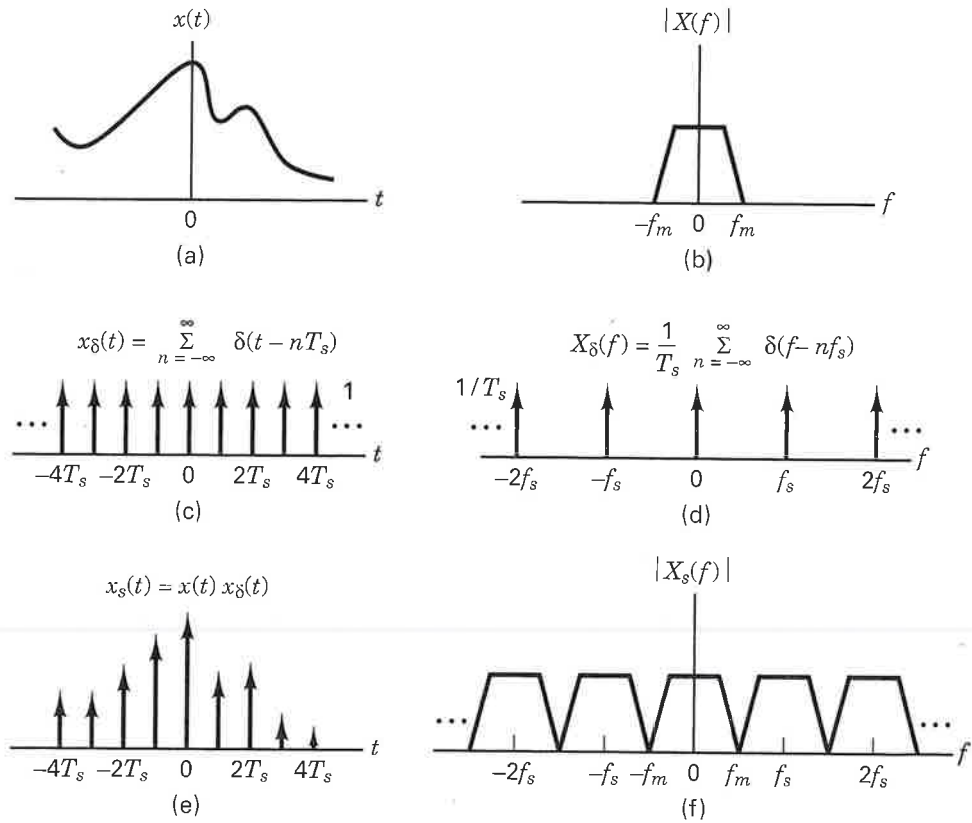


Figure 2.6 Sampling theorem using the frequency convolution property of the Fourier transform.

where T_s is the sampling period and $\delta(t)$ is the unit impulse or Dirac delta function defined in Section 1.2.5. Let us choose $T_s = 1/2f_m$, so that the Nyquist criterion is just satisfied.

The *sifting property* of the impulse function (see Section A.4.1) states that

$$x(t)\delta(t - t_0) = x(t_0)\delta(t - t_0) \quad (2.4)$$

Using this property, we can see that $x_s(t)$, the sampled version of $x(t)$ shown in Figure 2.6e, is given by

$$\begin{aligned} x_s(t) &= x(t)x_\delta(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s) \\ &= \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \end{aligned} \quad (2.5)$$

Using the *frequency convolution property* of the Fourier transform (see Section A.5.3), we can transform the time-domain product $x(t)x_\delta(t)$ of Equation (2.5) to the frequency-domain convolution $X(f) * X_\delta(f)$, where

$$X_s(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \quad (2.6)$$

is the Fourier transform of the impulse train $x_s(t)$ and where $f_s = 1/T_s$ is the sampling frequency. Notice that the Fourier transform of an impulse train is another impulse train; the values of the periods of the two trains are reciprocally related to one another. Figures 2.6c and d illustrate the impulse train $x_s(t)$ and its Fourier transform $X_s(f)$, respectively.

Convolution with an impulse function simply shifts the original function as follows:

$$X(f) * \delta(f - nf_s) = X(f - nf_s) \quad (2.7)$$

We can now solve for the transform $X_s(f)$ of the sampled waveform:

$$\begin{aligned} X_s(f) &= X(f) * X_s(f) = X(f) * \left[\frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] \\ &= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s) \end{aligned} \quad (2.8)$$

We therefore conclude that within the original bandwidth, the spectrum $X_s(f)$ of the sampled signal $x_s(t)$ is, to within a constant factor ($1/T_s$), exactly the same as that of $x(t)$. In addition, the spectrum repeats itself periodically in frequency every f_s hertz. The sifting property of an impulse function makes the convolving of an impulse train with another function easy to visualize. The impulses act as sampling functions. Hence, convolution can be performed graphically by sweeping the impulse train $X_s(f)$ in Figure 2.6d past the transform $|X(f)|$ in Figure 2.6b. This sampling of $|X(f)|$ at each step in the sweep replicates $|X(f)|$ at each of the frequency positions of the impulse train, resulting in $|X_s(f)|$, shown in Figure 2.6f.

When the sampling rate is chosen, as it has been here, such that $f_s = 2f_m$, each spectral replicate is separated from each of its neighbors by a frequency band exactly equal to f_s hertz, and the analog waveform can theoretically be completely recovered from the samples, by the use of filtering. However, a filter with infinitely steep sides would be required. It should be clear that if $f_s > 2f_m$, the replications will move farther apart in frequency, as shown in Figure 2.7a, making it easier to perform the filtering operation. A typical low-pass filter characteristic that might be used to separate the baseband spectrum from those at higher frequencies is shown in the figure. When the sampling rate is reduced, such that $f_s < 2f_m$, the replications will overlap, as shown in Figure 2.7b, and some information will be lost. The phenomenon, the result of undersampling (sampling at too low a rate), is called *aliasing*. The Nyquist rate, $f_s = 2f_m$, is the sampling rate below which aliasing occurs; to avoid aliasing, the Nyquist criterion, $f_s \geq 2f_m$, must be satisfied.

As a matter of practical consideration, neither waveforms of engineering interest nor realizable bandlimiting filters are strictly bandlimited. Perfectly bandlimited signals do not occur in nature (see Section 1.7.2); thus, realizable signals, even though we may think of them as bandlimited, always contain some aliasing. These signals and filters can, however, be considered to be "essentially" bandlimited. By

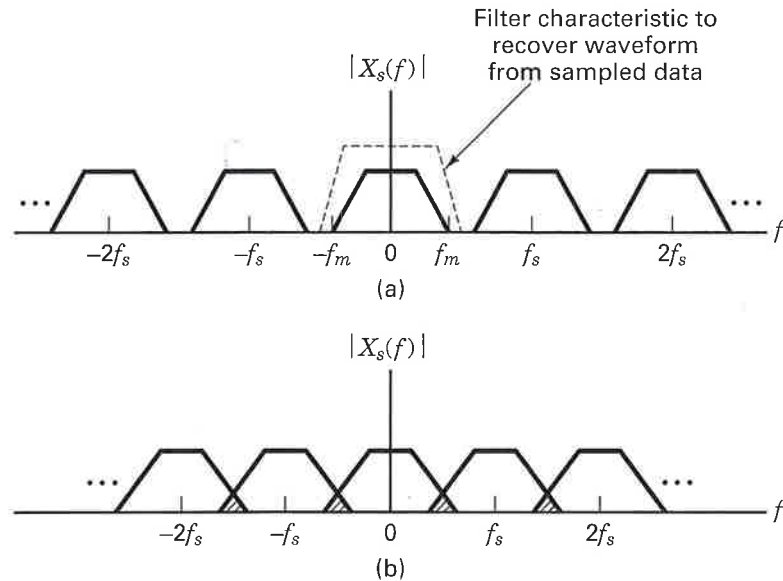


Figure 2.7 Spectra for various sampling rates. (a) Sampled spectrum ($f_s > 2f_m$). (b) Sampled spectrum ($f_s < 2f_m$).

this we mean that a bandwidth can be determined beyond which the spectral components are attenuated to a level that is considered negligible.

2.4.1.2 Natural Sampling

Here we demonstrate the validity of the sampling theorem using the frequency shifting property of the Fourier transform. Although instantaneous sampling is a convenient model, a more practical way of accomplishing the sampling of a bandlimited analog signal $x(t)$ is to multiply $x(t)$, shown in Figure 2.8a, by the pulse train or switching waveform $x_p(t)$, shown in Figure 2.8c. Each pulse in $x_p(t)$ has width T and amplitude $1/T$. Multiplication by $x_p(t)$ can be viewed as the opening and closing of a switch. As before, the sampling frequency is designated f_s , and its reciprocal, the time period between samples, is designated T_s . The resulting sampled-data sequence, $x_s(t)$, is illustrated in Figure 2.8e and is expressed as

$$x_s(t) = x(t)x_p(t) \quad (2.9)$$

The sampling here is termed *natural sampling*, since the top of each pulse in the $x_s(t)$ sequence retains the shape of its corresponding analog segment during the pulse interval. Using Equation (A.13), we can express the periodic pulse train as a Fourier series in the form

$$x_p(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_s t} \quad (2.10)$$

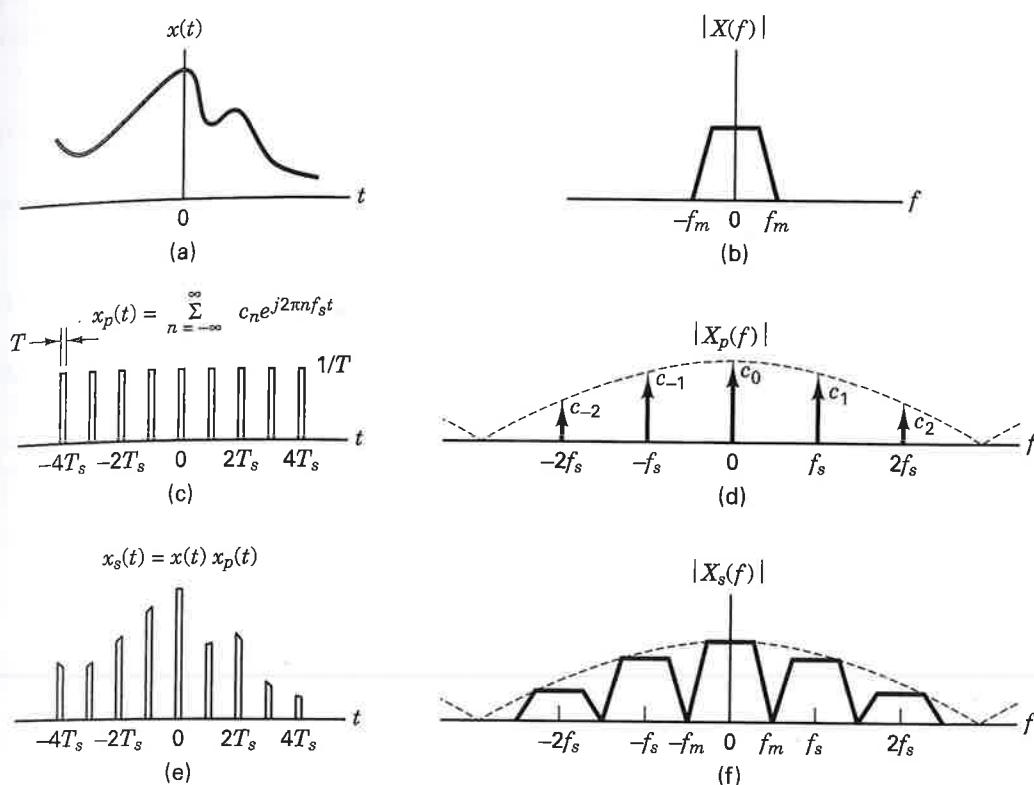


Figure 2.8 Sampling theorem using the frequency shifting property of the Fourier transform.

where the sampling rate, $f_s = 1/T_s$, is chosen equal to $2f_m$, so that the Nyquist criterion is just satisfied. From Equation (A.24), $c_n = (1/T_s) \text{sinc}(nT/T_s)$, where T is the pulse width, $1/T$ is the pulse amplitude, and

$$\text{sinc } y = \frac{\sin \pi y}{\pi y}$$

The envelope of the magnitude spectrum of the pulse train, seen as a dashed line in Figure 2.8d, has the characteristic sinc shape. Combining Equations (2.9) and (2.10) yields

$$x_s(t) = x(t) \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_s t} \quad (2.11)$$

The transform $X_s(f)$ of the sampled waveform is found as follows:

$$X_s(f) = \mathcal{F} \left\{ x(t) \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_s t} \right\} \quad (2.12)$$

For linear systems, we can interchange the operations of summation and Fourier transformation. Therefore, we can write

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_n \mathcal{F}\{x(t)e^{j2\pi n f_s t}\} \quad (2.13)$$

Using the *frequency translation* property of the Fourier transform (see Section A.3.2), we solve for $X_s(f)$ as follows:

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_n X(f - n f_s) \quad (2.14)$$

Similar to the unit impulse sampling case, Equation (2.14) and Figure 2.8f illustrate that $X_s(f)$ is a replication of $X(f)$, periodically repeated in frequency every f_s hertz. In this natural-sampled case, however, we see that $X_s(f)$ is weighted by the Fourier series coefficients of the pulse train, compared with a constant value in the impulse-sampled case. It is satisfying to note that *in the limit*, as the pulse width, T , approaches zero, c_n approaches $1/T_s$ for all n (see the example that follows), and Equation (2.14) converges to Equation (2.8).

Example 2.1 Comparison of Impulse Sampling and Natural Sampling

Consider a given waveform $x(t)$ with Fourier transform $X(f)$. Let $X_{s1}(f)$ be the spectrum of $x_{s1}(t)$, which is the result of sampling $x(t)$ with a unit impulse train $x_8(t)$. Let $X_{s2}(f)$ be the spectrum of $x_{s2}(t)$, the result of sampling $x(t)$ with a pulse train $x_p(t)$ with pulse width T , amplitude $1/T$, and period T_s . Show that in the limit, as T approaches zero, $X_{s1}(f) = X_{s2}(f)$.

Solution

From Equation (2.8),

$$X_{s1}(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - n f_s)$$

and from Equation (2.14),

$$X_{s2}(f) = \sum_{n=-\infty}^{\infty} c_n X(f - n f_s)$$

As the pulse with $T \rightarrow 0$, and the pulse amplitude approaches infinity (the area of the pulse remains unity), $x_p(t) \rightarrow x_8(t)$. Using Equation (A.14), we can solve for c_n in the limit as follows:

$$\begin{aligned} c_n &= \lim_{T \rightarrow 0} \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_p(t) e^{-j2\pi n f_s t} dt \\ &= \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_8(t) e^{-j2\pi n f_s t} dt \end{aligned}$$

Since, within the range of integration, $-T_s/2$ to $T_s/2$, the only contribution of $x_8(t)$ is that due to the impulse at the origin, we can write

and Fourier

(2.13)

(see Section

(2.14)

2.8f illustrate every f_s hertz. by the Fourier value in the pulse width, T , follows), and

$X_{s1}(f)$ be the pulse train $x_s(t)$. pulse train $x_p(t)$ the limit, as T

(the area of the lve for c_n in the

ution of $x_s(t)$ is

ion Chap. 2

$$c_n = \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} \delta(t) e^{-j 2\pi n f_s t} dt = \frac{1}{T_s}$$

Therefore, in the limit, $X_{s1}(f) = X_{s2}(f)$ for all n .

2.4.1.3 Sample-and-Hold Operation

The simplest and thus most popular sampling method, *sample and hold*, can be described by the convolution of the sampled pulse train, $[x(t)x_s(t)]$, shown in Figure 2.6e, with a unity amplitude rectangular pulse $p(t)$ of pulse width T_s . This time, convolution results in the *flattop* sampled sequence

$$\begin{aligned} x_s(t) &= p(t) * [x(t)x_s(t)] \\ &= p(t) * \left[x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \right] \end{aligned} \quad (2.15)$$

The Fourier transform, $X_s(f)$, of the time convolution in Equation (2.15) is the frequency-domain product of the transform $P(f)$ of the rectangular pulse and the periodic spectrum, shown in Figure 2.6f, of the impulse-sampled data:

$$\begin{aligned} X_s(f) &= P(f) \mathcal{F} \left\{ x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \right\} \\ &= P(f) \left\{ X(f) * \left[\frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] \right\} \\ &= P(f) \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s) \end{aligned} \quad (2.16)$$

Here, $P(f)$ is of the form $T_s \text{sinc } fT_s$. The effect of this product operation results in a spectrum similar in appearance to the natural-sampled example presented in Figure 2.8f. The most obvious effect of the hold operation is the significant attenuation of the higher-frequency spectral replicates (compare Figure 2.8f to Figure 2.6f), which is a desired effect. Additional analog postfiltering is usually required to finish the filtering process by further attenuating the residual spectral components located at the multiples of the sample rate. A secondary effect of the hold operation is the nonuniform spectral gain $P(f)$ applied to the desired baseband spectrum shown in Equation (2.16). The postfiltering operation can compensate for this attenuation by incorporating the inverse of $P(f)$ over the signal passband.

2.4.2 Aliasing

Figure 2.9 is a detailed view of the positive half of the baseband spectrum and one of the replicates from Figure 2.7b. It illustrates aliasing in the frequency domain. The overlapped region, shown in Figure 2.9b, contains that part of the spectrum which is aliased due to *undersampling*. The aliased spectral components represent ambiguous data that appear in the frequency band between $(f_s - f_m)$ and f_m . Figure 2.10 illustrates that a higher sampling rate f'_s , can eliminate the aliasing by separat-

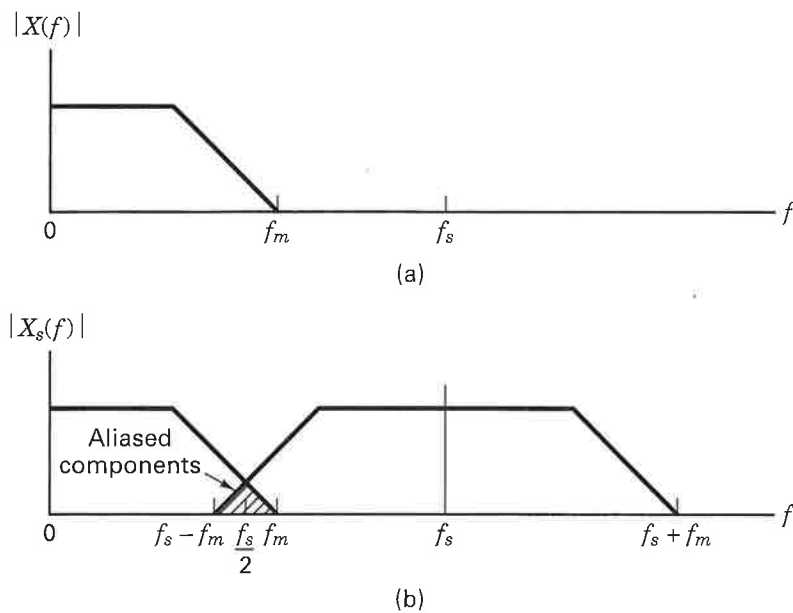


Figure 2.9 Aliasing in the frequency domain. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

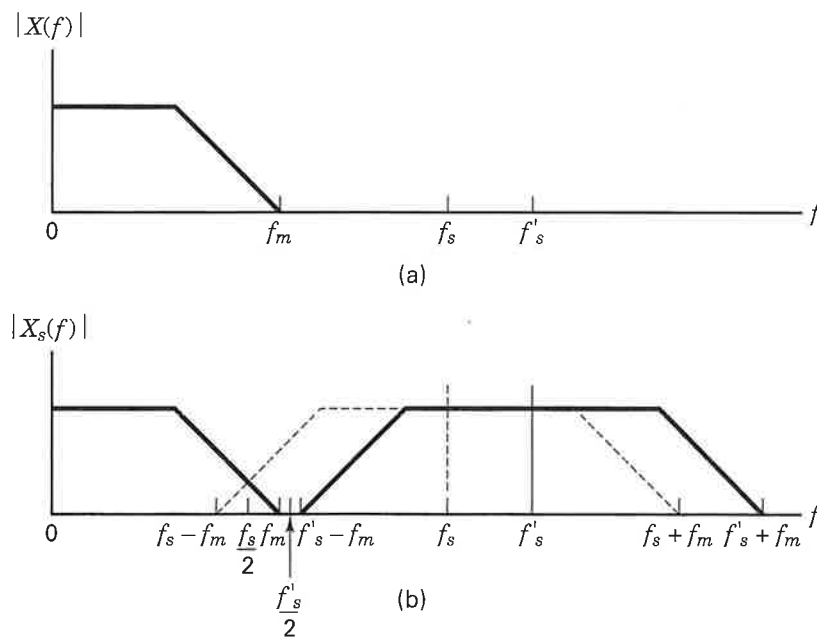


Figure 2.10 Higher sampling rate eliminates aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

ing the spectral replicates; the resulting spectrum in Figure 2.10b corresponds to the case in Figure 2.7a. Figures 2.11 and 2.12 illustrate two ways of eliminating aliasing using *antialiasing filters*. In Figure 2.11 the analog signal is *prefiltered* so that the new maximum frequency, f'_m , is reduced to $f_s/2$ or less. Thus there are no aliased components seen in Figure 2.11b, since $f_s > 2f'_m$. Eliminating the aliasing terms prior to sampling is good engineering practice. When the signal structure is well known, the aliased terms can be eliminated after sampling, with a low-pass filter operating on the sampled data [2]. In Figure 2.12 the aliased components are removed by *postfiltering* after sampling; the filter cutoff frequency, f''_m , removes the aliased components; f''_m needs to be less than $(f_s - f_m)$. Notice that the filtering techniques for eliminating the aliased portion of the spectrum in Figures 2.11 and 2.12 will result in a loss of some of the signal information. For this reason, the sample rate, cutoff bandwidth, and filter type selected for a particular signal bandwidth are all interrelated.

Realizable filters require a nonzero bandwidth for the transition between the passband and the required out-of-band attenuation. This is called the *transition bandwidth*. To minimize the system sample rate, we desire that the antialiasing filter have a small transition bandwidth. Filter complexity and cost rise sharply with narrower transition bandwidth, so a trade-off is required between the cost of a small transition bandwidth and the costs of the higher sampling rate, which are those of more storage and higher transmission rates. In many systems the answer has been to make the transition bandwidth between 10 and 20% of the signal band-

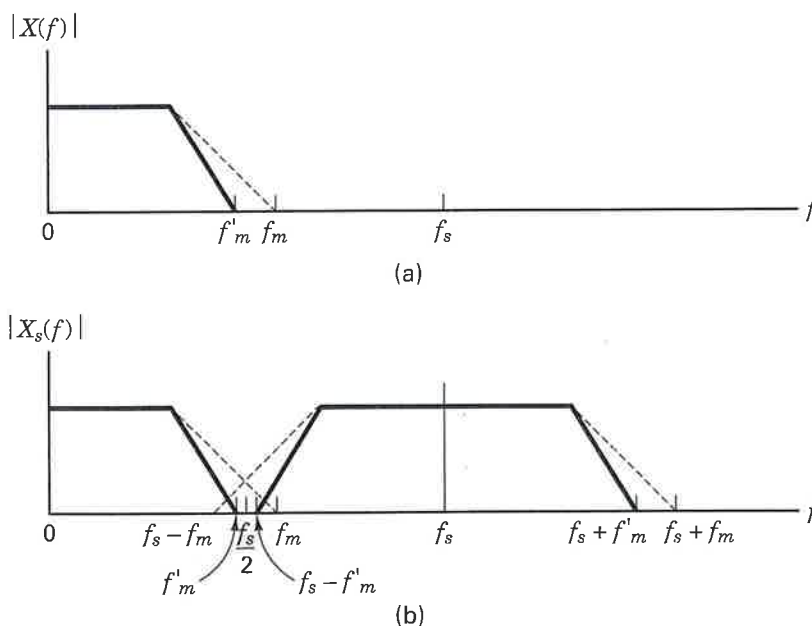


Figure 2.11 Sharper-cutoff filters eliminate aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

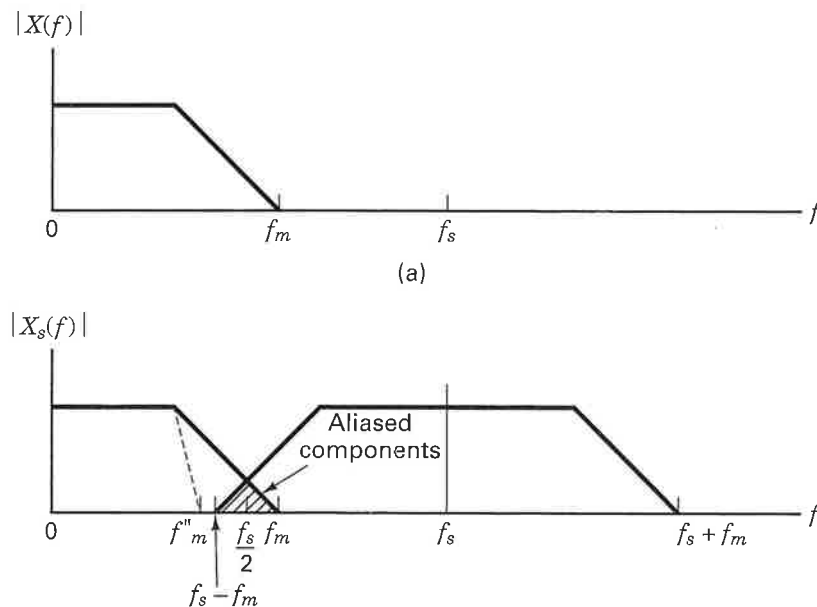


Figure 2.12 Postfilter eliminates aliased portion of spectrum. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

width. If we account for the 20% transition bandwidth of the antialiasing filter, we have an *engineer's version* of the Nyquist sampling rate:

$$f_s \geq 2.2f_m \quad (2.17)$$

Figure 2.13 provides some insight into aliasing as seen in the time domain. The sampling instants of the solid-line sinusoid have been chosen so that the sinusoidal signal is undersampled. Notice that the resulting ambiguity allows one to draw a totally different (dashed-line) sinusoid, following the undersampled points.

Example 2.2 Sampling Rate for a High-Quality Music System

We wish to produce a high-quality digitization of a 20-kHz bandwidth music source. We are to determine a reasonable sample rate for this source. By the engineer's version of the Nyquist rate, in Equation (2.17), the sampling rate should be greater than 44.0 ksamples/s. As a matter of comparison, the standard sampling rate for the compact disc digital audio player is 44.1 ksamples/s, and the standard sampling rate for studio-quality audio is 48.0 ksamples/s.

2.4.3 Why Oversample?

Oversampling is the most economic solution for the task of transforming an analog signal to a digital signal, or the reverse, transforming a digital signal to an analog signal. This is so because signal processing performed with high performance ana-

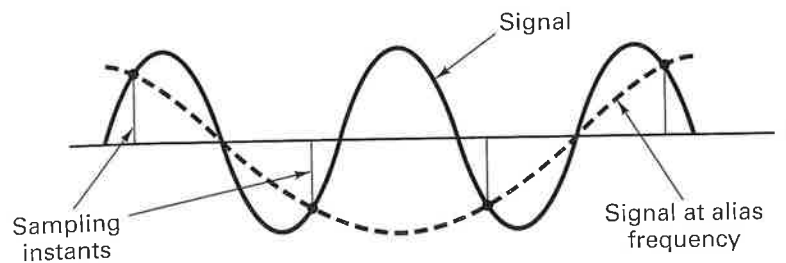


Figure 2.13 Alias frequency generated by sub-Nyquist sampling rate.

log equipment is typically much more costly than using digital signal processing equipment to perform the same task. Consider the task of transforming analog signals to digital signals. When this task is performed without the benefit of over-sampling, the process is characterized by three simple steps, performed in the order that follows.

Without Oversampling

1. The signal passes through a high performance analog lowpass filter to limit its bandwidth.
2. The filtered signal is **sampled** at the Nyquist rate for the (approximated) bandlimited signal. As described in Section 1.7.2, a strictly bandlimited signal is not realizable.
3. The samples are processed by an analog-to-digital converter that maps the continuous-valued samples to a finite list of discrete output levels.

When this task is performed with the benefit of over-sampling, the process is best described as five simple steps, performed in the order that follows.

With Oversampling

1. The signal is passed through a low performance (less costly) analog low-pass filter (prefilter) to limit its bandwidth.
2. The pre-filtered signal is sampled at the (now higher) Nyquist rate for the (approximated) bandlimited signal.
3. The samples are processed by an analog-to-digital converter that maps the continuous-valued samples to a finite list of discrete output levels.
4. The digital samples are then processed by a high performance digital filter to reduce the bandwidth of the digital samples.
5. The sample rate at the output of the digital filter is reduced in proportion to the bandwidth reduction obtained by this digital filter.

The next two sections examine the benefits of over-sampling.

2.4.3.1 Analog Filtering, Sampling, and Analog to Digital Conversion

The analog filter that limits the bandwidth of an input signal has a passband frequency equal to the signal bandwidth, followed by a transition to a stop band. The bandwidth of the transition region results in an increase in bandwidth of the output signal by some amount f_t . The Nyquist rate f_s for the filtered output, nominally equal to $2f_m$ (twice the highest frequency in the sampled signal) must now be increased to $2f_m + f_t$. The transition bandwidth of the filter represents an overhead in the sampling process. This additional spectral interval does not represent useful signal bandwidth but rather protects the signal bandwidth by reserving a spectral region for the aliased spectrum due to the sampling process. The aliasing stems from the fact that real signals cannot be strictly bandlimited. Typical transition bandwidths represent a 10- to 20-percent increase of the sample rate relative to that dictated by the Nyquist criterion. Examples of this overhead are seen in the compact disc (CD) digital audio system, for which the two-sided bandwidth is 40 kHz and the sample rate is 44.1 kHz, and also in the digital audio tape (DAT) system, which also has a two-sided bandwidth of 40 kHz with a sample rate of 48.0 kHz.

Our intuition and initial impulse is to keep the sample rate as low as possible by building analog filters with narrow transition bandwidths. However, analog filters can exhibit two undesirable characteristics. First, they can exhibit distortion (nonlinear phase versus frequency) due to narrow transition bandwidths. Second, the cost can be high because narrow transition bandwidths dictate high-order filters (see Section 1.6.3.2) requiring a large number of high-quality components. Our quandary is that we wish to operate the sampler at the lowest possible rate to reduce the data-storage cost. To meet this goal we might build a sophisticated analog filter with a narrow transition bandwidth. But such a filter is not only expensive, it also distorts the very signal it has been designed to protect (from undesired aliasing).

The solution (oversampling) is elegant—having been given a problem that we can't solve, we convert it to one that we can solve. We elect to use a low-cost, less sophisticated analog prefilter to limit the bandwidth of the input signal. This analog filter has been simplified by choosing a wider transition bandwidth. With a wider transition bandwidth, the required sample rate must now be increased to accommodate this larger spectrum. We typically start by selecting the higher sample rate to be 4 times the original sample rate, and then we design the analog filter to have a transition bandwidth that matches the increased sample rate. As an example, rather than sampling a CD signal at 44.1 kHz with a transition bandwidth of 4.1 kHz implemented with a sophisticated 10th order elliptic filter (implying that the filter includes 10 energy storage elements, such as capacitors and inductors), we might choose the option to employ oversampling. In that case, we could operate the sampler at 176.4 kHz with a transition bandwidth of 136.4 kHz implemented with a simpler 4th-order elliptic filter (having only 4 energy storage elements).

2.4.3.2 Digital Filtering and Resampling

Now that we have the sampled data, with its higher-than-desired sample rate, we pass the sampled data through a high-performance, low-cost, digital filter to perform the desired anti-alias filtering. The digital filter can realize the narrow

version

a passband fre-
stop band. The
th of the output
nominally equal
be increased to
l in the sampling
ignal bandwidth
n for the aliased
act that real sig-
represent a 10- to
e Nyquist crite-
digital audio sys-
e is 44.1 kHz, and
led bandwidth of

ow as possible by
er, analog filters
stortion (nonlin-
Second, the cost
order filters (see
ts. Our quandary
reduce the data-
g filter with a nar-
t also distorts the
).

problem that we
e a low-cost, less
gnal. This analog
lth. With a wider
reased to accom-
higher sample rate
og filter to have a
As an example,
bandwidth of 4.1
implying that the
nd inductors), we
we could operate
Hz implemented
ge elements).

sired sample rate,
st, digital filter to
ealize the narrow

ulation Chap. 2

transition bandwidth without the distortion associated with analog filters, and it can operate at low cost. We next reduce the sample rate of the signal (resample) after the digital filtering operation that had reduced the transition bandwidth. Good digital signal processing techniques combine the filtering and the resampling in a single structure.

Now we address a system consideration to further improve the quality of the data collection process. The analog prefilter induces some amplitude and phase distortion. We know precisely what this distortion is, and we design the digital filter so that it not only completes the anti-aliasing task of the analog prefilter, but also compensates for its gain and phase distortion. The composite response can be made as good as we want it to be. Thus we obtain a collected signal of higher quality (less distortion) at reduced cost. Digital signal processing hardware, an extension of the computer industry, is characterized by significantly lower prices each year, which has not been the case with analog processing.

In a similar fashion, oversampling is employed in the process of converting the digital signal to an analog signal (DAC). The analog filter following the DAC suffers from distortion if it has a sharp transition bandwidth. But the transition bandwidth will not be narrow if the output data presented to the DAC has been digitally oversampled.

2.4.4 Signal Interface for a Digital System

Let us examine four ways in which analog source information can be described. Figure 2.14 illustrates the choices. Let us refer to the waveform in Figure 2.14a as the *original analog waveform*. Figure 2.14b represents a sampled version of the original waveform, typically referred to as *natural-sampled data* or PAM (pulse amplitude modulation). Do you suppose that the sampled data in Figure 2.14b are compatible with a digital system? No, they are not, because the amplitude of each natural sample still has an infinite number of possible values; a digital system deals with a finite number of values. Even if the sampling is flat-top sampling, the possible pulse values form an infinite set, since they reflect all the possible values of the continuous analog waveform. Figure 2.14c illustrates the original waveform represented by discrete pulses. Here the pulses have flat tops and the pulse amplitude values are limited to a finite set. Each pulse is expressed as a level from a finite number of predetermined levels; each such level can be represented by a symbol from a finite alphabet. The pulses in Figure 2.14c are referred to as *quantized samples*; such a format is the obvious choice for interfacing with a digital system. The format in Figure 2.14d may be construed as the output of a sample-and-hold circuit. When the sample values are quantized to a finite set, this format can also interface with a digital system. After quantization, the analog waveform can still be recovered, but not precisely; improved reconstruction fidelity of the analog waveform can be achieved by increasing the number of quantization levels (requiring increased system bandwidth). Signal distortion due to quantization is treated in the following sections (and later in Chapter 13).

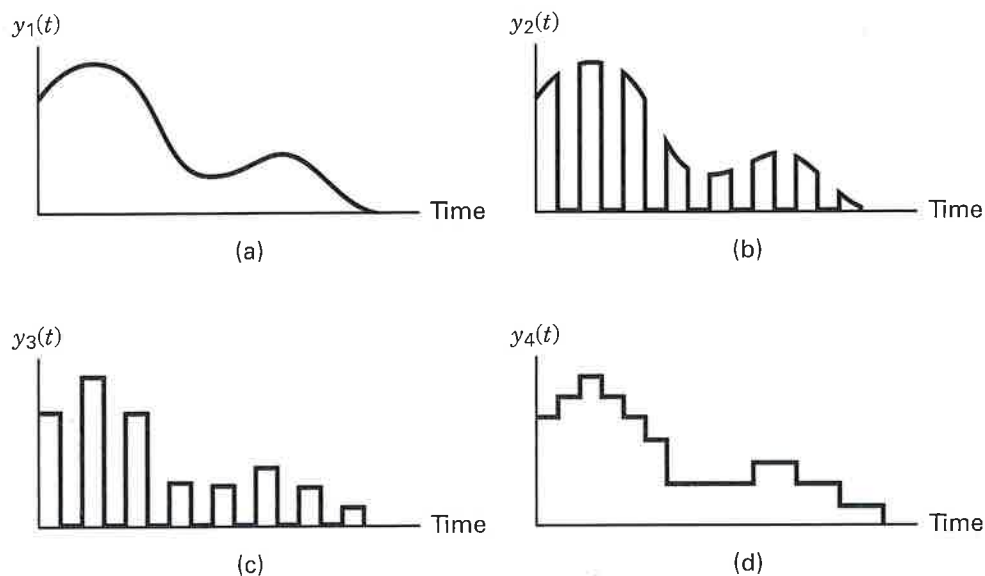


Figure 2.14 Amplitude and time coordinates of source data. (a) Original analog waveform. (b) Natural-sampled data. (c) Quantized samples. (d) Sample and hold.

2.5 SOURCES OF CORRUPTION

The analog signal recovered from the sampled, quantized, and transmitted pulses will contain corruption from several sources. The sources of corruption are related to (1) sampling and quantizing effects, and (2) channel effects. These effects are considered in the sections that follow.

2.5.1 Sampling and Quantizing Effects

2.5.1.1 Quantization Noise

The distortion inherent in quantization is a round-off or truncation error. The process of encoding the PAM signal into a quantized PAM signal involves discarding some of the original analog information. This distortion, introduced by the need to approximate the analog waveform with quantized samples, is referred to as *quantization noise*; the amount of such noise is inversely proportional to the number of levels employed in the quantization process. (The signal-to-noise ratio of quantized pulses is treated in Sections 2.5.3 and 13.2.)

2.5.1.2 Quantizer Saturation

The quantizer (or analog-to-digital converter) allocates L levels to the task of approximating the continuous range of inputs with a finite set of outputs. The range of inputs for which the difference between the input and output is small is

called the *operating range* of the converter. If the input exceeds this range, the difference between the input and the output becomes large, and we say that the converter is operating in *saturation*. Saturation errors, being large, are more objectionable than quantizing noise. Generally, saturation is avoided by the use of automatic gain control (AGC), which effectively extends the operating range of the converter. (Chapter 13 covers quantizer saturation in greater detail.)

2.5.1.3 Timing Jitter

Our analysis of the sampling theorem predicted precise reconstruction of the signal based on uniformly spaced samples of the signal. If there is a slight jitter in the position of the sample, the sampling is no longer uniform. Although exact reconstruction is still possible if the sample positions are accurately known, the jitter is usually a random process and thus the sample positions are not accurately known. The effect of the jitter is equivalent to frequency modulation (FM) of the baseband signal. If the jitter is random, a low-level wideband spectral contribution is induced whose properties are very close to those of the quantizing noise. If the jitter exhibits periodic components, as might be found in data extracted from a tape recorder, the periodic FM will induce low-level spectral lines in the data. Timing jitter can be controlled with very good power supply isolation and stable clock references.

2.5.2 Channel Effects

2.5.2.1 Channel Noise

Thermal noise, interference from other users, and interference from circuit switching transients can cause errors in detecting the pulses carrying the digitized samples. Channel-induced errors can degrade the reconstructed signal quality quite quickly. This rapid degradation of output signal quality with channel-induced errors is called a *threshold effect*. If the channel noise is small, there will be no problem detecting the presence of the waveforms. Thus, small noise does not corrupt the reconstructed signals. In this case, the only noise present in the reconstruction is the quantization noise. On the other hand, if the channel noise is large enough to affect our ability to detect the waveforms, the resulting detection error causes reconstruction errors. A large difference in behavior can occur for very small changes in channel noise level.

2.5.2.2 Intersymbol Interference

The channel is always bandlimited. A bandlimited channel disperses or spreads a pulse waveform passing through it (see Section 1.6.4). When the channel bandwidth is much greater than the pulse bandwidth, the spreading of the pulse will be slight. When the channel bandwidth is close to the signal bandwidth, the spreading will exceed a symbol duration and cause signal pulses to overlap. This overlapping is called *intersymbol interference* (ISI). Like any other source of interference, ISI causes system degradation (higher error rates); it is a particularly



(a) Original samples.

transmitted pulses
ption are related
These effects are

quantization error. The
involves discarding
introduced by the
, is referred to as
proportional to the num-
ber-to-noise ratio of

relates to the task of
of outputs. The
output is small is

insidious form of interference because raising the signal power to overcome the interference will not always improve the error performance. (Details of how ISI is handled are presented in the next chapter, in Sections 3.3 and 3.4.)

2.5.3 Signal-to-Noise Ratio for Quantized Pulses

Figure 2.15 illustrates an L -level linear quantizer for an analog signal with a peak-to-peak voltage range of $V_{pp} = V_p - (-V_p) = 2V_p$ volts. The quantized pulses assume positive and negative values, as shown in the figure. The step size between quantization levels, called the *quantile interval*, is denoted q volts. When the quantization levels are uniformly distributed over the full range, the quantizer is called a *uniform or linear quantizer*. Each sample value of the analog waveform is approximated with a quantized pulse; the approximation will result in an error no larger than $q/2$ in the positive direction or $-q/2$ in the negative direction. The degradation of the signal due to quantization is therefore limited to half a quantile interval, $\pm q/2$ volts.

A useful figure of merit for the uniform quantizer is the quantizer variance (mean-square error assuming zero mean). If we assume that the quantization error, e , is uniformly distributed over a single quantile interval q -wide (i.e., the analog input takes on all values with equal probability), the quantizer error variance is found to be

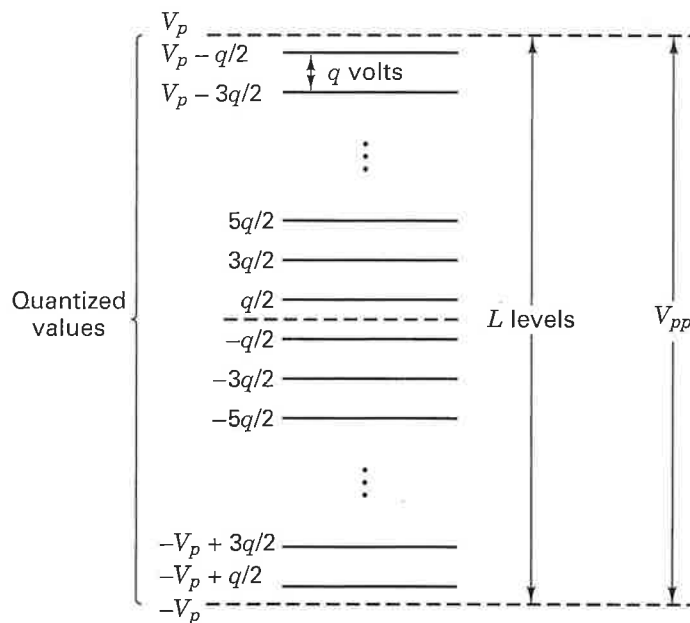


Figure 2.15 Quantization levels.

$$\sigma^2 = \int_{-q/2}^{+q/2} e^2 p(e) de \quad (2.18a)$$

$$= \int_{-q/2}^{+q/2} e^2 \frac{1}{q} de = \frac{q^2}{12} \quad (2.18b)$$

where $p(e) = 1/q$ is the (uniform) probability density function of the quantization error. The variance, σ^2 , corresponds to the *average quantization noise power*. The peak power of the analog signal (normalized to 1 Ω) can be expressed as

$$V_p^2 = \left(\frac{V_{pp}}{2} \right)^2 = \left[\frac{q(L-1)}{2} \right]^2 \approx \left(\frac{Lq}{2} \right)^2 = \frac{L^2 q^2}{4} \quad (2.19)$$

where L is the number of quantization levels. Equations (2.18) and (2.19) combined yield the ratio of *peak* signal power to *average* quantization noise power

$$\left(\frac{S}{N} \right)_q = \frac{L^2 q^2 / 4}{q^2 / 12} = 3L^2 \quad (2.20)$$

where N is the average quantization noise power, σ^2 . We see that $(S/N)_q$ increases as a function of the number of quantization levels squared. In the limit (as $L \rightarrow \infty$), the signal becomes analog (infinite quantization and zero quantization noise). Note that for random signals, the important $(S/N)_q$ parameter deals with *average* rather than *peak* signal power. In that case, to obtain average signal power, one needs to know the signal's pdf.

2.6 PULSE CODE MODULATION

Pulse code modulation (PCM) is the name given to the class of baseband signals obtained from the quantized PAM signals by encoding each quantized sample into a *digital word* [3]. The source information is sampled and quantized to one of L levels; then each quantized sample is digitally encoded into an ℓ -bit ($\ell = \log_2 L$) codeword. For baseband transmission, the codeword bits will then be transformed to pulse waveforms. The essential features of binary PCM are shown in Figure 2.16. Assume that an analog signal $x(t)$ is limited in its excursions to the range -4 to $+4$ V. The step size between quantization levels has been set at 1 V. Thus, eight quantization levels are employed; these are located at $-3.5, -2.5, \dots, +3.5$ V. We assign the code number 0 to the level at -3.5 V, the code number 1 to the level at -2.5 V, and so on, until the level at 3.5 V, which is assigned the code number 7. Each code number has its representation in binary arithmetic, ranging from 000 for code number 0 to 111 for code number 7. Why have the voltage levels been chosen in this manner, compared with using a sequence of consecutive integers, 1, 2, 3, $\dots$? The choice of voltage levels is guided by two constraints. First, the quantile intervals between the levels should be equal; and second, it is convenient for the levels to be symmetrical about zero.

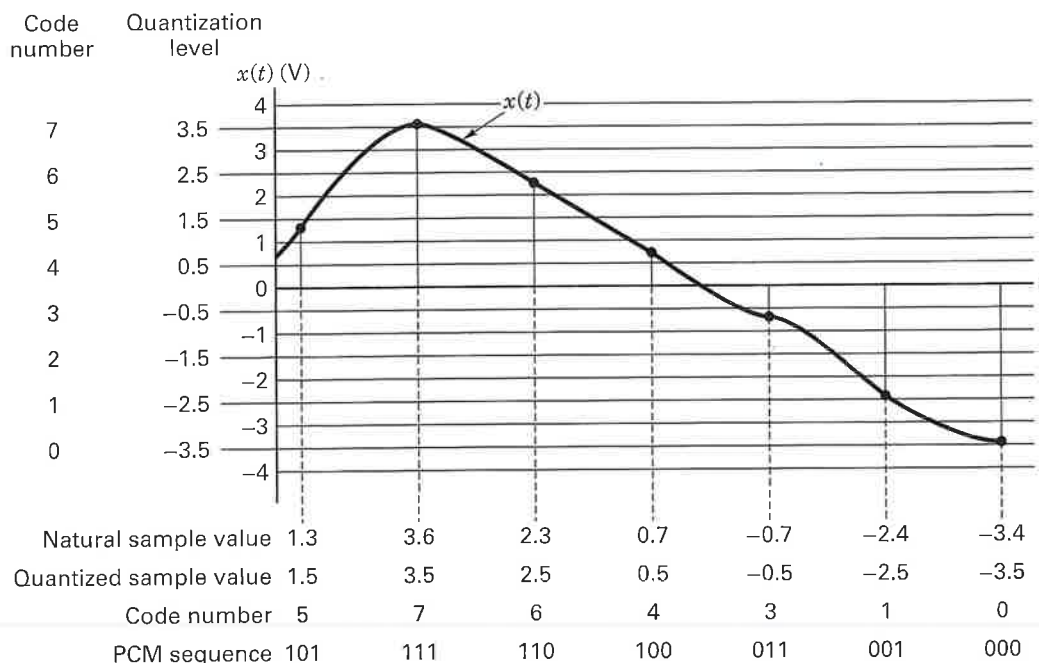


Figure 2.16 Natural samples, quantized samples, and pulse code modulation. (Reprinted with permission from Taub and Schilling, *Principles of Communications Systems*, McGraw-Hill Book Company, New York, 1971, Fig. 6.5-1, p. 205.)

The ordinate in Figure 2.16 is labeled with quantization levels and their code numbers. Each sample of the analog signal is assigned to the quantization level closest to the value of the sample. Beneath the analog waveform $x(t)$ are seen four representations of $x(t)$, as follows: the natural sample values, the quantized sample values, the code numbers, and the PCM sequence.

Note, that in the example of Figure 2.16, each sample is assigned to one of eight levels or a three-bit PCM sequence. Suppose that the analog signal is a musical passage, which is sampled at the Nyquist rate. And, suppose that when we listen to the music in digital form, it sounds terrible. What could we do to improve the fidelity? Recall that the process of quantization replaces the true signal with an approximation (i.e., adds quantization noise). Thus, increasing the number of levels will reduce the quantization noise. If we double the number of levels to 16, what are the consequences? In that case, each analog sample will be represented as a four-bit PCM sequence. Will that cost anything? In a real-time communication system, the messages must not be delayed. Hence, the transmission time for each sample must be the same, regardless of how many bits represent the sample. Hence, when there are more bits per sample, the bits must move faster; in other words, they must be replaced by "skinnier" bits. The data rate is thus increased, and the cost is a greater transmission bandwidth. This explains how one can generally obtain better fidelity at the cost of more transmission bandwidth. Be aware, however,

that there are some communication applications where delay is permissible. For example, consider the transmission of planetary images from a spacecraft. The Galileo project, launched in 1989, was on such a mission to photograph and transmit images of the planet Jupiter. The Galileo spacecraft arrived at its Jupiter destination in 1995. The journey took several years; therefore, any excess signal delay of several minutes (or hours or days) would certainly not be a problem. In such cases, the cost of more quantization levels and greater fidelity need not be bandwidth; it can be time delay.

In Figure 2.1, the term "PCM" appears in two places. First, it is a formatting topic, since the process of analog-to-digital (A/D) conversion involves sampling, quantization, and ultimately yields binary digits via the conversion of quantized PAM to PCM. Here, PCM digits are just binary numbers—a baseband carrier wave has not yet been discussed. The second appearance of PCM in Figure 2.1 is under the heading *Baseband Signaling*. Here, we list various PCM waveforms (line codes) that can be used to "carry" the PCM digits. Therefore, note that the difference between PCM and a PCM waveform is that the former represents a bit sequence, and the latter represents a particular waveform conveyance of that sequence.

2.7 UNIFORM AND NONUNIFORM QUANTIZATION

2.7.1 Statistics of Speech Amplitudes

Speech communication is a very important and specialized area of digital communications. Human speech is characterized by unique statistical properties; one such property is illustrated in Figure 2.17. The abscissa represents speech signal magnitudes, normalized to the root-mean-square (rms) value of such magnitudes through a typical communication channel, and the ordinate is probability. For most voice

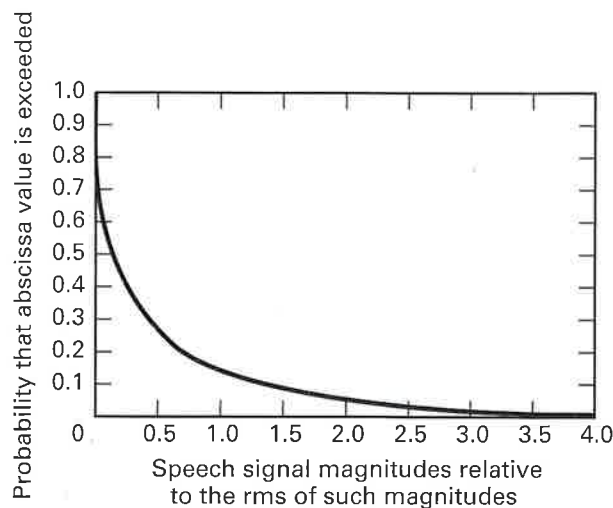


Figure 2.17 Statistical distribution of single-talker speech signal magnitudes.

communication channels, very low speech volumes predominate; 50% of the time, the voltage characterizing detected speech energy is less than one-fourth of the rms value. Large amplitude values are relatively rare; only 15% of the time does the voltage exceed the rms value. We see from Equation (2.18b) that the quantization noise depends on the step size (size of the quantile interval). When the steps are uniform in size the quantization is known as *uniform quantization*. Such a system would be wasteful for speech signals; many of the quantizing steps would rarely be used. In a system that uses equally spaced quantization levels, the quantization noise is the same for all signal magnitudes. Therefore, with uniform quantization, the signal-to-noise (SNR) is worse for low-level signals than for high-level signals. *Nonuniform quantization* can provide fine quantization of the weak signals and coarse quantization of the strong signals. Thus in the case of nonuniform quantization, quantization noise can be made proportional to signal size. The effect is to improve the overall SNR by reducing the noise for the predominant weak signals, at the expense of an increase in noise for the rarely occurring strong signals. Figure 2.18 compares the quantization of a strong versus a weak signal for uniform and nonuniform quantization. The staircase-like waveforms represent the approximations to the analog waveforms (after quantization distortion has been introduced). The SNR improvement that nonuniform quantization provides for the weak signal should be apparent. Nonuniform quantization can be used to make the SNR a constant for all signals within the input range. For voice signals, the typical input signal

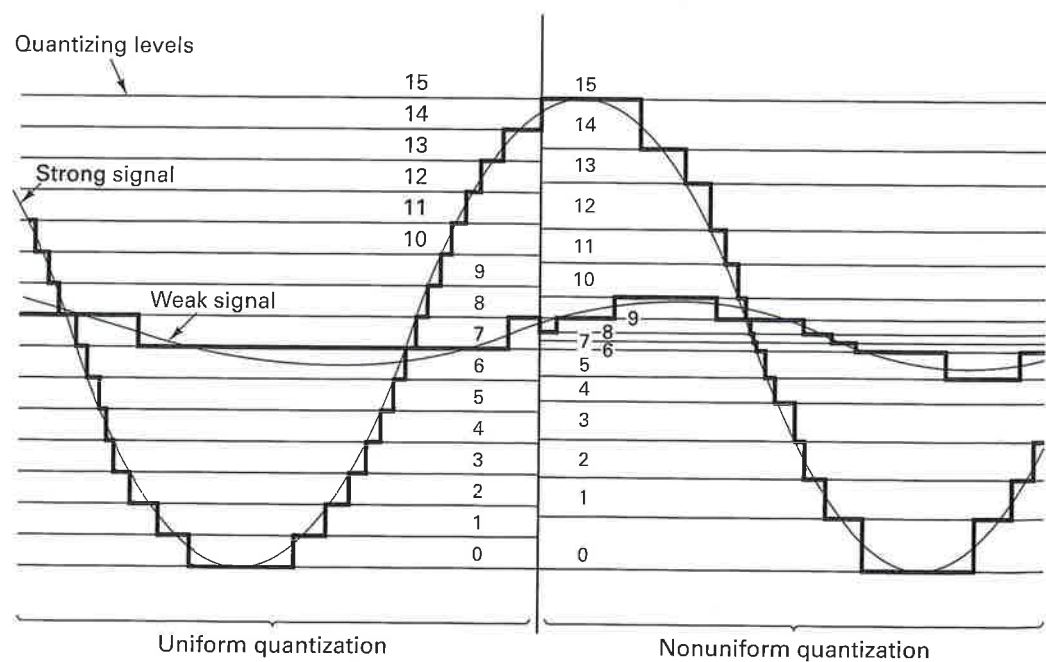


Figure 2.18 Uniform and nonuniform quantization of signals.

dynamic range is 40 decibels (dB), where a decibel is defined in terms of the ratio of power P_2 to power P_1 :

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} \quad (2.21)$$

With a uniform quantizer, weak signals would experience a 40-dB-poorer SNR than that of strong signals. The standard telephone technique of handling the large range of possible input signal levels is to use a *logarithmic-compressed* quantizer instead of a uniform one. With such a nonuniform compressor the output SNR is independent of the distribution of input signal levels.

2.7.2 Nonuniform Quantization

One way of achieving nonuniform quantization is to use a nonuniform quantizer characteristic, shown in Figure 2.19a. More often, nonuniform quantization is achieved by first distorting the original signal with a logarithmic compression characteristic, as shown in Figure 2.19b, and then using a uniform quantizer. For small magnitude signals the compression characteristic has a much steeper slope than for large magnitude signals. Thus, a given signal change at small magnitudes will carry the uniform quantizer through more steps than the same change at large magnitudes. The compression characteristic effectively changes the distribution of the

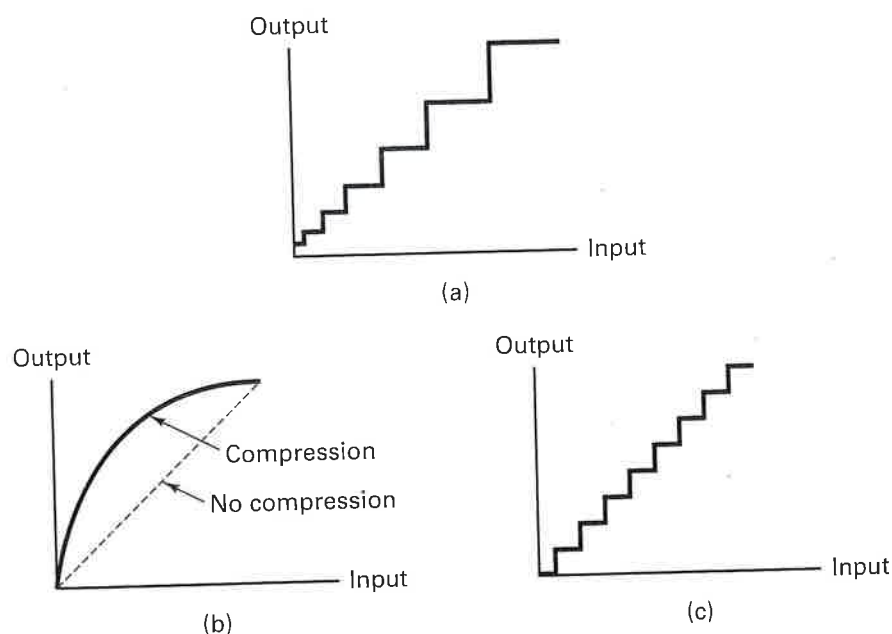


Figure 2.19 (a) Nonuniform quantizer characteristic. (b) Compression characteristic. (c) Uniform quantizer characteristic.

input signal magnitudes so that there is not a preponderance of *low* magnitude signals at the output of the compressor. After compression, the distorted signal is used as the input to a uniform (linear) quantizer characteristic, shown in Figure 2.19c. At the receiver, an inverse compression characteristic, called *expansion*, is applied so that the overall transmission is not distorted. The processing pair (compression and expansion) is usually referred to as *companding*.

2.7.3 Companding Characteristics

The early PCM systems implemented a smooth logarithmic compression function. Today, most PCM systems use a piecewise linear approximation to the logarithmic compression characteristic. In North America, a μ -law compression characteristic

$$y = y_{\max} \frac{\log_e [1 + \mu(|x|/x_{\max})]}{\log_e(1 + \mu)} \operatorname{sgn} x \quad (2.22)$$

is used, where

$$\operatorname{sgn} x = \begin{cases} +1 & \text{for } x \geq 0 \\ -1 & \text{for } x < 0 \end{cases}$$

and where μ is a positive constant, x and y represent input and output voltages, and $x_{\max}$ and $y_{\max}$ are the maximum positive excursions of the input and output voltages, respectively. The compression characteristic is shown in Figure 2.20a for several values of μ . In North America, the standard value for μ is 255. Notice that $\mu = 0$ corresponds to linear amplification (uniform quantization).

Another compression characteristic, used mainly in Europe, is the A -law characteristic, defined as

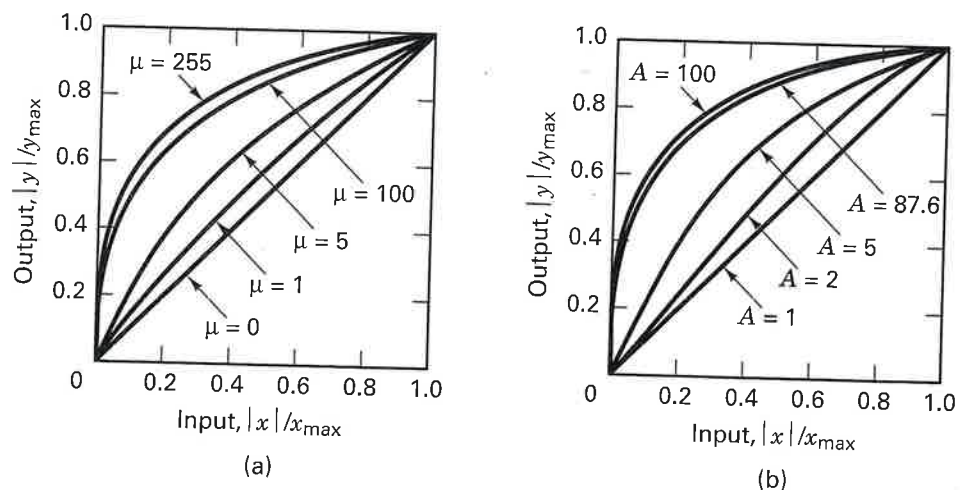


Figure 2.20 Compression characteristics. (a) μ -law characteristic. (b) A -law characteristic.

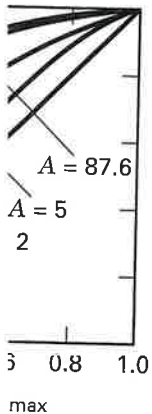
magnitude sig-
d signal is used
n Figure 2.19c.
sion, is applied
r (compression

ssion function.
he logarithmic
characteristic

(2.22)

t voltages, and
input voltages,
0a for several
ice that $\mu = 0$

is the A -law



istic.

1 Chap. 2

$$y = \begin{cases} y_{\max} \frac{A(|x|/x_{\max})}{1 + \log_e A} \operatorname{sgn} x & 0 < \frac{|x|}{x_{\max}} \leq \frac{1}{A} \\ y_{\max} \frac{1 + \log_e [A(|x|/x_{\max})]}{1 + \log_e A} \operatorname{sgn} x & \frac{1}{A} < \frac{|x|}{x_{\max}} < 1 \end{cases} \quad (2.23)$$

where A is a positive constant and x and y are defined the same way as they were for Equation (2.22). The A -law compression characteristic is shown in Figure 2.20b for several values of A . A standard value for A is 87.6. (The subjects of uniform and nonuniform quantization are treated further in Chapter 13, Section 13.2.)

2.8 BASEBAND TRANSMISSION

2.8.1 Waveform Representation of Binary Digits

In Section 2.6, it was shown how analog waveforms are transformed into binary digits via the use of PCM. There is nothing “physical” about the digits resulting from this process. Digits are just abstractions—a way to describe the message information. Thus, we need something physical that will represent or “carry” the digits.

We will represent the binary digits with electrical pulses in order to transmit them through a baseband channel. Such a representation is shown in Figure 2.21. Codeword time slots are shown in Figure 2.21a, where the codeword is a 4-bit representation of each quantized sample. In Figure 2.21b, each binary one is represented by a pulse and each binary zero is represented by the absence of a pulse. Thus a sequence of electrical pulses having the pattern shown in Figure 2.21b can be used to transmit the information in the PCM bit stream, and hence the information in the quantized samples of a message.

At the receiver, a determination must be made as to the presence or absence of a pulse in each bit time slot. It will be shown in Section 2.9 that the likelihood of correctly detecting the presence of a pulse is a function of the received pulse energy (or area under the pulse). Thus there is an advantage in making the pulse width T' in Figure 2.21b as wide as possible. If we increase the pulse width to the maximum possible (equal to the bit time T), we have the waveform shown in Figure 2.21c. Rather than describe this waveform as a sequence of present or absent pulses (unipolar), we can describe it as a sequence of transitions between two levels (bipolar). When the waveform occupies the upper voltage level it represents a binary one; when it occupies the lower voltage level it represents a binary zero.

2.8.2 PCM Waveform Types

When pulse modulation is applied to a *binary* symbol, the resulting binary waveform is called a pulse-code modulation (PCM) waveform. There are several types of PCM waveforms that are described below and illustrated in Figure 2.22; in telephony applications, these waveforms are often called *line codes*. When pulse modulation is applied to a *nonbinary* symbol, the resulting waveform is called an M -ary

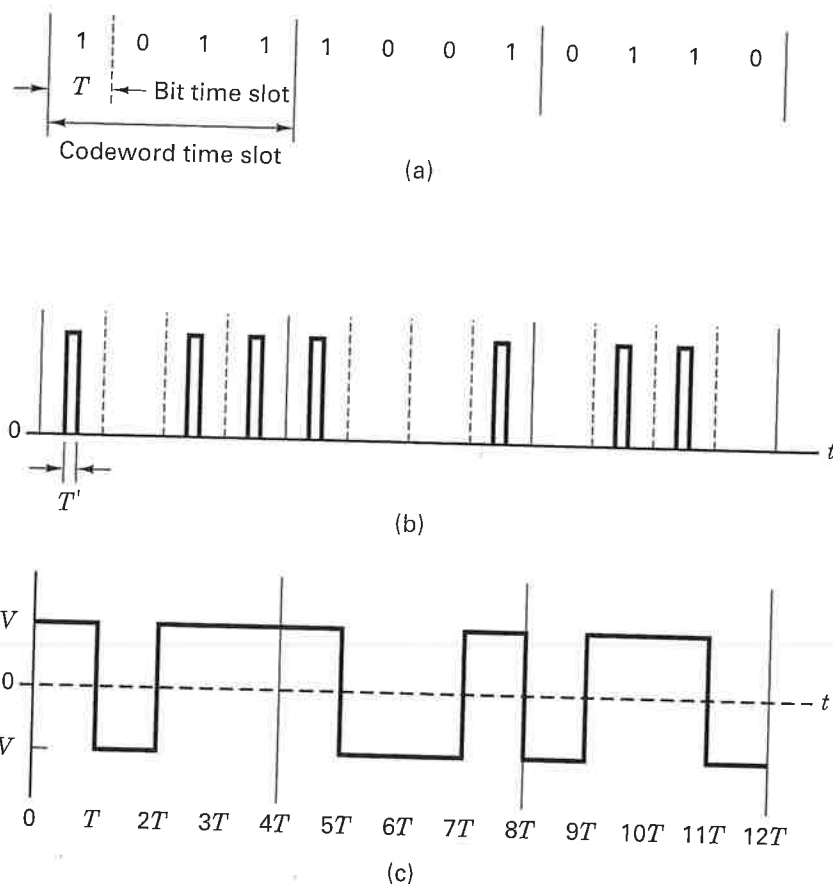
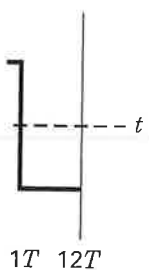
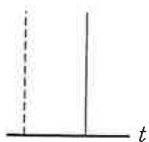


Figure 2.21 Example of waveform representation of binary digits. (a) PCM sequence. (b) Pulse representation of PCM. (c) Pulse waveform (transition between two levels).

pulse-modulation waveform, of which there are several types. They are described in Section 2.8.5, where one of them, called pulse-amplitude modulation (PAM), is emphasized. In Figure 2.1, the highlighted block, labeled *Baseband Signaling*, shows the basic classification of the PCM waveforms and the M -ary pulse waveforms. The PCM waveforms fall into the following four groups.

1. Nonreturn-to-zero (NRZ)
2. Return-to-zero (RZ)
3. Phase encoded
4. Multilevel binary

The NRZ group is probably the most commonly used PCM waveform. It can be partitioned into the following subgroups: NRZ-L (L for level), NRZ-M (M for mark), and NRZ-S (S for space). NRZ-L is used extensively in digital logic circuits.



digits.
wave-

y are described
tion (PAM), is
and Signaling,
ry pulse wave-

veform. It can
NRZ-M (M for
logic circuits.

ion Chap. 2

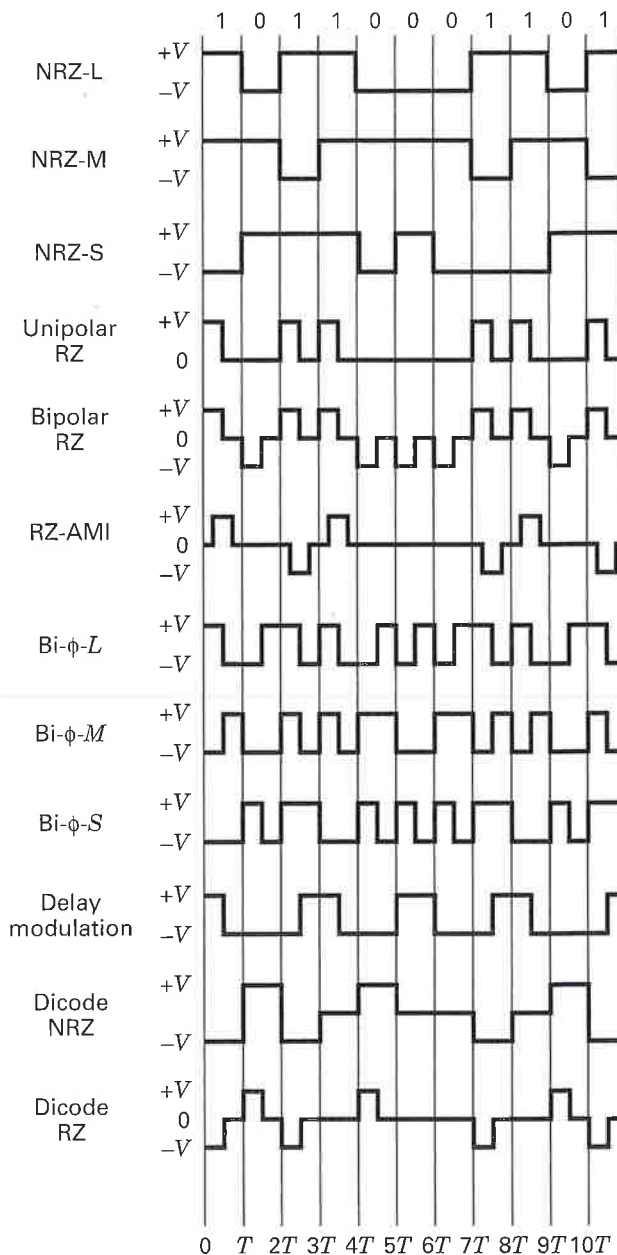


Figure 2.22 Various PCM waveforms.

A binary one is represented by one voltage level and a binary zero is represented by another voltage level. There is a change in level whenever the data change from a one to a zero or from a zero to a one. With NRZ-M, the one, or *mark*, is represented by a change in level, and the zero, or *space*, is represented by no change in level. This is often referred to as *differential encoding*. NRZ-M is used primarily in

magnetic tape recording. NRZ-S is the complement of NRZ-M: A one is represented by no change in level, and a zero is represented by a change in level.

The RZ waveforms consist of unipolar-RZ, bipolar-RZ, and RZ-AMI. These codes find application in baseband data transmission and in magnetic recording. With unipolar-RZ, a one is represented by a half-bit-wide pulse, and a zero is represented by the absence of a pulse. With bipolar-RZ, the ones and zeros are represented by opposite-level pulses that are one-half bit wide. There is a pulse present in each bit interval. RZ-AMI (AMI for "alternate mark inversion") is a signaling scheme used in telephone systems. The ones are represented by equal-amplitude alternating pulses. The zeros are represented by the absence of pulses.

The phase-encoded group consists of bi- ϕ -L (bi-phase-level), better known as *Manchester coding*; bi- ϕ -M (bi-phase-mark); bi- ϕ -S (bi-phase-space); and *delay modulation* (DM), or *Miller coding*. The phase-encoding schemes are used in magnetic recording systems and optical communications and in some satellite telemetry links. With bi- ϕ -L, a one is represented by a half-bit-wide pulse positioned during the first half of the bit interval; a zero is represented by a half-bit-wide pulse positioned during the second half of the bit interval. With bi- ϕ -M, a transition occurs at the beginning of every bit interval. A one is represented by a second transition one-half bit interval later; a zero is represented by no second transition. With bi- ϕ -S, a transition also occurs at the beginning of every bit interval. A one is represented by no second transition; a zero is represented by a second transition one-half bit interval later. With delay modulation [4], a one is represented by a transition at the midpoint of the bit interval. A zero is represented by no transition, unless it is followed by another zero. In this case, a transition is placed at the end of the bit interval of the first zero. Reference to the illustration in Figure 2.22 should help to make these descriptions clear.

Many binary waveforms use three levels, instead of two, to encode the binary data. Bipolar RZ and RZ-AMI belong to this group. The group also contains formats called *dicode* and *duobinary*. With dicode-NRZ, the one-to-zero or zero-to-one data transition changes the pulse polarity; without a data transition, the zero level is sent. With dicode-RZ, the one-to-zero or zero-to-one transition produces a half-duration polarity change; otherwise, a zero level is sent. The three-level duobinary signaling scheme is treated in Section 2.9.

One might ask why there are so many PCM waveforms. Are there really so many unique applications necessitating such a variety of waveforms to represent digits? The reason for the large selection relates to the differences in performance that characterize each waveform [5]. In choosing a PCM waveform for a particular application, some of the parameters worth examining are the following:

1. *Dc component.* Eliminating the dc energy from the signal's power spectrum enables the system to be ac coupled. Magnetic recording systems, or systems using transformer coupling, have little sensitivity to very low frequency signal components. Thus low-frequency information could be lost.
2. *Self-Clocking.* Symbol or bit synchronization is required for any digital communication system. Some PCM coding schemes have inherent synchronizing

or clocking features that aid in the recovery of the clock signal. For example, the Manchester code has a transition in the middle of every bit interval whether a one or a zero is being sent. This guaranteed transition provides a clocking signal.

3. *Error detection.* Some schemes, such as duobinary, provide the means of detecting data errors without introducing additional error-detection bits into the data sequence.
4. *Bandwidth compression.* Some schemes, such as multilevel codes, increase the efficiency of bandwidth utilization by allowing a reduction in required bandwidth for a given data rate; thus there is more information transmitted per unit bandwidth.
5. *Differential encoding.* This technique is useful because it allows the polarity of differentially encoded waveforms to be inverted without affecting the data detection. In communication systems where waveforms sometimes experience inversion, this is a great advantage. (Differential encoding is treated in greater detail in Chapter 4, Section 4.5.2.)
6. *Noise immunity.* The various PCM waveform types can be further characterized by probability of bit error versus signal-to-noise ratio. Some of the schemes are more immune than others to noise. For example, the NRZ waveforms have better error performance than does the unipolar RZ waveform.

2.8.3 Spectral Attributes of PCM Waveforms

The most common criteria used for comparing PCM waveforms and for selecting one waveform type from the many available are spectral characteristics, bit synchronization capabilities, error-detecting capabilities, interference and noise immunity, and cost and complexity of implementation. Figure 2.23 shows the spectral characteristics of some of the most popular PCM waveforms. The figure plots power spectral density in watts/hertz versus normalized bandwidth, WT , where W is bandwidth, and T is the duration of the pulse. WT is often referred to as the *time-bandwidth product*, of the signal. Since the pulse or symbol rate R_s is the reciprocal of T , normalized bandwidth can also be expressed as W/R_s . From this latter expression, we see that the units of normalized bandwidth are hertz/(pulse/s) or hertz/(symbol/s). This is a relative measure of bandwidth; it is valuable because it describes how efficiently the transmission bandwidth is being utilized for each waveform of interest. Any waveform type that requires less than 1.0 Hz for sending 1 symbol/s is relatively bandwidth efficient. Examples would be delay modulation and duobinary (see Section 2.9). By comparison, any waveform type that requires more than 1.0 Hz for sending 1 symbol/s is relatively bandwidth inefficient. An example of this would be bi-phase (Manchester) signaling. From Figure 2.23, we can also see the spectral concentration of signaling energy for each waveform type. For example, NRZ and duobinary schemes have large spectral components at dc and low frequency, while bi-phase has no energy at dc.

An important parameter for measuring *bandwidth efficiency* is R/W having units of bits/s/hz. This measure involves data rate rather than symbol rate. For a

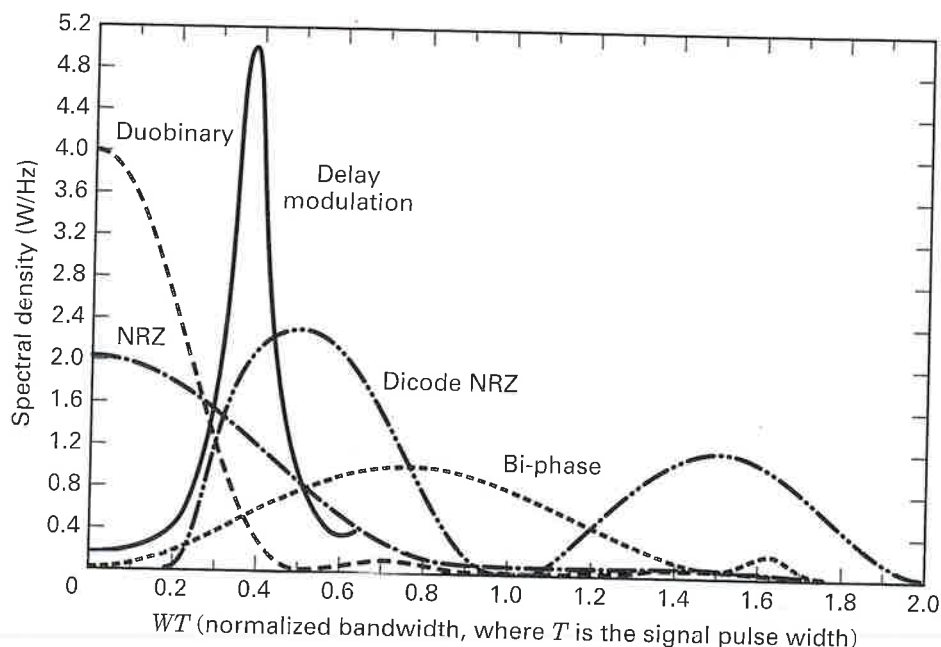


Figure 2.23 Spectral densities of various PCM waveforms.

given signaling scheme, R/W describes how much data throughput can be transmitted for each Hertz of available bandwidth. (Bandwidth efficiency is treated in greater detail in Chapter 9.)

2.8.4 Bits per PCM Word and Bits per Symbol

Throughout Chapters 1 and 2, the idea of binary partitioning ($M = 2^k$) is used to relate the grouping of bits to form symbols for the purpose of signal processing and transmission. We now examine an analogous application where the $M = 2^k$ concept is also applicable. Consider the process of formatting analog information into a bit stream via sampling, quantization, and coding. Each analog sample is transformed into a PCM word made up of groups of bits. The PCM word size can be described by the number of quantization levels allowed for each sample; this is identical to the number of values that the PCM word can assume. Or, the quantization can be described by the number of bits required to identify that set of levels. The relationship between the number of levels per sample and the number of bits needed to represent those levels is the same as the $M = 2^k$ relationship between the size of a set of message symbols and the number of bits needed to represent the symbol. To distinguish between the two applications, the notation is changed for the PCM case. Instead of $M = 2^k$, we use $L = 2^\ell$, where L is the number of quantization levels in the PCM word, and ℓ is the number of bits needed to represent those levels.

2.8.4.1 PCM Word Size

How many bits shall we assign to each analog sample? For digital telephone channels, each speech sample is PCM encoded using 8 bits, yielding 2^8 or 256 levels per sample. The choice of the number of levels, or bits per sample, depends on how much quantization distortion we are willing to tolerate with the PCM format. It is useful to develop a general relationship between the required number of bits per analog sample (the PCM word size), and the allowable quantization distortion. Let the magnitude of the quantization distortion error, $|e|$, be specified as a fraction p of the peak-to-peak analog voltage V_{pp} as follows:

$$|e| \leq p V_{pp} \quad (2.24)$$

Since the quantization error can be no larger than $q/2$, where q is the quantile interval, we can write

$$|e|_{\max} = \frac{q}{2} = \frac{V_{pp}}{2(L-1)} \approx \frac{V_{pp}}{2L} \quad (2.25)$$

where L is the number of quantization levels. For most applications the number of levels is large enough so that $L-1$ can be replaced by L , as was done above. Then, from Equations (2.24) and (2.25), we can write

$$\frac{V_{pp}}{2L} \leq p V_{pp} \quad (2.26)$$

$$2^\ell = L \geq \frac{1}{2p} \quad \text{levels} \quad (2.27)$$

and

$$\ell \geq \log_2 \frac{1}{2p} \quad \text{bits} \quad (2.28)$$

It is important that we do not confuse the idea of bits per PCM word, denoted by ℓ in Equation (2.28), with the M -level transmission concept of k data bits per symbol. (Example 2.3, presented shortly, should clarify the distinction.)

2.8.5 M -ary Pulse-Modulation Waveforms

There are three basic ways to modulate information on to a sequence of pulses: we can vary the pulse's amplitude, position, or duration, which leads to the names *pulse-amplitude modulation* (PAM), *pulse-position modulation* (PPM), and *pulse-duration modulation* (PDM), respectively. PDM is sometimes called pulse-width modulation (PWM). When information samples without any quantization are modulated on to pulses, the resulting pulse modulation can be called *analog pulse modulation*. When the information samples are first quantized, yielding symbols from an M -ary alphabet set, and then modulated on to pulses, the resulting pulse modulation is digital and we refer to it as *M -ary pulse modulation*. In the case of M -ary PAM, one of M allowable amplitude levels are assigned to each of the M possible symbol values. Earlier we described PCM waveforms as binary waveforms having

two amplitude values (e.g., NRZ, RZ). Note that such PCM waveforms requiring only two levels represent the special case ($M = 2$) of the general M -ary PAM that requires M levels. In this book, the PCM waveforms are grouped separately (see Figure 2.1 and Section 2.8.2) and are emphasized because they are the most popular of the pulse-modulation schemes.

In the case of M -ary PPM waveforms, modulation is effected by delaying (or advancing) a pulse occurrence, by an amount that corresponds to the value of the information symbols. In the case of M -ary PDM waveforms, modulation is effected by varying the pulse width by an amount that corresponds to the value of the symbols. For both PPM and PDM, the pulse amplitude is held constant. Baseband modulation with pulses have analogous counterparts in the area of bandpass modulation. PAM is similar to amplitude modulation, while PPM and PDM are similar to phase and frequency modulation respectively. In this section, we only address M -ary PAM waveforms as they compare to PCM waveforms.

The transmission bandwidth required for binary digital waveforms such as PCM may be very large. What might we do to reduce the required bandwidth? One possibility is to use *multilevel signaling*. Consider a bit stream with data rate, R bits per second. Instead of transmitting a pulse waveform for each bit, we might first partition the data into k -bit groups, and then use ($M = 2^k$)-level pulses for transmission. With such multilevel signaling or M -ary PAM, each pulse waveform can now represent a k -bit symbol in a symbol stream moving at the rate of R/k symbols per second (a factor k slower than the bit stream). Thus for a given data rate, multilevel signaling, where $M > 2$, can be used to reduce the number of symbols transmitted per second; or, in other words, M -ary PAM as opposed to binary PCM can be used to reduce the transmission bandwidth requirements of the channel. Is there a price to be paid for such bandwidth reduction? Of course, and that is discussed below.

Consider the task that the pulse receiver must perform: It must distinguish between the possible levels of each pulse. Can the receiver distinguish among the eight possible levels of each octal pulse in Figure 2.24a as easily as it can distinguish between the two possible levels of each binary pulse in Figure 2.24b? The transmission of an 8-level (compared with a 2-level) pulse requires a greater amount of energy for equivalent detection performance. (It is the amount of received E_b/N_0 that determines how reliably a signal will be detected). For equal average power in the binary and the octal pulses, it is easier to detect the binary pulses because the detector has more signal energy per level for making a binary decision than an 8-level decision. What price does a system designer pay if he or she chooses the transmission waveform to be the easier-to-detect binary PCM rather than the 8-level PAM? The engineer pays the price of needing three times as much transmission bandwidth for a given data rate, compared with the octal pulses, since each octal pulse must be replaced with three binary pulses (each one-third as wide as the octal pulses). One might ask, Why not use binary pulses with the same pulse duration as the original octal pulses and suffer the information delay? For some cases, this might be appropriate, but for real-time communication systems, such an increase in delay cannot be tolerated—the 6 o'clock news *must* be received at 6 o'clock. (In Chapter 9, we examine in detail the trade-off between signal power and transmission bandwidth.)

forms requiring any PAM that separately (see the most popular

y delaying (or the value of the ion is effected ue of the sym- ant. Baseband ndpass modu- M are similar only address

forms such as ndwidth? One ata rate, R bits we might first s for transmis- form can now k symbols per ate, multilevel ls transmitted M can be used s there a price ssed below.

ust distinguish ish among the an distinguish

The transmis- ter amount of eceived E_b/N_0 rage power in es because the cision that an ie chooses the ther than the as much trans- ses, since each as wide as the ie same pulse lay? For some systems, such be received at n signal power

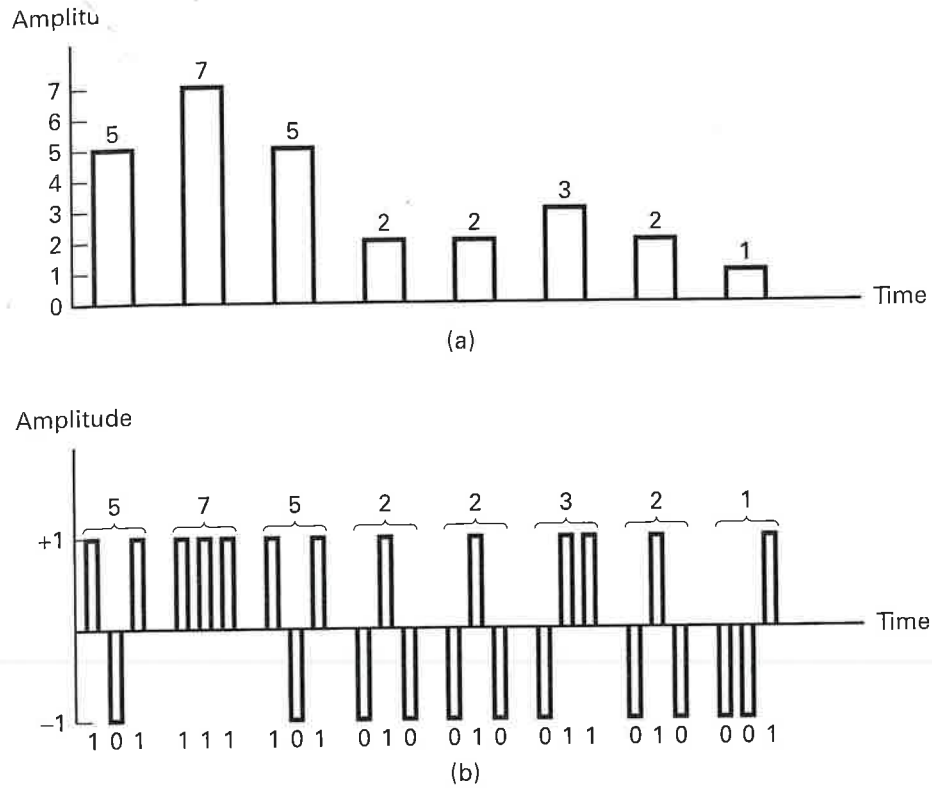


Figure 2.24 Pulse code modulation signaling. (a) Eight-level signaling. (b) Two-level signaling.

Example 2.3 Quantization Levels and Multilevel Signaling

The information in an analog waveform, with maximum frequency $f_m = 3$ kHz, is to be transmitted over an M -ary PAM system, where the number of pulse levels is $M = 16$. The quantization distortion is specified not to exceed $\pm 1\%$ of the peak-to-peak analog signal.

- What is the minimum number of bits/sample, or bits/PCM word that should be used in digitizing the analog waveform?
- What is the minimum required sampling rate, and what is the resulting bit transmission rate?
- What is the PAM pulse or symbol transmission rate?
- If the transmission bandwidth (including filtering) equals 12 kHz, determine the bandwidth efficiency for this system.

In this example we are concerned with two types of *levels*: the number of quantization levels for fulfilling the distortion requirement and the 16 levels of the multilevel PAM pulses.

Solution

(a) Using Equation (2.28), we calculate

$$\ell \geq \log_2 \frac{1}{0.02} = \log_2 50 \approx 5.6.$$

Therefore, use $\ell = 6$ bits/sample to meet the distortion requirement.

- (b) Using the Nyquist sampling criterion, the minimum sampling rate $f_s = 2f_m = 6000$ samples/second. From part (a), each sample will give rise to a PCM word composed of 6 bits. Therefore the bit transmission rate $R = \ell f_s = 36,000$ bits/sec.
- (c) Since multilevel pulses are to be used with $M = 2^k = 16$ levels, then $k = \log_2 16 = 4$ bits/symbol. Therefore, the bit stream will be partitioned into groups of 4 bits to form the new 16-level PAM digits, and the resulting symbol transmission rate R_s is $R/k = 36,000/4 = 9000$ symbols/s.
- (d) Bandwidth efficiency is described by data throughput per hertz, R/W . Since $R = 36,000$ bits/s, and $W = 12$ kHz, then $R/W = 3$ bits/s/hz.

2.9 CORRELATIVE CODING

In 1963, Adam Lender [6, 7] showed that it is possible to transmit $2W$ symbols/s with zero ISI, using the theoretical minimum bandwidth of W hertz, without infinitely sharp filters. Lender used a technique called *duobinary signaling*, also referred to as *correlative coding* and *partial response signaling*. The basic idea behind the duobinary technique is to introduce some controlled amount of ISI into the data stream rather than trying to eliminate it completely. By introducing correlated interference between the pulses, and by changing the detection procedure, Lender, in effect, "canceled out" the interference at the detector and thereby achieved the ideal symbol-rate packing of 2 symbols/s/Hz, an amount that had been considered unrealizable.

2.9.1 Duobinary Signaling

To understand how duobinary signaling introduces controlled ISI, let us look at a model of the process. We can think of the duobinary coding operation as if it were implemented as shown in Figure 2.25. Assume that a sequence of binary symbols $\{x_k\}$ is to be transmitted at the rate of R symbols/s over a system having an ideal rectangular spectrum of bandwidth $W = R/2 = 1/2T$ hertz. You might ask: How is this rectangular spectrum, in Figure 2.25, different from the unrealizable Nyquist characteristic? It has the same ideal characteristic; but we are not trying to implement the ideal rectangular filter. It is only the part of our equivalent model that is used for developing a filter that is easier to approximate. Before being shaped by the ideal filter, the pulses pass through a simple digital filter, as shown in the figure. The digital filter incorporates a one-digit delay; to each incoming pulse, the filter adds the value of the previous pulse. In other words, for every pulse into the digital filter, we get the summation of two pulses out. Each pulse of the sequence $\{y_k\}$ out of the digital filter can be expressed as

$$y_k = x_k + x_{k-1} \quad (2.29)$$

rement.
 ing rate $f_s = 2f_m = 6000$
 to a PCM word com-
 36,000 bits/sec.
 s, then $k = \log_2 16 = 4$
 to groups of 4 bits to
 transmission rate R_s is

er hertz, R/W . Since

smi $2W$ symbols/s
 hertz, without infi-
 signaling, also re-
 e basic idea behind
 int of ISI into the
 oducing correlated
 procedure, Lender,
 reby achieved the
 d been considered

II, let us look at a
 ration as if it were
 of binary symbols
 n having an ideal
 night ask: How is
 ealizable Nyquist
 t trying to imple-
 lent model that is
 e being shaped by
 own in the figure.
 g pulse, the filter
 se into the digital
 sequence $\{y_k\}$ out

(2.29)

lation Chap. 2

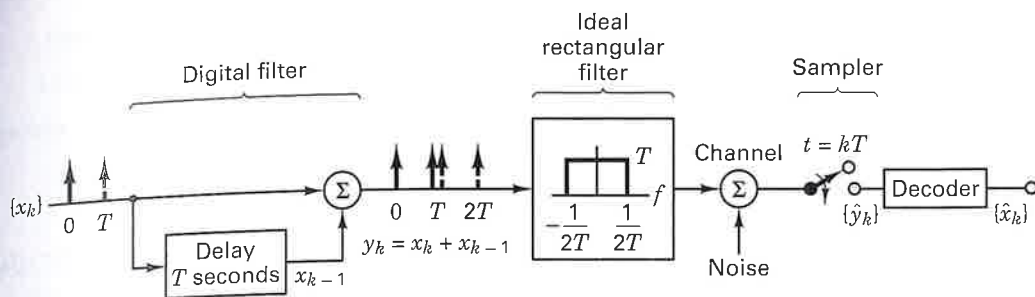


Figure 2.25 Duobinary signaling.

Hence, the $\{y_k\}$ amplitudes are not independent; each y_k digit carries with it the *memory* of the prior digit. The ISI introduced to each y_k digit comes only from the *preceding* x_{k-1} digit. This correlation between the pulse amplitudes of $\{y_k\}$ can be *thought of* as the controlled ISI introduced by the duobinary coding. Controlled interference is the essence of this novel technique because at the detector, such controlled interference can be removed as easily as it was added. The $\{y_k\}$ sequence is followed by the ideal Nyquist filter that does not introduce any ISI. In Figure 2.25, at the receiver sampler, we would expect to recover the sequence $\{y_k\}$ exactly in the absence of noise. Since all systems experience noise contamination, we shall refer to the *received* $\{y_k\}$ as the estimate of $\{y_k\}$ and denote it $\{\hat{y}_k\}$. Removing the controlled interference with the duobinary decoder yields an estimate of $\{x_k\}$ which we shall denote as $\{\hat{x}_k\}$.

2.9.2 Duobinary Decoding

If the binary digit x_k is equal to ± 1 , then using Equation (2.29), y_k has one of three possible values: $+2$, 0 , or -2 . The duobinary code results in a three-level output: in general, for M -ary transmission, partial response signaling results in $2M - 1$ output levels. The decoding procedure involves the inverse of the coding procedure, namely, subtracting the x_{k-1} decision from the y_k digit. Consider the following coding/decoding example.

Example 2.4 Duobinary Coding and Decoding

Use Equation (2.29) to demonstrate duobinary coding and decoding for the following sequence: $\{x_k\} = 0010110$. Consider the first bit of the sequence to be a startup digit, not part of the data.

Solution

Binary digit sequence $\{x_k\}$:	0	0	1	0	1	1	0
Bipolar amplitudes $\{x_k\}$:	-1	-1	+1	-1	+1	+1	-1
Coding rule: $y_k = x_k + x_{k-1}$:	-2	0	0	0	2	0	

Decoding decision rule:

If $\hat{y}_k = 2$, decide that $\hat{x}_k = +1$ (or binary one).

If $\hat{y}_k = -2$, decide that $\hat{x}_k = -1$ (or binary zero).

If $\hat{y}_k = 0$, decide opposite of the previous decision.

Decoded bipolar sequence $\{\hat{x}_k\}$: -1 +1 -1 +1 +1 -1

Decoded binary sequence $\{\hat{x}_k\}$: 0 1 0 1 1 0

The decision rule simply implements the subtraction of each $\hat{x}_{k-1}$ decision from each $\hat{y}_k$. One drawback of this detection technique is that once an error is made, it tends to propagate, causing further errors, since present decisions depend on prior decisions. A means of avoiding this error propagation is known as *precoding*.

2.9.3 Precoding

Precoding is accomplished by first differentially encoding the $\{x_k\}$ binary sequence into a new $\{w_k\}$ binary sequence by means of the equation:

$$w_k = x_k \oplus w_{k-1} \quad (2.30)$$

where the symbol $\oplus$ represents modulo-2 addition (equivalent to the logical *exclusive-or* operation) of the binary digits. The rules of modulo-2 addition are as follows:

$$0 \oplus 0 = 0$$

$$0 \oplus 1 = 1$$

$$1 \oplus 0 = 1$$

$$1 \oplus 1 = 0$$

The $\{w_k\}$ binary sequence is then converted to a bipolar pulse sequence, and the coding operation proceeds in the same way as it did in Example 2.4. However, with precoding, the detection process is quite different from the detection of ordinary duobinary, as shown below in Example 2.5: The precoding model is shown in Figure 2.26; in this figure it is implicit that the modulo-2 addition producing the precoded $\{w_k\}$ sequence is performed on the *binary* digits, while the digital filtering producing the $\{y_k\}$ sequence is performed on the *bipolar* pulses.

Example 2.5 Duobinary Precoding

Illustrate the duobinary coding and decoding rules when using the differential precoding of Equation (2.30). Assume the same $\{x_k\}$ sequence as that given in Example 2.4.

Solution

Binary digit sequence $\{x_k\}$: 0 0 1 0 1 1 0

Precoded sequence $w_k = x_k \oplus w_{k-1}$: 0 0 1 1 0 1 1

Bipolar sequence $\{w_k\}$: -1 -1 +1 +1 -1 +1 +1

Coding rule: $y_k = w_k + w_{k-1}$: -2 0 +2 0 0 +2

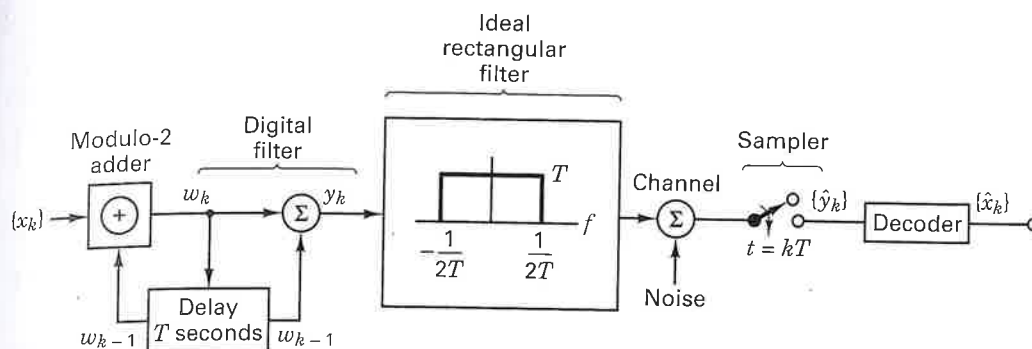


Figure 2.26 Precoded duobinary signaling.

Decoding decision rule:

If $\hat{y}_k = \pm 2$, decide that $\hat{x}_k = \text{binary zero}$.
 If $\hat{y}_k = 0$, decide that $\hat{x}_k = \text{binary one}$.

Decoded binary sequence $\{x_k\}$:

0 1 0 1 1 0

The differential precoding enables us to decode the $\{\hat{y}_k\}$ sequence by making a decision on each received sample singly, without resorting to prior decisions that could be in error. The major advantage is that in the event of a digit error due to noise, such an error does not propagate to other digits. Notice that the first bit in the differentially precoded binary sequence $\{w_k\}$ is an arbitrary choice. If the startup bit in $\{w_k\}$ had been chosen to be a binary one instead of a binary zero, the decoded result would have been the same.

2.9.4 Duobinary Equivalent Transfer Function

In Section 2.9.1, we described the duobinary transfer function as a digital filter incorporating a one-digit delay followed by an ideal rectangular transfer function. Let us now examine an equivalent model. The Fourier transform of a delay can be described as $e^{-j2\pi fT}$ (see Section A.3.1); therefore, the input digital filter of Figure 2.25 can be characterized as the frequency transfer function

$$H_1(f) = 1 + e^{-j2\pi fT} \quad (2.31)$$

The transfer function of the ideal rectangular filter, is

$$H_2(f) = \begin{cases} T & \text{for } |f| < \frac{1}{2T} \\ 0 & \text{elsewhere} \end{cases} \quad (2.32)$$

The overall equivalent transfer function of the digital filter cascaded with the ideal rectangular filter is then given by

$$\begin{aligned}
H_e(f) &= H_1(f)H_2(f) \quad \text{for } |f| < \frac{1}{2T} \\
&= (1 + e^{-j2\pi fT})T \\
&= T(e^{j\pi fT} + e^{-j\pi fT})e^{-j\pi fT}
\end{aligned} \tag{2.33}$$

so that

$$|H_e(f)| = \begin{cases} 2T \cos \pi fT & \text{for } |f| < \frac{1}{2T} \\ 0 & \text{elsewhere} \end{cases} \tag{2.34}$$

Thus $H_e(f)$, the composite transfer function for the cascaded digital and rectangular filters, has a gradual roll-off to the band edge, as can be seen in Figure 2.27a. The transfer function can be approximated by using realizable analog filtering; a separate digital filter is not needed. The duobinary equivalent $H_e(f)$ is called a *cosine filter* [8]. The cosine filter should not be confused with the *raised cosine filter* (described in Chapter 3, Section 3.3.1). The corresponding impulse response $h_e(t)$, found by taking the inverse Fourier transform of $H_e(f)$ in Equation (2.33) is

$$h_e(t) = \text{sinc}\left(\frac{t}{T}\right) + \text{sinc}\left(\frac{t - T}{T}\right) \tag{2.35}$$

and is plotted in Figure 2.27b. For every impulse $\delta(t)$ at the input of Figure 2.25, the output is $h_e(t)$ with an appropriate polarity. Notice that there are only two nonzero samples at T -second intervals, giving rise to controlled ISI from the adjacent bit. The introduced ISI is eliminated by use of the decoding procedure discussed in Section 2.9.2. Although the cosine filter is noncausal and therefore nonrealizable, it can be easily approximated. The implementation of the precoded duobinary technique described in Section 2.9.3 can be accomplished by first differentially encoding the binary sequence $\{x_k\}$ into the sequence $\{w_k\}$ (see Example 2.5). The pulse sequence $\{w_k\}$ is then filtered by the equivalent cosine characteristic described in Equation (2.34).

2.9.5 Comparison of Binary with Duobinary Signaling

The duobinary technique introduces correlation between pulse amplitudes, whereas the more restrictive Nyquist criterion assumes that the transmitted pulse amplitudes are independent of one another. We have shown that duobinary signaling can exploit this introduced correlation to achieve zero ISI signal transmission, using a smaller system bandwidth than is otherwise possible. Do we get this performance improvement without paying a price? No, such is rarely the case with engineering design options—there is almost always a trade-off involved. We saw that duobinary coding requires three levels, compared with the usual two levels for binary coding. Recall our discussion in Section 2.8.5, where we compared the performance and the required signal power for making eight-level PAM decisions versus two-level (PCM) decisions. For a fixed amount of signal power, the ease of making reliable decisions is inversely related to the number of levels that must be

(2.33)

(2.34)

l and rectangu-
n Figure 2.27a.
alog filtering; a
 f is called a
sed cosine filter
response $h_e(t)$,
(2.33) is

(2.35)

Figure 2.25, the
ly two nonzero
ne adjacent bit.
re discussed in
nonrealizable, it
duobinary tech-
entially encod-
2.5). The pulse
tic described in

lse amplitudes,
ansmitted pulse
uobinary signal-
al transmission,
get this perfor-
the case with
volved. We saw
al two levels for
compared the
PAM decisions
wer, the ease of
els that must be

ation Chap. 2

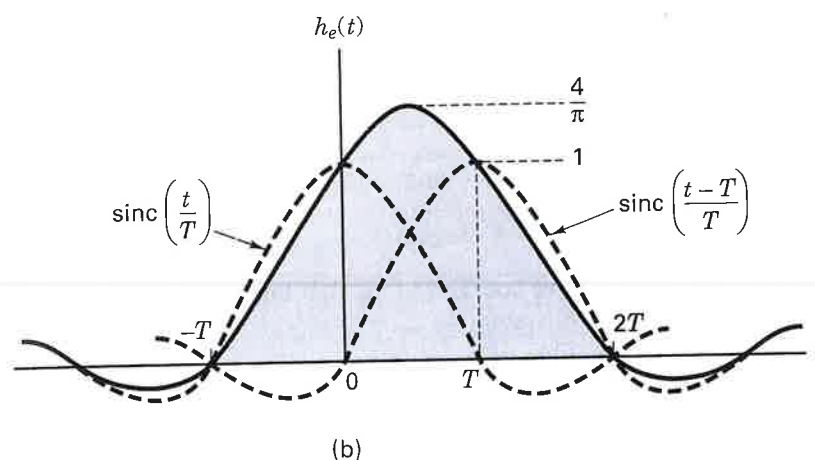
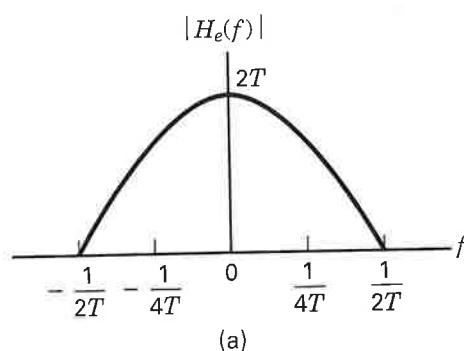


Figure 2.27 Duobinary transfer function and pulse shape. (a) Cosine filter. (b) Impulse response of the cosine filter.

distinguished in each waveform. Therefore, it should be no surprise that although duobinary signaling accomplishes the zero ISI requirement with minimum bandwidth, duobinary signaling also requires more power than binary signaling, for equivalent performance against noise. For a given probability of bit error (P_B), duobinary signaling requires approximately 2.5 dB greater SNR than binary signaling, while using only $1/(1+r)$ the bandwidth that binary signaling requires [7], where r is the filter roll-off.

2.9.6 Polybinary Signaling

Duobinary signaling can be extended to more than three digits or levels, resulting in greater bandwidth efficiency; such systems are called *polybinary* [7, 9]. Consider that a binary message with two signaling levels is transformed into a signal with j signaling levels numbered consecutively from zero to $(j-1)$. The transformation from binary to polybinary takes place in two steps. First, the original sequence $\{x_k\}$, consisting of binary ones and zeros, is converted into another binary sequence $\{y_k\}$,

as follows. The present binary digit of sequence $\{y_k\}$ is formed from the modulo-2 addition of the $(j - 2)$ immediately preceding digits of sequence $\{y_k\}$ and the present digit x_k . For example, let

$$y_k = x_k \oplus y_{k-1} \oplus y_{k-2} \oplus y_{k-3} \quad (2.36)$$

Here x_k represents the input binary digit and y_k the k th encoded digit. Since the expression involves $(j - 2) = 3$ bits preceding y_k , there are $j = 5$ signaling levels. Next, the binary sequence $\{y_k\}$ is transformed into a polybinary pulse train $\{z_k\}$ by adding algebraically the present bit of sequence $\{y_k\}$ to the $(j - 2)$ preceding bits of $\{y_k\}$. Therefore, $z_k \text{ modulo-2} = x_k$, and the binary elements one and zero are mapped into even- and odd-valued pulses in the sequence $\{z_k\}$. Note that each digit in $\{z_k\}$ can be independently detected despite the strong correlation between bits. The primary advantage of such a signaling scheme is the redistribution of the spectral density of the original sequence $\{x_k\}$, so as to favor the low frequencies, thus improving system bandwidth efficiency.

2.10 CONCLUSION

In this chapter we have considered the first important step in any digital communication system, transforming the source information (both textual and analog) to a form that is compatible with a digital system. We treated various aspects of sampling, quantization (both uniform and nonuniform), and pulse code modulation (PCM). We considered the selection of pulse waveforms for the transmission of baseband signals through the channel. We also introduced the duobinary concept of adding a controlled amount of ISI to achieve an improvement in bandwidth efficiency at the expense of an increase in power.

REFERENCES

1. Black, H. S., *Modulation Theory*, D. Van Nostrand Company, Princeton, N.J., 1953.
2. Oppenheim, A. V., *Applications of Digital Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.
3. Stiltz, H., ed., *Aerospace Telemetry*, Vol. 1, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1961, p. 179.
4. Hecht, M., and Guida, A., "Delay Modulation," *Proc. IEEE*, vol. 57, no. 7, July 1969, pp. 1314-1316.
5. Deffebach, H. L., and Frost, W. O., "A Survey of Digital Baseband Signaling Techniques," *NASA Technical Memorandum NASATM X-64615*, June 30, 1971.
6. Lender, A., "The Duobinary Technique for High Speed Data Transmission," *IEEE Trans. Commun. Electron.*, vol. 82, May 1963, pp. 214-218.
7. Lender, A., "Correlative (Partial Response) Techniques and Applications to Digital Radio Systems," in K. Feher, *Digital Communications: Microwave Applications*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981, Chap. 7.

8. Couch, L. W., II, *Digital and Analog Communication Systems*, Macmillan Publishing Company, New York, 1982.
9. Lender, A., "Correlative Digital Communication Techniques," *IEEE Trans. Commun. Technol.*, Dec. 1964, pp. 128-135.

PROBLEMS

- 2.1. You want to transmit the word "HOW" using an 8-ary system.
 - (a) Encode the word "HOW" into a sequence of bits, using 7-bit ASCII coding, followed by an eighth bit for error detection, per character. The eighth bit is chosen so that the number of ones in the 8 bits is an even number. How many total bits are there in the message?
 - (b) Partition the bit stream into $k = 3$ bit segments. Represent each of the 3-bit segments as an octal number (symbol). How many octal symbols are there in the message?
 - (c) If the system were designed with 16-ary modulation, how many symbols would be used to represent the word "HOW"?
 - (d) If the system were designed with 256-ary modulation, how many symbols would be used to represent the word "HOW"?
- 2.2. We want to transmit 800 characters/s, where each character is represented by its 7-bit ASCII codeword, followed by an eighth bit for error detection, per character, as in Problem 2.1. A multilevel PAM waveform with $M = 16$ levels is used.
 - (a) What is the effective transmitted bit rate?
 - (b) What is the symbol rate?
- 2.3. We wish to transmit a 100-character alphanumeric message in 2 s, using 7-bit ASCII coding, followed by an eighth bit for error detection, per character, as in Problem 2.1. A multilevel PAM waveform with $M = 32$ levels is used.
 - (a) Calculate the effective transmitted bit rate and the symbol rate.
 - (b) Repeat part (a) for 16-level PAM, eight-level PAM, four-level PAM, and PCM (binary) waveforms.
- 2.4. Given an analog waveform that has been sampled at its Nyquist rate, f_s , using natural sampling, prove that a waveform (proportional to the original waveform) can be recovered from the samples, using the recovery techniques shown in Figure P2.1. The parameter mf_s is the frequency of the local oscillator, where m is an integer.

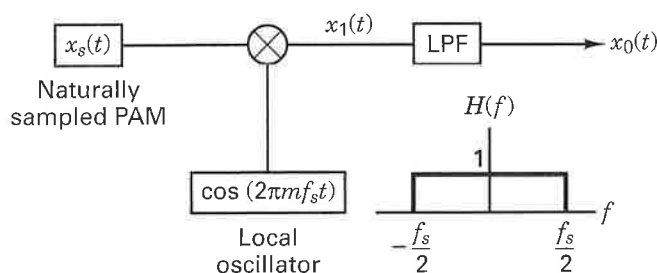


Figure P2.1

- 2.5. An analog signal is sampled at its Nyquist rate $1/T_s$, and quantized using L quantization levels. The derived digital signal is then transmitted on some channel.
- Show that the time duration, T , of one bit of the transmitted binary encoded signal must satisfy $T \leq T_s/(\log_2 L)$.
 - When is the equality sign valid?
- 2.6. Determine the number of quantization levels that are implied if the number of bits per sample in a given PCM code is (a) 5; (b) 8; (c) x .
- 2.7. Determine the minimum sampling rate necessary to sample and perfectly reconstruct the signal $x(t) = \sin(6280t)/(6280t)$.
- 2.8. Consider an audio signal with spectral components limited to the frequency band 300 to 3300 Hz. Assume that a sampling rate of 8000 samples/s will be used to generate a PCM signal. Assume that the ratio of peak signal power to average quantization noise power at the output needs to be 30 dB.
- What is the minimum number of uniform quantization levels needed, and what is the minimum number of bits per sample needed?
 - Calculate the system bandwidth (as specified by the main spectral lobe of the signal) required for the detection of such a PCM signal.
- 2.9. A waveform, $x(t) = 10 \cos(1000t + \pi/3) + 20 \cos(2000t + \pi/6)$ is to be uniformly sampled for digital transmission.
- What is the maximum allowable time interval between sample values that will ensure perfect signal reproduction?
 - If we want to reproduce 1 hour of this waveform, how many sample values need to be stored?
- 2.10. (a) A waveform that is bandlimited to 50 kHz is sampled every 10 μ s. Show graphically that these samples uniquely characterize the waveform. (Use a sinusoidal example for simplicity. Avoid sampling at points where the waveform equals zero.)
- (b) If samples are taken 30 μ s apart instead of 10 μ s, show graphically that waveforms other than the original can be characterized by the samples.
- 2.11. Use the method of convolution to illustrate the effect of undersampling the waveform $x(t) = \cos 2\pi f_0 t$ for a sampling rate of $f_s = \frac{3}{2} f_0$.
- 2.12. Aliasing will not occur if the sampling rate is greater than twice the signal bandwidth. However, perfectly bandlimited signals do not occur in nature. Hence, there is always some aliasing present.
- Suppose that a filtered signal has a spectrum described by a Butterworth filter with order $n = 6$, and upper cutoff frequency $f_u = 1000$ Hz. What sampling rate is required so that aliasing is reduced to the -50 dB point in the power spectrum?
 - Repeat for a Butterworth filter with order $n = 12$.
- 2.13. (a) Sketch the complete $\mu = 10$ compression characteristic that will handle input voltages in the range -5 to +5 V.
- (b) Plot the corresponding expansion characteristic.
- (c) Draw a 16-level nonuniform quantizer characteristic that corresponds to the $\mu = 10$ compression characteristic.
- 2.14. The information in an analog waveform, whose maximum frequency $f_m = 4000$ Hz, is to be transmitted using a 16-level PAM system. The quantization distortion must not exceed $\pm 1\%$ of the peak-to-peak analog signal.
- What is the minimum number of bits per sample or bits per PCM word that should be used in this PAM transmission system?
 - What is the minimum required sampling rate, and what is the resulting bit rate?
 - What is the 16-ary PAM symbol transmission rate?

- 2.15. A signal in the frequency range 300 to 3300 Hz is limited to a peak-to-peak swing of 10 V. It is sampled at 8000 samples/s and the samples are quantized to 64 evenly spaced levels. Calculate and compare the bandwidths and ratio of peak signal power to rms quantization noise if the quantized samples are transmitted either as binary pulses or as four-level pulses. Assume that the system bandwidth is defined by the main spectral lobe of the signal.
- 2.16. In the compact disc (CD) digital audio system, an analog signal is digitized so that the ratio of the peak-signal power to the peak-quantization noise power is at least 96 dB. The sampling rate is 44.1 kilosamples/s.
- How many quantization levels of the analog signal are needed for $(S/N_q)_{\text{peak}} = 96$ dB?
 - How many bits per sample are needed for the number of levels found in part (a)?
 - What is the data rate in bits/s?
- 2.17. Calculate the difference in required signal power between two PCM waveforms, unipolar RZ and bipolar RZ (see Figure 2.22), assuming that each signaling scheme has the same requirements for data-rate and bit-error probability. Also assume equally likely signaling, and that the difference between the high-voltage and low-voltage levels is the same for both schemes. If there is a power advantage in using one of the signaling schemes, what, if any, is the disadvantage in using it?
- 2.18. In the year 1962, AT&T first offered digital telephone transmission referred to as T1 service. With this service, each T1 frame is partitioned into 24 channels or time slots. Each time slot contains 8 bits (one speech sample), and there is one additional bit per frame for alignment. The frame is sampled at the Nyquist rate of 8000 samples/s, and the bandwidth used for transmitting the composite signal is 386 kHz. Find the bandwidth efficiency (bits/s/Hz) for this signaling scheme.
- 2.19. (a) Consider that you desire a digital transmission system, such that the quantization distortion of any audio source does not exceed $\pm 2\%$ of the peak-to-peak analog signal voltage. If the audio signal bandwidth and the allowable transmission bandwidth are each 4000 Hz, and sampling takes place at the Nyquist rate, what value of bandwidth efficiency (bits/s/Hz) is required?
- (b) Repeat part (a) except that the audio signal bandwidth is 20 kHz (high fidelity), yet the available transmission bandwidth is still 4000 Hz.

QUESTIONS

- What are the similarities and differences between the terms "formatting" and "source coding"? (See Chapter 2, introduction.)
- In the process of *formatting* information, why is it often desirable to perform *oversampling*? (See Section 2.4.3.)
- In using pulse code modulation (PCM) for digitizing analog information, explain how the parameters *fidelity*, *bandwidth*, and *time delay* can be traded off. (See Section 2.6.)
- Why is it often preferred to use units of normalized bandwidth, WT (or time-bandwidth product), compared with bandwidth alone? (See Section 2.8.3.)

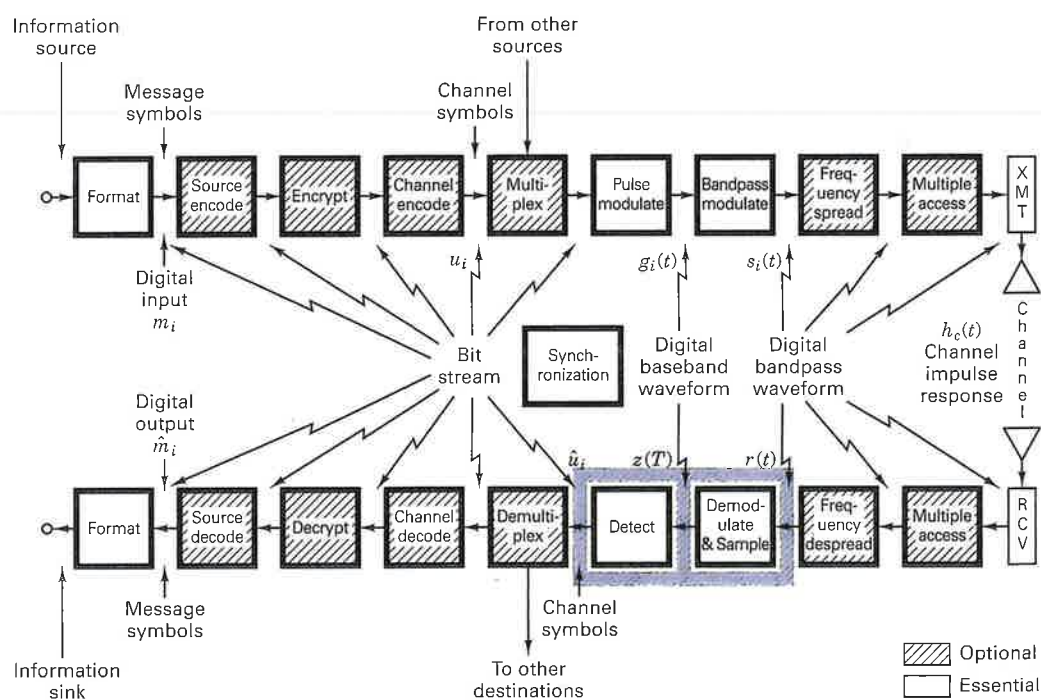
EXERCISES

Using the Companion CD, run the exercises associated with Chapter 2.

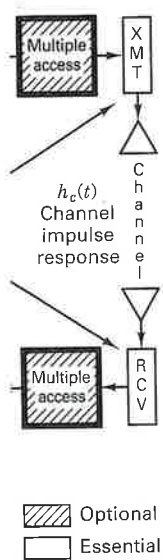
Questions

CHAPTER 3

Baseband Demodulation/Detection



n



In the case of baseband signaling, the received waveforms are already in a pulse-like form. One might ask, why then, is a demodulator needed to recover the pulse waveforms? The answer is that the arriving baseband pulses are not in the form of ideal pulse shapes, each one occupying its own symbol interval. The filtering at the transmitter and the channel typically cause the received pulse sequence to suffer from intersymbol interference (ISI) and thus appear as an amorphous "smeared" signal, not quite ready for sampling and detection. The goal of the demodulator (receiving filter) is to recover a baseband pulse with the best possible signal-to-noise ratio (SNR), free of any ISI. Equalization, covered in this chapter, is a technique used to help accomplish this goal. The equalization process is not required for every type of communication channel. However, since equalization embodies a sophisticated set of signal-processing techniques, making it possible to compensate for channel-induced interference, it is an important area for many systems.

The bandpass model of the detection process, covered in Chapter 4, is virtually identical to the baseband model considered in this chapter. That is because a received bandpass waveform is first transformed to a baseband waveform before the final detection step takes place. For linear systems, the mathematics of detection is unaffected by a shift in frequency. In fact, we can define an *equivalence theorem* as follows: Performing bandpass linear signal processing followed by heterodyning the signal to baseband, yields the same results as heterodyning the bandpass signal to baseband, followed by baseband linear signal processing. The term "heterodyning" refers to a frequency *conversion* or *mixing* process that yields

a spectral shift in the signal. As a result of this equivalence theorem, all linear signal-processing simulations can take place at baseband (which is preferred for simplicity) with the same results as at bandpass. This means that the performance of most digital communication systems will often be described and analyzed as if the transmission channel is a baseband channel.

3.1 SIGNALS AND NOISE

3.1.1 Error-Performance Degradation in Communication Systems

The task of the detector is to retrieve the bit stream from the received waveform, as error free as possible, notwithstanding the impairments to which the signal may have been subjected. There are two primary causes for error-performance degradation. The first is the effect of filtering at the transmitter, channel, and receiver, discussed in Section 3.3, below. As described there, a nonideal system transfer function causes symbol “smearing” or *intersymbol interference* (ISI).

Another cause for error-performance degradation is electrical noise and interference produced by a variety of sources, such as galaxy and atmospheric noise, switching transients, intermodulation noise, as well as interfering signals from other sources. (These are discussed in Chapter 5.) With proper precautions, much of the noise and interference entering a receiver can be reduced in intensity or even eliminated. However, there is one noise source that cannot be eliminated, and that is the noise caused by the thermal motion of electrons in any conducting media. This motion produces *thermal noise* in amplifiers and circuits, and corrupts the signal in an additive fashion. The statistics of thermal noise have been developed using quantum mechanics, and are well known [1].

The primary statistical characteristic of thermal noise is that the noise amplitudes are distributed according to a normal or Gaussian distribution, discussed in Section 1.5.5, and shown in Figure 1.7. In this figure, it can be seen that the most probable noise amplitudes are those with small positive or negative values. In theory, the noise can be infinitely large, but very large noise amplitudes are rare. The primary spectral characteristic of thermal noise in communication systems, is that its two-sided power spectral density $G_n(f) = N_0/2$ is flat for all frequencies of interest. In other words, the thermal noise, on the average, has just as much power per hertz in low-frequency fluctuations as in high-frequency fluctuations—up to a frequency of about 10^{12} hertz. When the noise power is characterized by such a constant-power spectral density, we refer to it as *white noise*. Since thermal noise is present in all communication systems and is the predominant noise source for many systems, the thermal noise characteristics (additive, white, and Gaussian, giving rise to the name AWGN) are most often used to model the noise in the detection process and in the design of receivers. Whenever a channel is designated as an AWGN channel (with no other impairments specified), we are in effect being told that its impairments are limited to the degradation caused by this unavoidable thermal noise.

3.1.2 Demodulation and Detection

During a given signaling interval T , a binary baseband system will transmit one of two waveforms, denoted $g_1(t)$ and $g_2(t)$. Similarly, a binary bandpass system will transmit one of two waveforms, denoted $s_1(t)$ and $s_2(t)$. Since the general treatment of demodulation and detection are essentially the same for baseband and bandpass systems, we use $s_i(t)$ here as a generic designation for a transmitted waveform, whether the system is baseband or bandpass. This allows much of the baseband demodulation/detection treatment in this chapter to be consistent with similar bandpass descriptions in Chapter 4. Then, for any binary channel, the transmitted signal over a symbol interval $(0, T)$ is represented by

$$s_i(t) = \begin{cases} s_1(t) & 0 \leq t \leq T & \text{for a binary 1} \\ s_2(t) & 0 \leq t \leq T & \text{for a binary 0} \end{cases}$$

The received signal $r(t)$ degraded by noise $n(t)$ and possibly degraded by the impulse response of the channel $h_c(t)$ was described in Equation (1.1) and is rewritten as

$$r(t) = s_i(t) * h_c(t) + n(t) \quad i = 1, \dots, M \quad (3.1)$$

where $n(t)$ is here assumed to be a zero mean AWGN process, and $*$ represents a convolution operation. For binary transmission over an ideal distortionless channel where convolution with $h_c(t)$ produces no degradation (since for the ideal case $h_c(t)$ is an impulse function), the representation of $r(t)$ can be simplified to

$$r(t) = s_i(t) + n(t) \quad i = 1, 2, \quad 0 \leq t \leq T \quad (3.2)$$

Figure 3.1 shows the typical demodulation and detection functions of a digital receiver. Some authors use the terms “demodulation” and “detection” interchangeably. This book makes a distinction between the two. We define *demodulation* as recovery of a waveform (to an undistorted baseband pulse), and we designate *detection* to mean the decision-making process of selecting the digital meaning of that waveform. If error-correction coding is *not* present, the detector output consists of estimates of message symbols (or bits), $\hat{m}_i$ (also called *hard decisions*). If error-correction coding is used, the detector output consists of estimates of channel symbols (or coded bits) $\hat{u}_i$, which can take the form of *hard* or *soft decisions* (see Section 7.3.2). For brevity, the term “detection” is occasionally used loosely to encompass all the receiver signal-processing steps through the decision making step. The *frequency down-conversion* block, shown in the demodulator portion of Figure 3.1, performs frequency translation for bandpass signals operating at some radio frequency (RF). This function may be configured in a variety of ways. It may take place within the front end of the receiver, within the demodulator, shared between the two locations, or not at all.

Within the *demodulate* and *sample* block of Figure 3.1 is the *receiving filter* (essentially the demodulator), which performs waveform recovery in preparation for the next important step—detection. The filtering at the transmitter and the channel typically cause the received pulse sequence to suffer from ISI, and thus it is

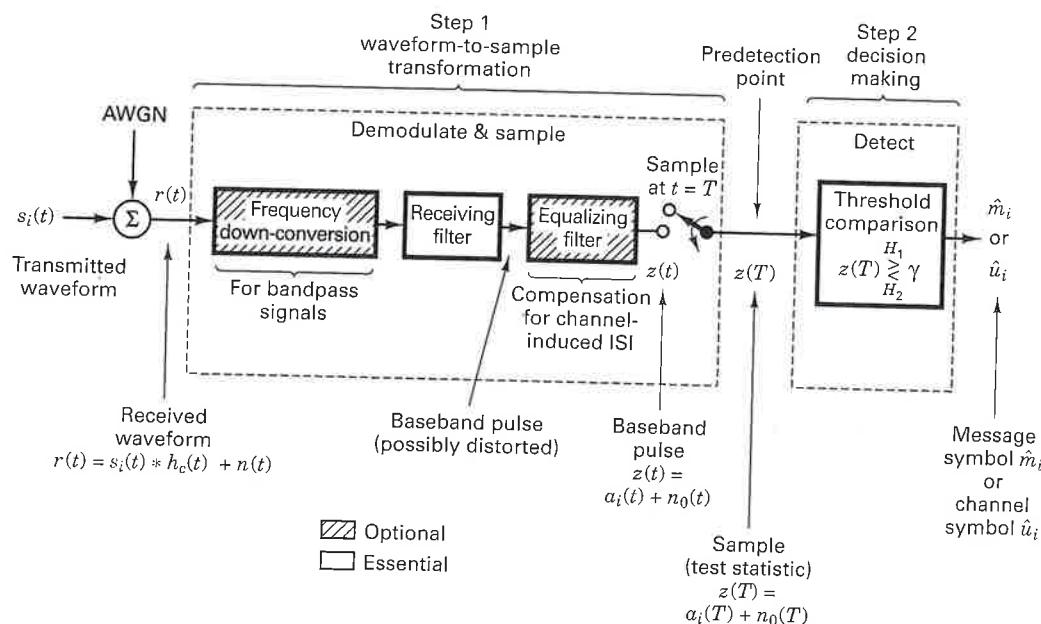
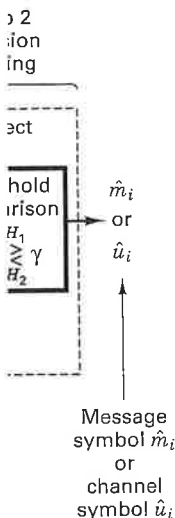


Figure 3.1 Two basic steps in the demodulation/detection of digital signals.

not quite ready for sampling and detection. The goal of the receiving filter is to recover a baseband pulse with the best possible signal-to-noise ratio (SNR), free of any ISI. The optimum receiving filter for accomplishing this is called a *matched filter* or *correlator*, described in Sections 3.2.2 and 3.2.3. An optional *equalizing filter* follows the receiving filter; it is only needed for those systems where channel-induced ISI can distort the signals. The receiving filter and equalizing filter are shown as two separate blocks in order to emphasize their separate functions. In most cases, however, when an equalizer is used, a single filter would be designed to incorporate both functions and thereby *compensate* for the distortion caused by both the transmitter and the channel. Such a *composite* filter is sometimes referred to simply as the *equalizing filter* or the *receiving and equalizing filter*.

Figure 3.1 highlights two steps in the demodulation/detection process. Step 1, the waveform-to-sample transformation, is made up of the demodulator followed by a sampler. At the end of each symbol duration T , the output of the sampler, the *predetection point*, yields a sample $z(T)$, sometimes called the test statistic. $z(T)$ has a voltage value directly proportional to the energy of the received symbol and that of the noise. In step 2, a decision (detection) is made regarding the digital meaning of that sample. We assume that the input noise is a Gaussian random process and that the receiving filter in the demodulator is linear. A linear operation performed on a Gaussian random process will produce a second Gaussian random process [2]. Thus, the filter output noise is Gaussian. The output of step 1 yields the test statistic

$$z(T) = a_i(T) + n_0(T) \quad i = 1, 2 \quad (3.3)$$



where $a_i(T)$ is the desired signal component, and $n_0(T)$ is the noise component. To simplify the notation, we sometimes express Equation (3.3) in the form of $z = a_i + n_0$. The noise component n_0 is a zero mean Gaussian random variable, and thus $z(T)$ is a Gaussian random variable with a mean of either a_1 or a_2 depending on whether a binary one or binary zero was sent. As described in Section 1.5.5, the probability density function (pdf) of the Gaussian random noise n_0 can be expressed as

$$p(n_0) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{n_0}{\sigma_0} \right)^2 \right] \quad (3.4)$$

where σ_0^2 is the noise variance. Thus it follows from Equations (3.3) and (3.4) that the conditional pdfs $p(z|s_1)$ and $p(z|s_2)$ can be expressed as

$$p(z|s_1) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{z - a_1}{\sigma_0} \right)^2 \right] \quad (3.5)$$

and

$$p(z|s_2) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{z - a_2}{\sigma_0} \right)^2 \right] \quad (3.6)$$

These conditional pdfs are illustrated in Figure 3.2. The rightmost conditional pdf, $p(z|s_1)$, called the *likelihood* of s_1 , illustrates the probability density function of the random variable $z(T)$, given that symbol s_1 was transmitted. Similarly, the leftmost conditional pdf, $p(z|s_2)$, called the *likelihood* of s_2 , illustrates the pdf of $z(T)$, given that symbol s_2 was transmitted. The abscissa, $z(T)$, represents the full range of possible sample output values from step 1 of Figure 3.1.

After a received waveform has been transformed to a sample, the actual shape of the waveform is no longer important; all waveform types that are transformed to the same value of $z(T)$ are identical for detection purposes. Later it is shown that an optimum receiving filter (matched filter) in step 1 of Figure 3.1 maps all signals of equal energy into the same point $z(T)$. Therefore, the received *signal energy* (not its shape) is the important parameter in the detection process. This is why the detection analysis for baseband signals is the same as that for bandpass sig-

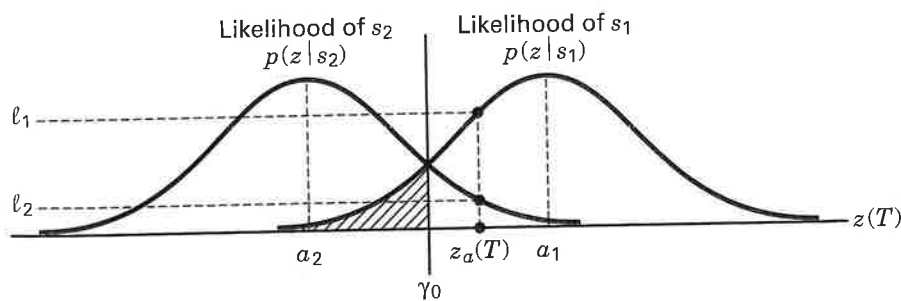


Figure 3.2 Conditional probability density functions: $p(z|s_1)$ and $p(z|s_2)$.

(3.3)

nals. Since $z(T)$ is a voltage signal that is proportional to the energy of the received symbol, the larger the magnitude of $z(T)$, the more error free will be the decision-making process. In step 2, detection is performed by choosing the hypothesis that results from the threshold measurement

$$z(T) \underset{H_2}{\overset{H_1}{\geq}} \gamma \quad (3.7)$$

where H_1 and H_2 are the two possible (binary) hypotheses. The inequality relationship indicates that hypothesis H_1 is chosen if $z(T) > \gamma$, and hypothesis H_2 is chosen if $z(T) < \gamma$. If $z(T) = \gamma$, the decision can be an arbitrary one. Choosing H_1 is equivalent to deciding that signal $s_1(t)$ was sent and hence a binary 1 is detected. Similarly, choosing H_2 is equivalent to deciding that signal $s_2(t)$ was sent, and hence a binary 0 is detected.

3.1.3 A Vectorial View of Signals and Noise

We now present a geometric or vectorial view of signal waveforms that are useful for either baseband or bandpass signals. We define an N -dimensional *orthogonal space* as a space characterized by a set of N linearly independent functions $\{\psi_j(t)\}$, called *basis functions*. Any arbitrary function in the space can be generated by a linear combination of these basis functions. The basis functions must satisfy the conditions

$$\int_0^T \psi_j(t) \psi_k(t) dt = K_j \delta_{jk} \quad 0 \leq t \leq T \quad j, k = 1, \dots, N \quad (3.8a)$$

where the operator

$$\delta_{jk} = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{otherwise} \end{cases} \quad (3.8b)$$

is called the *Kronecker delta function* and is defined by Equation (3.8b). When the K_j constants are nonzero, the signal space is called *orthogonal*. When the basis functions are normalized so that each $K_j = 1$, the space is called an *orthonormal* space. The principal requirement for orthogonality can be stated as follows. Each $\psi_j(t)$ function of the set of basis functions must be independent of the other members of the set. Each $\psi_j(t)$ must not interfere with any other members of the set in the detection process. From a geometric point of view, each $\psi_j(t)$ is mutually perpendicular to each of the other $\psi_k(t)$ for $j \neq k$. An example of such a space with $N = 3$ is shown in Figure 3.3, where the mutually perpendicular axes are designated $\psi_1(t)$, $\psi_2(t)$, and $\psi_3(t)$. If $\psi_j(t)$ corresponds to a real-valued voltage or current waveform component, associated with a $1\text{-}\Omega$ resistive load, then using Equations (1.5) and (3.8), the normalized energy in joules dissipated in the load in T seconds, due to ψ_j , is

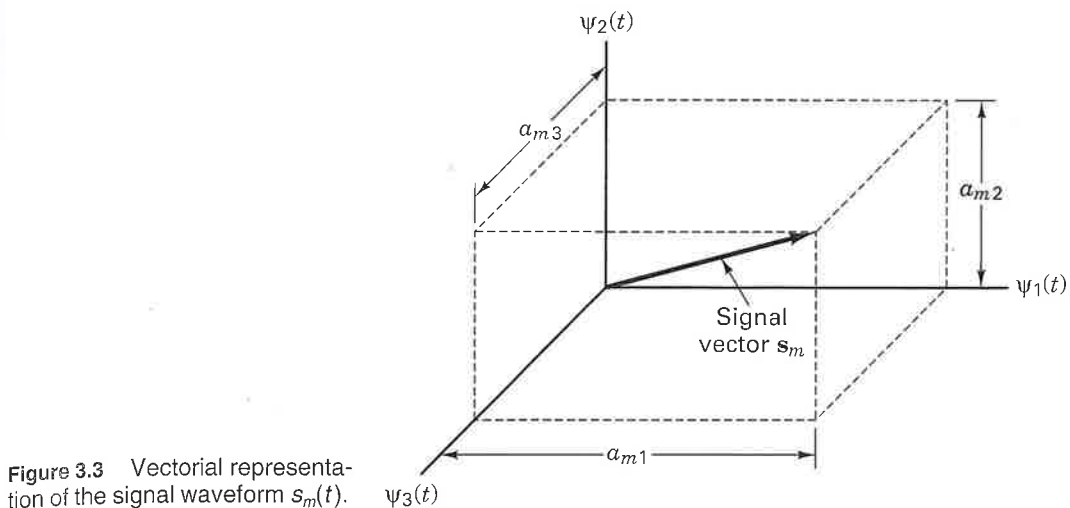


Figure 3.3 Vectorial representation of the signal waveform $s_m(t)$.

$$E_j = \int_0^T \psi_j^2(t) dt = K_j \quad (3.9)$$

One reason we focus on an *orthogonal signal space* is that Euclidean distance measurements, fundamental to the detection process, are easily formulated in such a space. However, even if the signaling waveforms do not make up such an orthogonal set, they can be transformed into linear combinations of orthogonal waveforms. It can be shown [3] that *any arbitrary* finite set of waveforms $\{s_i(t)\}$ ($i = 1, \dots, M$), where each member of the set is physically realizable and of duration T , can be expressed as a linear combination of N orthogonal waveforms $\psi_1(t), \psi_2(t), \dots, \psi_N(t)$, where $N \leq M$, such that

$$\begin{aligned} s_1(t) &= a_{11}\psi_1(t) + a_{12}\psi_2(t) + \dots + a_{1N}\psi_N(t) \\ s_2(t) &= a_{21}\psi_1(t) + a_{22}\psi_2(t) + \dots + a_{2N}\psi_N(t) \\ &\vdots \\ s_M(t) &= a_{M1}\psi_1(t) + a_{M2}\psi_2(t) + \dots + a_{MN}\psi_N(t) \end{aligned}$$

These relationships are expressed in more compact notation as

$$s_i(t) = \sum_{j=1}^N a_{ij} \psi_j(t) \quad \begin{matrix} i = 1, \dots, M \\ N \leq M \end{matrix} \quad (3.10)$$

where

$$a_{ij} = \frac{1}{K_j} \int_0^T s_i(t) \psi_j(t) dt \quad \begin{matrix} i = 1, \dots, M \\ j = 1, \dots, N \end{matrix} \quad 0 \leq t \leq T \quad (3.11)$$

The coefficient a_{ij} is the value of the $\psi_j(t)$ component of signal $s_i(t)$. The form of the $\{\psi_j(t)\}$ is not specified; it is chosen for convenience and will depend on the form of the signal waveforms. The set of signal waveforms, $\{s_i(t)\}$, can be viewed as a set of

vectors, $\{s_i\} = \{a_{i1}, a_{i2}, \dots, a_{iN}\}$. If, for example, $N = 3$, we may plot the vector s_m corresponding to the waveform

$$s_m(t) = a_{m1}\psi_1(t) + a_{m2}\psi_2(t) + a_{m3}\psi_3(t)$$

as a point in a three-dimensional Euclidean space with coordinates (a_{m1}, a_{m2}, a_{m3}) , as shown in Figure 3.3. The orientation among the signal vectors describes the relation of the signals to one another (with respect to phase or frequency), and the amplitude of each vector in the set $\{s_i\}$ is a measure of the signal energy transmitted during a symbol duration. In general, once a set of N orthogonal functions has been adopted, each of the transmitted signal waveforms, $s_i(t)$, is completely determined by the vector of its coefficients,

$$s_i = (a_{i1}, a_{i2}, \dots, a_{iN}) \quad i = 1, \dots, M \quad (3.12)$$

We shall employ the notation of signal vectors, $\{s\}$, or signal waveforms, $\{s(t)\}$, as best suits the discussion. A typical detection problem, conveniently viewed in terms of signal vectors, is illustrated in Figure 3.4. Vectors s_j and s_k represent *prototype* or *reference signals* belonging to the set of M waveforms, $\{s_i(t)\}$. The receiver knows, a priori, the location in the signal space of each prototype vector belonging to the M -ary set. During the transmission of any signal, the signal is perturbed by noise so that the resultant vector that is actually received is a perturbed version (e.g., $s_j + \mathbf{n}$ or $s_k + \mathbf{n}$) of the original one, where $\mathbf{n}$ represents a noise vector. The noise is **additive** and has a Gaussian distribution; therefore, the resulting distribution of possible received signals is a cluster or cloud of points around s_j and s_k . The cluster is dense in the center and **becomes** sparse with increasing distance from the prototype. The arrow marked " $\mathbf{r}$ " represents a signal vector that might arrive at the receiver during some symbol interval. The task of the receiver is to decide whether $\mathbf{r}$ has a close "resem-

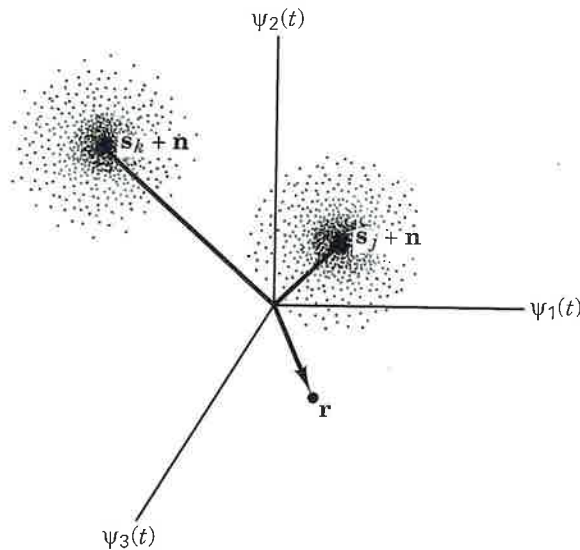


Figure 3.4 Signals and noise in a three-dimensional vector space.

blance" to the prototype s_j , whether it more closely resembles s_k , or whether it is closer to some other prototype signal in the M -ary set. The measurement can be thought of as a *distance measurement*. The receiver or detector must decide which of the prototypes within the signal space is *closest* in distance to the received vector $\mathbf{r}$. The analysis of all demodulation or detection schemes involves this concept of distance between a received waveform and a set of possible transmitted waveforms. A simple rule for the detector to follow is to decide that $\mathbf{r}$ belongs to the same class as its nearest neighbor (nearest prototype vector).

3.1.3.1 Waveform Energy

Using Equations (1.5), (3.10), and (3.8), the normalized energy E_i , associated with the waveform $s_i(t)$ over a symbol interval T can be expressed in terms of the orthogonal components of $s_i(t)$ as follows:

$$E_i = \int_0^T s_i^2(t) dt = \int_0^T \left[\sum_j a_{ij} \psi_j(t) \right]^2 dt \quad (3.13)$$

$$= \int_0^T \sum_j a_{ij} \psi_j(t) \sum_k a_{ik} \psi_k(t) dt \quad (3.14)$$

$$= \sum_j \sum_k a_{ij} a_{ik} \int_0^T \psi_j(t) \psi_k(t) dt \quad (3.15)$$

$$= \sum_j \sum_k a_{ij} a_{ik} K_{jk} \delta_{jk} \quad (3.16)$$

$$= \sum_{j=1}^N a_{ij}^2 K_j \quad i = 1, \dots, M \quad (3.17)$$

Equation (3.17) is a special case of Parseval's theorem relating the integral of the square of the waveform $s_i(t)$ to the sum of the square of the orthogonal series coefficients. If orthonormal functions are used (i.e., $K_j = 1$), the normalized energy over a symbol duration T is given by

$$E_i = \sum_{j=1}^N a_{ij}^2 \quad (3.18)$$

If there is equal energy E in each of the $s_i(t)$ waveforms, we can write Equation (3.18) in the form

$$E = \sum_{j=1}^N a_{ij}^2 \quad \text{for all } i \quad (3.19)$$

3.1.3.2 Generalized Fourier Transforms

The transformation described by Equations (3.8), (3.10), and (3.11) is referred to as the *generalized Fourier transformation*. In the case of ordinary Fourier transforms, the $\{\psi_j(t)\}$ set comprises sine and cosine harmonic functions. But in the

case of generalized Fourier transforms, the $\{\psi_j(t)\}$ set is not constrained to any specific form; it must only satisfy the orthogonality statement of Equation (3.8). Any arbitrary integrable waveform set, as well as noise, can be represented as a linear combination of orthogonal waveforms through such a generalized Fourier transformation [3]. Therefore, in such an orthogonal space, we are justified in using distance (Euclidean distance) as a decision criterion for the detection of *any* signal set in the presence of AWGN. The most important application of this orthogonal transformation has to do with the way in which signals are actually transmitted and received. The transmission of a nonorthogonal signal set is generally accomplished by the appropriate weighting of the orthogonal carrier components.

Example 3.1 Orthogonal Representation of Waveforms

Figure 3.5 illustrates the statement that any arbitrary integrable waveform set can be represented as a linear combination of orthogonal waveforms. Figure 3.5a shows a set of three waveforms, $s_1(t)$, $s_2(t)$, and $s_3(t)$.

- Demonstrate that these waveforms *do not* form an orthogonal set.
- Figure 3.5b shows a set of two waveforms, $\psi_1(t)$ and $\psi_2(t)$. Verify that these waveforms form an orthogonal set.
- Show how the nonorthogonal waveform set in part (a) can be expressed as a linear combination of the orthogonal set in part (b).
- Figure 3.5c illustrates another orthogonal set of two waveforms, $\psi'_1(t)$ and $\psi'_2(t)$. Show how the nonorthogonal set in Figure 3.5a can be expressed as a linear combination of the set in Figure 3.5c.

Solution

- $s_1(t)$, $s_2(t)$, and $s_3(t)$ are clearly not orthogonal, since they do not meet the requirements of Equation (3.8); that is, the time integrated value (over a symbol duration) of the product of any two of the three waveforms is not zero. Let us verify this for $s_1(t)$ and $s_2(t)$:

$$\begin{aligned} \int_0^T s_1(t)s_2(t) dt &= \int_0^{T/2} s_1(t)s_2(t) dt + \int_{T/2}^T s_1(t)s_2(t) dt \\ &= \int_0^{T/2} (-1)(2) dt + \int_{T/2}^T (-3)(0) dt = -T \end{aligned}$$

Similarly, the integral over the interval T of each of the cross-products $s_1(t)s_3(t)$ and $s_2(t)s_3(t)$ results in nonzero values. Hence, the waveform set $\{s_i(t)\}$ ($i = 1, 2, 3$) in Figure 3.5a is not an orthogonal set.

- Using Equation (3.8), we verify that $\psi_1(t)$ and $\psi_2(t)$ form an orthogonal set as follows:

$$\int_0^T \psi_1(t)\psi_2(t) dt = \int_0^{T/2} (1)(1) dt + \int_{T/2}^T (-1)(1) dt = 0$$

- Using Equation (3.11) with $K_j = T$, we can express the nonorthogonal set $\{s_i(t)\}$ ($i = 1, 2, 3$) as a linear combination of the orthogonal basis waveforms $\{\psi_j(t)\}$ ($j = 1, 2$):

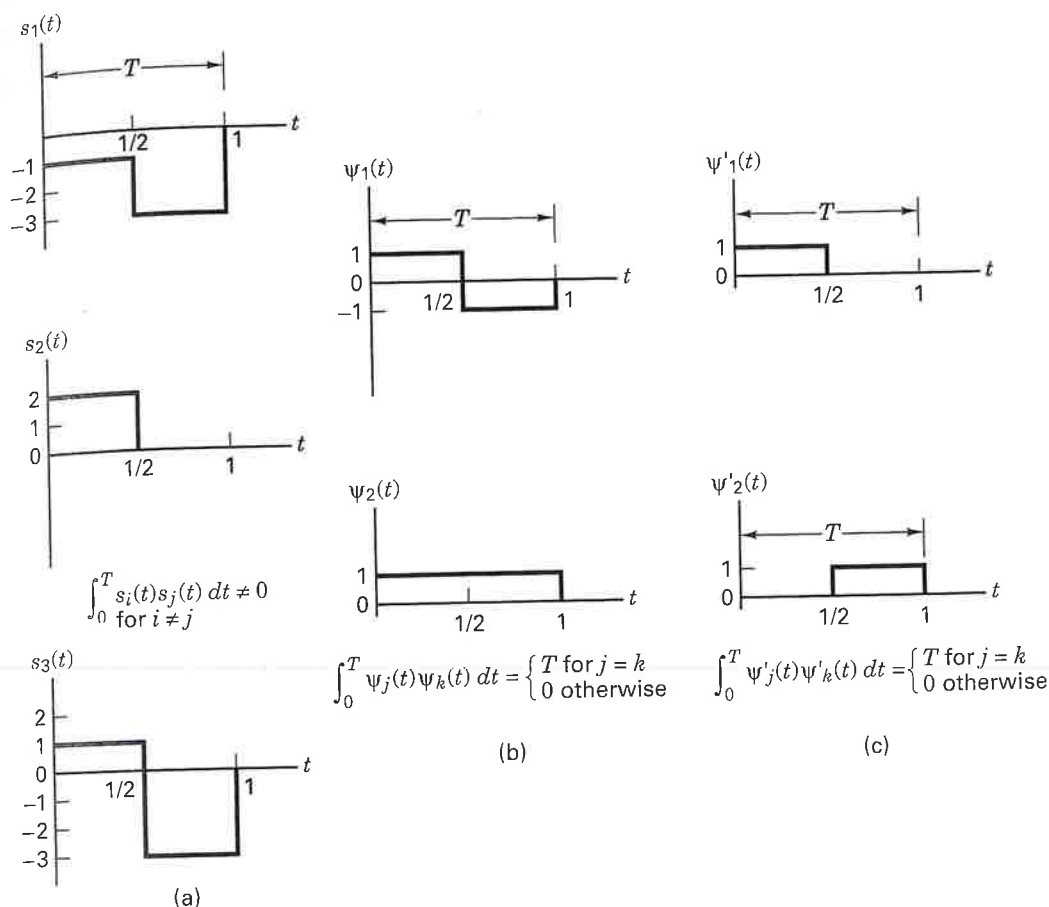


Figure 3.5 Example of an arbitrary signal set in terms of an orthogonal set. (a) Arbitrary signal set. (b) A set of orthogonal basis functions. (c) Another set of orthogonal basis functions.

$$s_1(t) = \psi_1(t) - 2\psi_2(t)$$

$$s_2(t) = \psi_1(t) + \psi_2(t)$$

$$s_3(t) = 2\psi_1(t) - \psi_2(t)$$

(d) Similar to part (c), the nonorthogonal set $\{s_i(t)\}$ ($i = 1, 2, 3$) can be expressed in terms of the simple orthogonal basis set $\{\psi'_j(t)\}$ ($j = 1, 2$) in Figure 3.5c, as follows:

$$s_1(t) = -\psi'_1(t) - 3\psi'_2(t)$$

$$s_2(t) = 2\psi'_1(t)$$

$$s_3(t) = \psi'_1(t) - 3\psi'_2(t)$$

These relationships illustrate how an arbitrary waveform set $\{s_i(t)\}$ can be expressed as a linear combination of an orthogonal set $\{\psi_j(t)\}$, as described in Equations (3.10) and (3.11). What are the practical applications of being able to describe $s_1(t)$, $s_2(t)$, and $s_3(t)$ in terms of $\psi_1(t)$, $\psi_2(t)$, and the appropriate coefficients?

cients? If we want a system for transmitting waveforms $s_1(t)$, $s_2(t)$, and $s_3(t)$, the transmitter and the receiver need only be implemented using the two basis functions $\psi_1(t)$ and $\psi_2(t)$ instead of the three original waveforms. The *Gram-Schmidt orthogonalization procedure* provides a convenient way in which an appropriate choice of a basis function set $\{\psi_j(t)\}$, can be obtained for any given signal set $\{s_i(t)\}$. (It is described in Appendix 4A of Reference [4].)

3.1.3.3 Representing White Noise with Orthogonal Waveforms

Additive white Gaussian noise (AWGN) can be expressed as a linear combination of orthogonal waveforms in the same way as signals. For the signal detection problem, the noise can be partitioned into two components,

$$n(t) = \hat{n}(t) + \tilde{n}(t) \quad (3.20)$$

where

$$\hat{n}(t) = \sum_{j=1}^N n_j \psi_j(t) \quad (3.21)$$

is taken to be the noise within the signal space, or the projection of the noise components on the signal coordinates $\psi_1(t), \dots, \psi_N(t)$, and

$$\tilde{n}(t) = n(t) - \hat{n}(t) \quad (3.22)$$

is defined as the noise outside the signal space. In other words, $\tilde{n}(t)$ may be thought of as the noise that is effectively tuned out by the detector. The symbol $\hat{n}(t)$ represents the noise that will interfere with the detection process. We can express the noise waveform $n(t)$ as

$$n(t) = \sum_{j=1}^N n_j \psi_j(t) + \tilde{n}(t) \quad (3.23)$$

where

$$n_j = \frac{1}{K_j} \int_0^T n(t) \psi_j(t) dt \quad \text{for all } j \quad (3.24)$$

and

$$\int_0^T \tilde{n}(t) \psi_j(t) dt = 0 \quad \text{for all } j \quad (3.25)$$

The interfering portion of the noise, $\hat{n}(t)$, expressed in Equation (3.21) will henceforth be referred to simply as $n(t)$. We can express $n(t)$ by a vector of its coefficients similar to the way we did for signals in Equation (3.12). We have

$$\mathbf{n} = (n_1, n_2, \dots, n_N) \quad (3.26)$$

where $\mathbf{n}$ is a random vector with zero mean and Gaussian distribution, and where the noise components n_i ($i = 1, \dots, N$) are independent.

3.1.3.4 Variance of White Noise

White noise is an *idealized process* with two-sided power spectral density equal to a constant $N_0/2$, for all frequencies from $-\infty$ to $+\infty$. Hence, the noise variance (average noise power, since the noise has zero mean) is

$$\sigma^2 = \text{var} [n(t)] = \int_{-\infty}^{\infty} \left(\frac{N_0}{2} \right) df = \infty \quad (3.27)$$

Although the variance for AWGN is infinite, the variance for *filtered* AWGN is finite. For example, if AWGN is correlated with one of a set of orthonormal functions $\psi_j(t)$, the variance of the correlator output is given by

$$\sigma^2 = \text{var} (n_j) = \mathbf{E} \left\{ \left[\int_0^T n(t) \psi_j(t) dt \right]^2 \right\} = \frac{N_0}{2} \quad (3.28)$$

The proof of Equation (3.28) is given in Appendix C. Henceforth we shall assume that the noise of interest in the **detection** process is the output noise of a correlator or matched filter with variance $\sigma^2 = N_0/2$ as expressed in Equation (3.28).

3.1.4 The Basic SNR Parameter for Digital Communication Systems

Anyone who has studied analog communications is familiar with the figure of merit, *average signal power to average noise power ratio* (S/N or SNR). In digital communications, we more often use E_b/N_0 , a normalized version of SNR, as a figure of merit. E_b is bit energy and can be described as **signal power S times the bit time T_b** . N_0 is noise power spectral density, and can be described as **noise power N divided by bandwidth W** . Since the bit time and bit rate R_b are reciprocal, we can replace T_b with $1/R_b$ and write

$$\frac{E_b}{N_0} = \frac{S T_b}{N/W} = \frac{S/R_b}{N/W} \quad (3.29)$$

Data rate, in units of bits per second, is one of the most recurring parameters in digital communications. We **therefore** simplify the notation throughout the book, by using R instead of R_b to **represent** bits/s, and we rewrite Equation (3.29) to emphasize that E_b/N_0 is just a version of S/N normalized by bandwidth and bit rate, as follows:

$$\frac{E_b}{N_0} = \frac{S}{N} \left(\frac{W}{R} \right) \quad (3.30)$$

One of the most important metrics of performance in digital communication systems is a plot of the bit-error probability P_B versus E_b/N_0 . Figure 3.6 illustrates the “waterfall-like” shape of most such curves. For $E_b/N_0 \geq x_0$, $P_B \leq P_0$. The dimensionless ratio E_b/N_0 is a standard quality measure for digital communications system performance. Therefore, required E_b/N_0 can be considered a metric that characterizes the performance of one system versus another; the smaller the required E_b/N_0 , the more efficient is the detection process for a given probability of error.

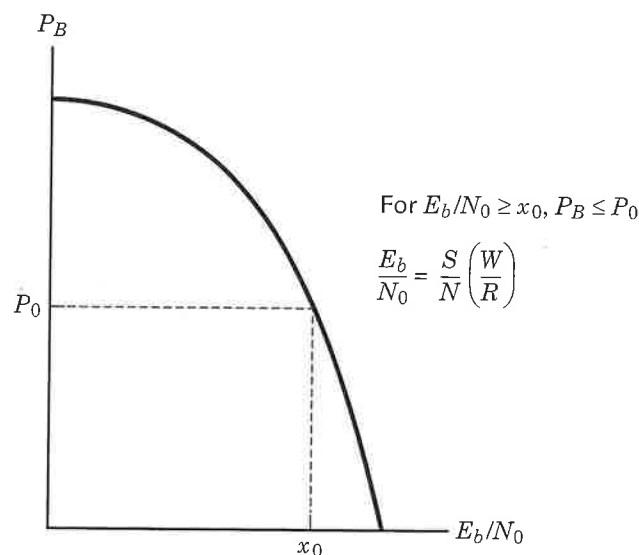


Figure 3.6 General shape of the P_B versus E_b/N_0 curve.

3.1.5 Why E_b/N_0 Is a Natural Figure of Merit

A newcomer to digital communications may question the usefulness of the parameter E_b/N_0 . After all, S/N is a useful figure of merit for analog communications—the numerator represents a power measurement of the signal we wish to preserve and deliver, and the denominator represents electrical noise degradation. Moreover, S/N is intuitively acceptable as a metric of goodness. Thus, why can't we continue to use S/N as a figure of merit for digital communications? Why has a different metric for digital systems—the ratio of bit energy to noise power spectral density—arisen? The explanation is given below.

In Section 1.2.4, a power signal was defined as a signal having finite average power and infinite energy. An energy signal was defined as a signal having zero average power and finite energy. These classifications are useful in comparing analog and digital waveforms. We classify an analog waveform as a power signal. Why does this make sense? We can think of an analog waveform as having infinite duration that need not be partitioned or windowed in time. An infinitely long electrical waveform has an infinite amount of energy; hence, energy is not a useful way to characterize this waveform. Power (or rate of delivering the energy) is a more useful parameter for analog waveforms.

However, in a digital communication system we transmit (and receive) a symbol by using a transmission waveform within a window of time, the symbol time T_s . Focusing on one symbol, we can see that the power (averaged over all time) goes to zero. Hence, power is not a useful way to characterize a digital waveform. What we need for such waveforms is a metric of the “good stuff” within the window. In other words, the symbol energy (power integrated over T_s) is a more useful parameter for characterizing digital waveforms.

The fact that a digital signal is best characterized by its received energy doesn't yet get to the crux of why E_b/N_0 is a natural metric for digital systems, so let us continue. The digital waveform is a vehicle that represents a digital message. The message may contain one bit (binary), two bits (4-ary), . . . , 10 bits (1024-ary). In analog systems, there is nothing akin to such a discretized message structure. An analog information source is an infinitely quantized continuous wave. For digital systems, a figure of merit should allow us to compare one system with another at the bit level. Therefore, a description of the digital waveform in terms of S/N is virtually useless, since the waveform may have a one-bit meaning, a two-bit meaning, or a 10-bit meaning. For example, suppose we are told that for a given error probability, the required S/N for a digital binary waveform is 20 units. Think of the waveform as being interchangeable with its meaning. Since the binary waveform has a one-bit meaning, then the S/N requirement per bit is equal to the same 20 units. However, suppose that the waveform is 1024-ary, with the same 20 units of required S/N . Now, since the waveform has a 10-bit meaning, the S/N requirement per bit is only 2 units. Why should we have to go through such computational manipulations to find a metric that represents a figure of merit? Why not immediately describe the metric in terms of what we need—an energy-related parameter at the bit level, E_b/N_0 ? Just as S/N is a dimensionless ratio, so too is E_b/N_0 . To verify this, consider the following units of measure:

$$\frac{E_b}{N_0} = \frac{\text{Joule}}{\text{Watt per Hz}} = \frac{\text{Watt-s}}{\text{Watt-s}}$$

3.2 DETECTION OF BINARY SIGNALS IN GAUSSIAN NOISE

3.2.1 Maximum Likelihood Receiver Structure

The decision-making criterion shown in step 2 of Figure 3.1 was described by Equation (3.7) as

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma$$

A popular criterion for choosing the threshold level γ for the binary decision in Equation (3.7) is based on minimizing the probability of error. The computation for this *minimum error* value of $\gamma = \gamma_0$ starts with forming an inequality expression between the ratio of conditional probability density functions and the signal a priori probabilities. Since the conditional density function $p(z|s_i)$ is also called the *likelihood* of s_i , the formulation

$$\frac{p(z|s_1)}{p(z|s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} \quad (3.31)$$

is called the *likelihood ratio test*. (See Appendix B.) In this inequality, $P(s_1)$ and $P(s_2)$ are the a priori probabilities that $s_1(t)$ and $s_2(t)$, respectively, are transmitted, and H_1 and H_2 are the two possible hypotheses. The rule for minimizing the error probability states that we should choose hypothesis H_1 if the ratio of likelihoods is greater than the ratio of a priori probabilities, as shown in Equation (3.31).

It is shown in Section B.3.1, that if $P(s_1) = P(s_2)$, and if the likelihoods, $p(z|s_i)$ ($i = 1, 2$), are symmetrical, the substitution of Equations (3.5) and (3.6) into (3.31) yields

$$z(T) \underset{H_2}{\overset{H_1}{\geq}} \frac{a_1 + a_2}{2} = \gamma_0 \quad (3.32)$$

where a_1 is the signal component of $z(T)$ when $s_1(t)$ is transmitted, and a_2 is the signal component of $z(T)$ when $s_2(t)$ is transmitted. The threshold level γ_0 , represented by $(a_1 + a_2)/2$, is the *optimum threshold* for minimizing the probability of making an incorrect decision for this important special case. This strategy is known as the *minimum error criterion*.

For equally likely signals, the optimum threshold γ_0 passes through the intersection of the likelihood functions, as shown in Figure 3.2. Thus by following Equation (3.32), the decision stage effectively selects the hypothesis that corresponds to the signal with the *maximum likelihood*. For example, given an arbitrary detector output value $z_a(T)$, for which there is a nonzero likelihood that $z_a(T)$ belongs to either signal class $s_1(t)$ or $s_2(t)$, one can think of the likelihood test as a comparison of the likelihood values $p(z_a|s_1)$ and $p(z_a|s_2)$. The signal corresponding to the maximum pdf is chosen as the most likely to have been transmitted. In other words, the detector chooses $s_1(t)$ if

$$p(z_a|s_1) > p(z_a|s_2) \quad (3.33)$$

Otherwise, the detector chooses $s_2(t)$. A detector that minimizes the error probability (for the case where the signal classes are equally likely) is also known as a *maximum likelihood detector*.

Figure 3.2 illustrates that Equation (3.33) is just a "common sense" way to make a decision when there exists statistical knowledge of the classes. Given the detector output value $z_a(T)$, we see in Figure 3.2 that $z_a(T)$ intersects the likelihood of $s_1(t)$ at a value ℓ_1 , and it intersects the likelihood of $s_2(t)$ at a value ℓ_2 . What is the most reasonable decision for the detector to make? For this example, choosing class $s_1(t)$, which has the greater likelihood, is the most sensible choice. If this was an M -ary instead of a binary example, there would be a total of M likelihood functions representing the M signal classes to which a received signal might belong. The maximum likelihood decision would then be to choose the class that had the greatest likelihood of all M likelihoods. (Refer to Appendix B for a review of decision theory fundamentals.)

3.2.1.1 Error Probability

For the binary decision-making depicted in Figure 3.2, there are two ways errors can occur. An error e will occur when $s_1(t)$ is sent, and channel noise results in the receiver output signal $z(t)$ being less than γ_0 . The probability of such an occurrence is

$$P(e|s_1) = P(H_2|s_1) = \int_{-\infty}^{\gamma_0} p(z|s_1) dz \quad (3.34)$$

This is illustrated by the shaded area to the left of γ_0 in Figure 3.2. Similarly, an error occurs when $s_2(t)$ is sent, and the channel noise results in $z(t)$ being greater than γ_0 . The probability of this occurrence is

$$P(e|s_2) = P(H_1|s_2) = \int_{\gamma_0}^{\infty} p(z|s_2) dz \quad (3.35)$$

The probability of an error is the sum of the probabilities of all the ways that an error can occur. For the binary case, we can express the probability of bit error as

$$P_B = \sum_{i=1}^2 P(e, s_i) = \sum_{i=1}^2 P(e|s_i) P(s_i) \quad (3.36)$$

Combining Equations (3.34) to (3.36), we can write

$$P_B = P(e|s_1)P(s_1) + P(e|s_2)P(s_2) \quad (3.37a)$$

or equivalently,

$$P_B = P(H_2|s_1)P(s_1) + P(H_1|s_2)P(s_2) \quad (3.37b)$$

That is, given that signal $s_1(t)$ was transmitted, an error results if hypothesis H_2 is chosen; or given that signal $s_2(t)$ was transmitted, an error results if hypothesis H_1 is chosen. For the case where the a priori probabilities are equal [that is, $P(s_1) = P(s_2) = \frac{1}{2}$],

$$P_B = \frac{1}{2} P(H_2|s_1) + \frac{1}{2} P(H_1|s_2) \quad (3.38)$$

and because of the symmetry of the probability density functions,

$$P_B = P(H_2|s_1) = P(H_1|s_2) \quad (3.39)$$

The probability of a bit error, P_B , is numerically equal to the area under the "tail" of either likelihood function, $p(z|s_1)$ or $p(z|s_2)$, falling on the "incorrect" side of the threshold. We can therefore compute P_B by integrating $p(z|s_1)$ between the limits $-\infty$ and γ_0 , or by integrating $p(z|s_2)$ between the limits γ_0 and ∞ :

$$P_B = \int_{\gamma_0=(a_1+a_2)/2}^{\infty} p(z|s_2) dz \quad (3.40)$$

Here, $\gamma_0 = (a_1 + a_2)/2$ is the optimum threshold from Equation (3.32). Replacing the likelihood $p(z|s_2)$ with its Gaussian equivalent from Equation (3.6), we have

$$P_B = \int_{\gamma_0 = (a_1 + a_2)/2}^{\infty} \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{z - a_2}{\sigma_0} \right)^2 \right] dz \quad (3.41)$$

where σ_0^2 is the variance of the noise out of the correlator.

Let $u = (z - a_2)/\sigma_0$. Then $\sigma_0 du = dz$ and

$$P_B = \int_{u = (a_1 - a_2)/2\sigma_0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{u^2}{2} \right) du = Q \left(\frac{a_1 - a_2}{2\sigma_0} \right) \quad (3.42)$$

where $Q(x)$, called the *complementary error function* or *co-error function*, is a commonly used symbol for the probability under the tail of the Gaussian pdf. It is defined as

$$Q(x) \approx \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp \left(-\frac{u^2}{2} \right) du \quad (3.43)$$

Note that the co-error function is defined in several ways (see Appendix B); however, all definitions are equally useful for determining probability of error in Gaussian noise. $Q(x)$ cannot be evaluated in closed form. It is presented in tabular form in Table B.1. Good approximations to $Q(x)$ by simpler functions can be found in Reference [5]. One such approximation, valid for $x > 3$, is

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} \exp \left(-\frac{x^2}{2} \right) \quad (3.44)$$

We have optimized (in the sense of minimizing P_B) the threshold level γ , but have not optimized the receiving filter in block 1 of Figure 3.1. We next consider optimizing this filter by maximizing the argument of $Q(x)$ in Equation (3.42).

3.2.2 The Matched Filter

A matched filter is a linear filter designed to provide the maximum signal-to-noise power ratio at its output for a given transmitted symbol waveform. Consider that a known signal $s(t)$ plus AWGN $n(t)$ is the input to a linear, time-invariant (receiving) filter followed by a sampler, as shown in Figure 3.1. At time $t = T$, the sampler output $z(T)$ consists of a signal component a_i and a noise component n_0 . The variance of the output noise (average noise power) is denoted by σ_0^2 , so that the ratio of the instantaneous signal power to average noise power, $(S/N)_T$, at time $t = T$, out of the sampler in step 1, is

$$\left(\frac{S}{N} \right)_T = \frac{a_i^2}{\sigma_0^2} \quad (3.45)$$

We wish to find the filter transfer function $H_0(f)$ that *maximizes* Equation (3.45). We can express the signal $a_i(t)$ at the filter output in terms of the filter transfer function $H(f)$ (before optimization) and the Fourier transform of the input signal, as

(3.41)

$$a_i(t) = \int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi ft} df \quad (3.46)$$

where $S(f)$ is the Fourier transform of the input signal, $s(t)$. If the two-sided power spectral density of the input noise is $N_0/2$ watts/hertz, then, using Equations (1.19) and (1.53), we can express the output noise power as

(3.42)

$$\sigma_0^2 = \frac{N_0}{2} \int_{-\infty}^{\infty} |H(f)|^2 df \quad (3.47)$$

(3.43)

We then combine Equations (3.45) to (3.47) to express $(S/N)_T$, as follows:

$$\left(\frac{S}{N}\right)_T = \frac{\left|\int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi fT} df\right|^2}{N_0/2 \int_{-\infty}^{\infty} |H(f)|^2 df} \quad (3.48)$$

(3.44)

We next find that value of $H(f) = H_0(f)$ for which the maximum $(S/N)_T$ is achieved, by using *Schwarz's inequality*. One form of the inequality can be stated as

$$\left|\int_{-\infty}^{\infty} f_1(x)f_2(x) dx\right|^2 \leq \int_{-\infty}^{\infty} |f_1(x)|^2 dx \int_{-\infty}^{\infty} |f_2(x)|^2 dx \quad (3.49)$$

(3.45)

The equality holds if $f_1(x) = kf_2^*(x)$, where k is an arbitrary constant and $*$ indicates complex conjugate. If we identify $H(f)$ with $f_1(x)$ and $S(f)e^{j2\pi fT}$ with $f_2(x)$, we can write

$$\left|\int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi fT} df\right|^2 \leq \int_{-\infty}^{\infty} |H(f)|^2 df \int_{-\infty}^{\infty} |S(f)|^2 df \quad (3.50)$$

(3.46)

Substituting into Equation (3.48) yields

$$\left(\frac{S}{N}\right)_T \leq \frac{2}{N_0} \int_{-\infty}^{\infty} |S(f)|^2 df \quad (3.51)$$

(3.47)

or

$$\max \left(\frac{S}{N}\right)_T = \frac{2E}{N_0} \quad (3.52)$$

(3.48)

where the energy E of the input signal $s(t)$ is

$$E = \int_{-\infty}^{\infty} |S(f)|^2 df \quad (3.53)$$

(3.49)

Thus, the maximum output $(S/N)_T$ depends on the input *signal energy* and the power spectral density of the noise, *not on the particular shape* of the waveform that is used.

The equality in Equation (3.52) holds only if the optimum filter transfer function $H_0(f)$ is employed, such that

(3.50)

(3.51)

(3.52)

(3.53)

(3.54)

(3.55)

(3.56)

(3.57)

(3.58)

(3.59)

(3.60)

(3.61)

(3.62)

(3.63)

(3.64)

(3.65)

(3.66)

(3.67)

(3.68)

(3.69)

(3.70)

(3.71)

(3.72)

(3.73)

(3.74)

(3.75)

(3.76)

(3.77)

(3.78)

(3.79)

(3.80)

(3.81)

(3.82)

(3.83)

(3.84)

(3.85)

(3.86)

(3.87)

(3.88)

(3.89)

(3.90)

(3.91)

(3.92)

(3.93)

(3.94)

(3.95)

(3.96)

(3.97)

(3.98)

(3.99)

(4.00)

(4.01)

(4.02)

(4.03)

(4.04)

(4.05)

(4.06)

(4.07)

(4.08)

(4.09)

(4.10)

(4.11)

(4.12)

(4.13)

(4.14)

(4.15)

(4.16)

(4.17)

(4.18)

(4.19)

(4.20)

(4.21)

(4.22)

(4.23)

(4.24)

(4.25)

(4.26)

(4.27)

(4.28)

(4.29)

(4.30)

(4.31)

(4.32)

(4.33)

(4.34)

(4.35)

(4.36)

(4.37)

(4.38)

(4.39)

(4.40)

(4.41)

(4.42)

(4.43)

(4.44)

(4.45)

(4.46)

(4.47)

(4.48)

(4.49)

(4.50)

(4.51)

(4.52)

(4.53)

(4.54)

(4.55)

(4.56)

(4.57)

(4.58)

(4.59)

(4.60)

(4.61)

(4.62)

(4.63)

(4.64)

(4.65)

(4.66)

(4.67)

(4.68)

(4.69)

(4.70)

(4.71)

(4.72)

(4.73)

(4.74)

(4.75)

(4.76)

(4.77)

(4.78)

(4.79)

(4.80)

(4.81)

(4.82)

(4.83)

(4.84)

(4.85)

(4.86)

(4.87)

(4.88)

(4.89)

(4.90)

(4.91)

(4.92)

(4.93)

(4.94)

(4.95)

(4.96)

(4.97)

(4.98)

(4.99)

(5.00)

(5.01)

(5.02)

(5.03)

(5.04)

(5.05)

(5.06)

(5.07)

(5.08)

(5.09)

(5.10)

(5.11)

(5.12)

(5.13)

(5.14)

(5.15)

(5.16)

(5.17)

(5.18)

(5.19)

(5.20)

(5.21)

(5.22)

(5.23)

(5.24)

(5.25)

(5.26)

(5.27)

(5.28)

(5.29)

(5.30)

(5.31)

(5.32)

(5.33)

(5.34)

(5.35)

(5.36)

(5.37)

(5.38)

(5.39)

(5.40)

(5.41)

(5.42)

(5.43)

(5.44)

(5.45)

(5.46)

(5.47)

(5.48)

(5.49)

(5.50)

(5.51)

(5.52)

(5.53)

(5.54)

(5.55)

(5.56)

(5.57)

(5.58)

(5.59)

(5.60)

(5.61)

(5.62)

(5.63)

(5.64)

(5.65)

(5.66)

(5.67)

(5.68)

(5.69)

(5.70)

(5.71)

(5.72)

(5.73)

(5.74)

(5.75)

(5.76)

(5.77)

(5.78)

(5.79)

(5.80)

(5.81)

(5.82)

(5.83)

(5.84)

(5.85)

(5.86)

(5.87)

(5.88)

(5.89)

(5.90)

(5.91)

(5.92)

(5.93)

(5.94)

(5.95)

(5.96)

(5.97)

(5.98)

(5.99)

(6.00)

(6.01)

(6.02)

(6.03)

(6.04)

(6.05)

(6.06)

(6.07)

(6.08)

(6.09)

(6.10)

(6.11)

(6.12)

(6.13)

(6.14)

(6.15)

(6.16)

(6.17)

(6.18)

(6.19)

(6.20)

(6.21)

(6.22)

(6.23)

(6.24)

(6.25)

(6.26)

(6.27)

(6.28)

(6.29)

(6.30)

(6.31)

(6.32)

(6.33)

(6.34)

(6.35)

(6.36)

(6.37)

(6.38)

(6.39)

(6.40)

(6.41)

(6.42)

(6.43)

(6.44)

(6.45)

(6.46)

(6.47)

(6.48)

(6.49)

(6.50)

(6.51)

(6.52)

(6.53)

(6.54)

(6.55)

(6.56)

(6.57)

(6.58)

(6.59)

(6.60)

(6.61)

$$H(f) = H_0(f) = kS^*(f)e^{-j2\pi fT} \quad (3.54)$$

or

$$h(t) = \mathcal{F}^{-1}\{kS^*(f)e^{-j2\pi fT}\} \quad (3.55)$$

Since $s(t)$ is a real-valued signal, we can write, from Equations (A.29) and (A.31),

$$h(t) = \begin{cases} ks(T-t) & 0 \leq t \leq T \\ 0 & \text{elsewhere} \end{cases} \quad (3.56)$$

Thus, the impulse response of a filter that produces the maximum output signal-to-noise ratio is the mirror image of the message signal $s(t)$, *delayed* by the symbol time duration T . Note that the delay of T seconds makes Equation (3.56) *causal*; that is, the delay of T seconds makes $h(t)$ a function of positive time in the interval $0 \leq t \leq T$. Without the **delay of T seconds**, the response $s(-t)$ is unrealizable because it describes a response as a function of negative time.

3.2.3 Correlation Realization of the Matched Filter

Equation (3.56) and Figure 3.7a illustrate the matched filter's basic property: The impulse response of the filter is a delayed version of the mirror image (rotated on the $t = 0$ axis) of the signal waveform. Therefore, if the signal waveform is $s(t)$, its mirror image is $s(-t)$, and the mirror image delayed by T seconds is $s(T-t)$. The output $z(t)$ of a causal filter can be described in the time domain as the convolution of a received input waveform $r(t)$ with the impulse response of the filter (see Section A.5):

$$z(t) = r(t) * h(t) = \int_0^t r(\tau)h(t-\tau) d\tau \quad (3.57)$$

Substituting $h(t)$ of Equation (3.56) into $h(t-\tau)$ of Equation (3.57) and arbitrarily setting the constant k equal to unity, we get

$$\begin{aligned} z(t) &= \int_0^t r(\tau)s[T-(t-\tau)] d\tau \\ &= \int_0^t r(\tau)s(T-t+\tau) d\tau \end{aligned} \quad (3.58)$$

When $t = T$, we can write Equation (3.58) as

$$z(T) = \int_0^T r(\tau)s(\tau) d\tau \quad (3.59)$$

The operation of Equation (3.59), the product **integration** of the received signal $r(t)$ with a replica of the transmitted waveform $s(t)$ **over one** symbol interval is known as the *correlation* of $r(t)$ with $s(t)$. Consider that a received signal $r(t)$ is correlated with each prototype signal $s_i(t)$ ($i = 1, \dots, M$), using a bank of M correlators. The signal $s_i(t)$ whose product integration or correlation with $r(t)$ yields the maximum

(3.54)

(3.55)

and (A.31),

(3.56)

output signal-to-noise ratio by the symbol (3.56) *causal*; in the interval s is unrealizable

property: The signal $s(t)$ is rotated on the time axis to $s(T-t)$. The matched filter (see Sec-

(3.57)

and arbitrarily

(3.58)

(3.59)

received signal $r(t)$ in the interval t is known $s(t)$ is correlated with $s(T-t)$. The maximum correlation occurs at $t = T$.

Chapter 3

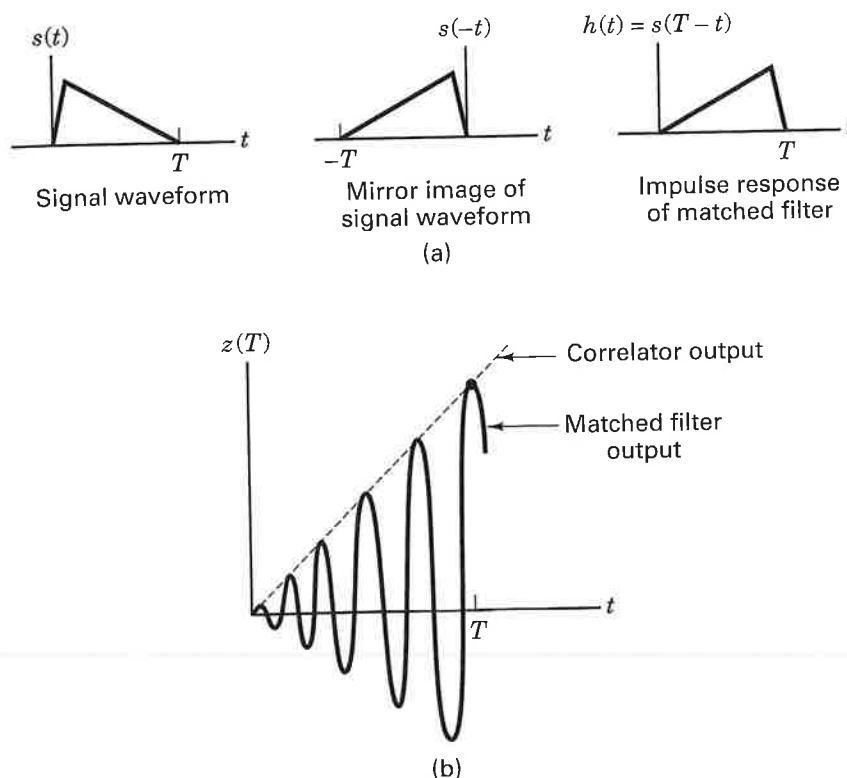


Figure 3.7 Correlator and matched filter. (a) Matched filter characteristic. (b) Comparison of correlator and matched filter outputs.

output $z_i(T)$ is the signal that matches $r(t)$ better than all the other $s_j(t)$, $j \neq i$. We will subsequently use this correlation characteristic for the optimum detection of signals.

3.2.3.1 Comparison of Convolution and Correlation

The mathematical operation of a matched filter (MF) is *convolution*; a signal is convolved with the impulse response of a filter. The mathematical operation of a correlator is *correlation*; a signal is correlated with a replica of itself. The term "matched filter" is often used synonymously with "correlator." How is that possible when their mathematical operations are different? Recall that the process of *convolving* two signals reverses one of them in time. Also, the impulse response of an MF is defined in terms of a signal that is reversed in time. Thus, convolution in the MF with a time-reversed function results in a second time-reversal, making the output (at the end of a symbol time) appear to be the result of a signal that has been correlated with its replica. Therefore, it is valid to implement the receiving filter in Figure 3.1 with either a matched filter or a correlator. It is important to note that the correlator output and the matched filter output are the same *only at time*

$t = T$. For a sine-wave input, the output of the correlator, $z(t)$, is approximately a linear ramp during the interval $0 \leq t \leq T$. However, the matched filter output is approximately a sine wave that is amplitude modulated by a linear ramp during the same time interval. The comparison is shown in Figure 3.7b. Since for comparable inputs, the MF output and the correlator output are identical at the sampling time $t = T$, the matched filter and correlator functions pictured in Figure 3.8 are often used interchangeably.

3.2.3.2 Dilemma in Representing Earliest versus Latest Event

A serious dilemma exists in representing timed events. This dilemma is undoubtedly the cause of a frequently made error in electrical engineering—confusing the most significant bit (MSB) with the least significant bit (LSB). Figure 3.9a illustrates how a function of time is typically plotted; the earliest event appears leftmost, and the latest event rightmost. In western societies, where we read from left to right, would there be any other way to plot timed events? Consider Figure 3.9b, where pulses are shown entering (and leaving) a network or circuit. Here, the earliest events are shown rightmost, and the latest are leftmost. From the figure, it should be clear that whenever we denote timed events, there is an inference that we are following one of the two formats described here. Often, we have to provide some descriptive words (e.g., the rightmost bit is the earliest bit) to avoid confusion.

Mathematical relationships often have built-in features guaranteeing the proper alignment of time events. For example, in Section 3.2.3, a matched filter is defined as having an impulse response $h(t)$ that is a delayed version of the time-reversed copy of the signal. That is, $h(t) = s(T - t)$. Delay of one symbol time T is needed for the filter to be causal (the output must occur in positive time). Time reversal can be thought of as a “precorrection” where the rightmost part of the time plot will now correspond to the earliest event. Since convolution dictates another time reversal, the arriving signal and the filter’s impulse response will be “in step” (earliest with earliest, and latest with latest).

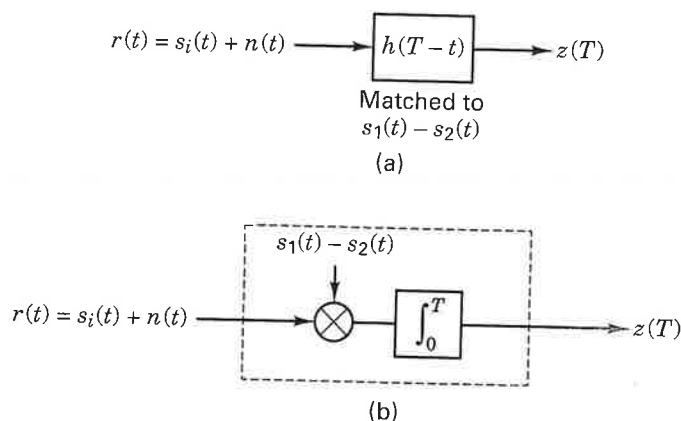


Figure 3.8 Equivalence of matched filter and correlator. (a) Matched filter. (b) Correlator.

approximately a
er output is ap-
amp during the
for comparable
e sampling time
re 3.8 are often

dilemma is un-
eering—confus-
3B). Figure 3.9a
ent appears left-
e read from left
der Figure 3.9b,
rcuit. Here, the
om the figure, it
n inference that
en, we have to
est bit) to avoid

uaranteeing the
matched filter is
ion of the time-
symbol time T is
e time). Time re-
part of the time
dictates another
will be “in step”

alence of matched
or. (a) Matched fil-

ection Chap. 3

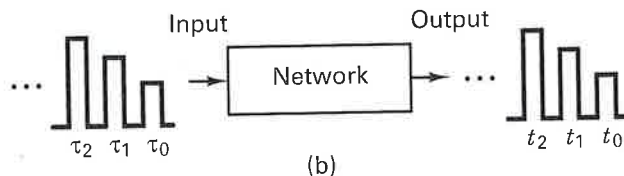
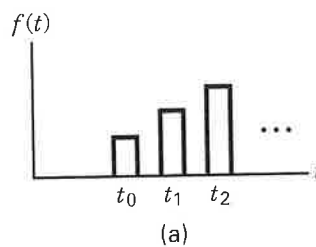


Figure 3.9 Dilemma in representing earliest versus latest events.

3.2.4 Optimizing Error Performance

To optimize (minimize) P_B in the context of an AWGN channel and the receiver shown in Figure 3.1, we need to select the optimum receiving filter in step 1 and the optimum decision threshold in step 2. For the binary case, the optimum decision threshold has already been chosen in Equation (3.32), and it was shown in Equation (3.42) that this threshold results in $P_B = Q[(a_1 - a_2)/2\sigma_0]$. Next, for minimizing P_B , it is necessary to choose the filter (matched filter) that maximizes the argument of $Q(x)$. Thus, we need to determine the linear filter that maximizes $(a_1 - a_2)/2\sigma_0$, or equivalently, that maximizes

$$\frac{(a_1 - a_2)^2}{\sigma_0^2} \quad (3.60)$$

where $(a_1 - a_2)$ is the difference of the desired signal components at the filter output at time $t = T$, and the square of this difference signal is the instantaneous power of the difference signal. In Section 3.2.2, a matched filter was described as one that maximizes the output signal-to-noise ratio (SNR) for a given known signal. Here, we continue that development for binary signaling, where we view the optimum filter as one that maximizes the difference between two possible signal outputs. Starting with Equations (3.45) and (3.47), it was shown in Equation (3.52) that a matched filter achieves the maximum possible output SNR equal to $2E/N_0$. Consider that the filter is matched to the input difference signal $[s_1(t) - s_2(t)]$; thus, we can write an output SNR at time $t = T$ as

$$\left(\frac{S}{N}\right)_T = \frac{(a_1 - a_2)^2}{\sigma_0^2} = \frac{2E_d}{N_0} \quad (3.61)$$

where $N_0/2$ is the two-sided power spectral density of the noise at the filter input and

$$E_d = \int_0^T [s_1(t) - s_2(t)]^2 dt \quad (3.62)$$

is the energy of the difference signal at the filter input. Note that Equation (3.61) does not represent the SNR for any single transmission, $s_1(t)$ or $s_2(t)$. This SNR yields a metric of signal difference for the filter output. By maximizing the output SNR as shown in Equation (3.61), the matched filter provides the maximum distance (normalized by noise) between the two candidate outputs—signal a_1 and signal a_2 .

Next, combining Equations (3.42) and (3.61) yields

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) \quad (3.63)$$

For the matched filter, Equation (3.63) is an important interim result in terms of the energy of the difference signal at the filter's input. From this equation, a more general relationship in terms of received bit energy can be developed. We start by defining a time cross-correlation coefficient ρ as a measure of similarity between two signals, $s_1(t)$ and $s_2(t)$. We have

$$\rho = \frac{1}{E_b} \int_0^T s_1(t) s_2(t) dt \quad (3.64a)$$

and

$$\rho = \cos \theta \quad (3.64b)$$

where $-1 \leq \rho \leq 1$. Equation (3.64a) is the classical mathematical way of expressing correlation. However, when $s_1(t)$ and $s_2(t)$ are viewed as signal vectors, $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively, then ρ is conveniently expressed by Equation (3.64b). This vector view provides a useful image. The vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ are separated by the angle θ ; for small angular separation, the vectors are quite similar (highly correlated) to each other, and for large angular separation, they are quite dissimilar. The cosine of this angle gives us the same normalized metric of correlation as Equation (3.64a).

Expanding Equation (3.62), we get

$$E_d = \int_0^T s_1^2(t) dt + \int_0^T s_2^2(t) dt - 2 \int_0^T s_1(t) s_2(t) dt \quad (3.65)$$

Recall that each of the first two terms in Equation (3.65) represents the energy associated with a bit, E_b ; that is,

$$E_b = \int_0^T s_1^2(t) dt = \int_0^T s_2^2(t) dt \quad (3.66)$$

Substituting Equations (3.64a) and (3.66) into Equation (3.65), we get

$$E_d = E_b + E_b - 2\rho E_b = 2E_b(1 - \rho) \quad (3.67)$$

Substituting Equation (3.67) into (3.63), we obtain

(3.62)

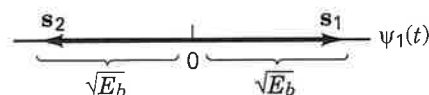
$$P_B = Q\left(\sqrt{\frac{E_b(1-\rho)}{N_0}}\right) \quad (3.68)$$

Consider the case of $\rho = 1$ corresponding to signals $s_1(t)$ and $s_2(t)$ being perfectly correlated over a symbol time (drawn as vectors, with the angle between them equal to zero). Would anyone use such waveforms for digital signaling? Of course not, because the communication signals (members of the alphabet set) need to be as disparate from one another as possible so that they are easily distinguished (detected). We are simply cataloging the possible values for the parameter ρ . Consider the case of $\rho = -1$ corresponding to $s_1(t)$ and $s_2(t)$ being "anticorrelated" over a symbol time. In other words, the angle between the signal vectors is 180° . In such a case, where the vectors are mirror images, we call the signals *antipodal*, as shown in Figure 3.10a. Also, consider the case of $\rho = 0$ corresponding to zero correlation between $s_1(t)$ and $s_2(t)$ (the angle between the vectors is 90°). We call such signals *orthogonal*, as seen in Figure 3.10b. For two waveforms to be orthogonal, they must be uncorrelated over a symbol interval; that is,

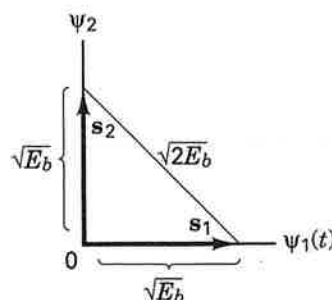
$$\int_0^T s_1(t) s_2(t) dt = 0 \quad (3.69)$$

The subject of orthogonality was treated earlier in Section 3.1.3. For the case of detecting antipodal signals (that is, $\rho = -1$) with a matched filter, Equation (3.68) can be written as

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \quad (3.70)$$



(a)



(b)

Figure 3.10 Binary signal vectors. (a) Antipodal. (b) Orthogonal.

Similarly, for the case of detecting orthogonal signals (that is, $\rho = 0$) with a matched filter, Equation (3.68) can be written as

$$P_B = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \quad (3.71)$$

Figure 3.10, where the signal magnitudes are each shown equal to $\sqrt{E_b}$, helps illustrate that the error performance described by Equations (3.70) and (3.71) is a function of the distance between s_1 and s_2 (the larger the distance, the smaller will be P_B). When the signals are antipodal, as in Figure 3.10a, the distance between them is $2\sqrt{E_b}$, and the energy E_d associated with that distance is characterized by the distance squared or $4E_b$. When we substitute $E_d = 4E_b$ into Equation (3.63), the result is Equation (3.70). When the signals are orthogonal as in Figure 3.10b, the distance between them is $\sqrt{2E_b}$ and thus $E_d = 2E_b$. When we substitute $E_d = 2E_b$ into Equation (3.63), the result is Equation (3.71).

Example 3.2 Matched Filter Detection of Antipodal Signals

Consider a binary communications system that receives equally likely signals $s_1(t)$ and $s_2(t)$ plus AWGN. See Figure 3.11. Assume that the receiving filter is a matched filter (MF), and that the noise-power spectral density N_0 is equal to 10^{-12} Watt/Hz. Use the values of received signal voltage and time shown on Figure 3.11 to compute the bit-error probability.

Solution

We can graphically determine the received energy per bit from the plot of either $s_1(t)$ or $s_2(t)$ shown in Figure 3.11 by integrating to find the energy (area under the voltage-squared pulse). Doing this in piecewise fashion, we get

$$\begin{aligned} E_b &= \int_0^3 v^2(t) dt \\ &= (10^{-3} \text{ V})^2 \times (10^{-6} \text{ s}) + (2 \times 10^{-3} \text{ V})^2 \times (10^{-6} \text{ s}) + (10^{-3} \text{ V})^2 \times (10^{-6} \text{ s}) \\ &= 6 \times 10^{-12} \text{ joule} \end{aligned}$$

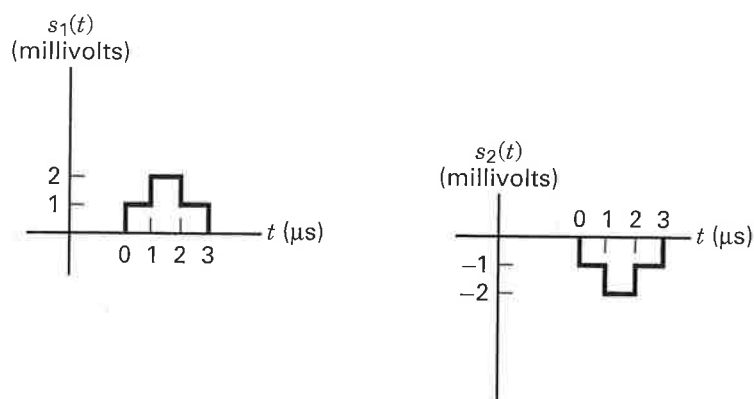


Figure 3.11 Baseband antipodal waveforms.

with a matched

(3.71)

al to $\sqrt{E_b}$, helps
) and (3.71) is a
the smaller will
istance between
characterized by
ation (3.63), the
figure 3.10b, the
stitute $E_d = 2 E_b$

ly signals $s_1(t)$ and
is a matched filter
Watt/Hz. Use the
, compute the bit-

plot of either $s_1(t)$
under the voltage-

$\sqrt{V}^2 \times (10^{-6} \text{ s})$

$\frac{3}{J} t \text{ (}\mu\text{s)}$

Since the waveforms depicted in Figure 3.11 are antipodal and are detected with a matched filter, we use Equation (3.70) to find the bit-error probability, as

$$Q\left(\sqrt{\frac{12 \times 10^{-12}}{10^{-12}}}\right) = Q(\sqrt{12}) = Q(3.46)$$

From Table B.1, we find that $P_B = 3 \times 10^{-4}$. Or, since the argument of $Q(x)$ is greater than 3, we can also use the approximate relationship in Equation (3.44) which yields $P_B \approx 2.9 \times 10^{-4}$.

Because the received signals are antipodal and are received by an MF, these are sufficient prerequisites such that Equation (3.70) provides the proper relationship for finding bit-error probability. The waveforms $s_1(t)$ and $s_2(t)$ could have been pictured in a much more bizarre fashion, but as long as they are antipodal and detected by an MF, their shapes do not enter into the P_B computations. The shapes of the waveforms, of course, *do matter* when it comes to specifying the impulse response of the MF needed to detect these waveforms.

3.2.5 Error Probability Performance of Binary Signaling

3.2.5.1 Unipolar Signaling

Figure 3.12a illustrates an example of baseband orthogonal signaling—namely, unipolar signaling, where

$$\begin{aligned} s_1(t) &= A & 0 \leq t \leq T & \text{for binary 1} \\ s_2(t) &= 0 & 0 \leq t \leq T & \text{for binary 0} \end{aligned} \quad (3.72)$$

and where $A > 0$ is the amplitude of symbol $s_1(t)$. The definition of orthogonal signaling described by Equation (3.69) requires that $s_1(t)$ and $s_2(t)$ have zero correlation over each symbol time duration. Because in Equation (3.72), $s_2(t)$ is equal to zero during the symbol time, this set of unipolar pulses clearly fulfills the condition

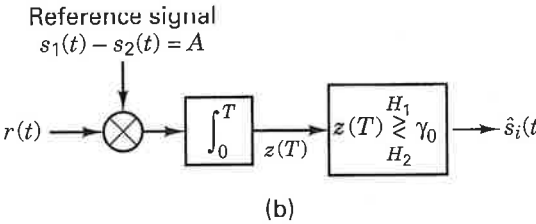
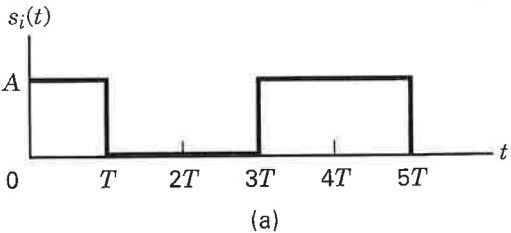


Figure 3.12 Detection of unipolar baseband signaling. (a) Unipolar signaling example. (b) Correlator detector.

shown in Equation (3.69), and hence, they form an orthogonal signal set. Consider such unipolar signaling, as illustrated in Figure 3.12a, as well as the correlator shown in Figure 3.12b, which can be used for detecting such pulses. The correlator multiplies and integrates the incoming signal $r(t)$ with the difference of the prototype signals, $[s_1(t) - s_2(t)] = A$. After a symbol duration T , a sampler (inferred by the upper limit of integration) yields the test statistic $z(T)$, which is then compared with the threshold γ_0 . For the case of $s_1(t)$ plus AWGN being received—that is, when $r(t) = s_1(t) + n(t)$ —the signal component of $z(T)$ is found, using Equation (3.59), to be

$$a_1(T) = E\{z(T)|s_1(t)\} = E\left\{\int_0^T A^2 + An(t) dt\right\} = A^2T$$

where $E\{z(T)|s_1(t)\}$ is the *expected value* of $z(T)$, given that $s_1(t)$ was sent. This follows since $E\{n(t)\} = 0$. Similarly, when $r(t) = s_2(t) + n(t)$, then $a_2(T) = 0$. Thus, in this case, the optimum decision threshold, from Equation (3.32), is given by $\gamma_0 = (a_1 + a_2)/2 = 1/2 A^2T$. If the test statistic $z(T)$ is greater than γ_0 , the signal is declared to be $s_1(t)$; otherwise, it is declared to be $s_2(t)$.

The energy difference signal, from Equation (3.62), is given by $E_d = A^2T$. Then the bit-error performance at the output is obtained from Equation (3.63) as

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) = Q\left(\sqrt{\frac{A^2T}{2N_0}}\right) = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \quad (3.73)$$

where, for the case of equally likely signaling, the *average* energy per bit is $E_b = A^2T/2$. Equation (3.73) corroborates Equation (3.71) where this relationship was established for orthogonal signaling in a more general way.

Note that the units out of a multiplier circuit, such as that seen in Figure 3.12b are volts. Therefore, for voltage signals on each of the two inputs, the multiplier transfer function must have units of 1/volt, and the measurable units of $r(t)s_i(t)$ out of the multiplier are volt/volt-squared. Similarly, the units out of an integrator circuit are also volts. Thus, for a voltage signal into an integrator, the integrator transfer function must have units of 1/second, and thus the overall transfer function of the product integrator has units of 1/volt-second. Then, for a signal into the product integrator having units of volt-squared-seconds (a measure of energy), we represent the output $z(T)$ as a voltage signal that is proportional to received signal energy (volt/joule).

3.2.5.2 Bipolar Signaling

Figure (3.13a) illustrates an example of baseband antipodal signaling—namely, bipolar signaling, where

$$s_1(t) = +A \quad 0 \leq t \leq T \quad \text{for binary 1}$$

and

$$s_2(t) = -A \quad 0 \leq t \leq T \quad \text{for binary 0}$$

(3.74)

nal set. Consider
s the correlator
s. The correlator
ice of the proto-
pler (inferred by
s then compared
eceived—that is,
, using Equation

$$A^2T$$

t) was sent. This
 $a_2(T) = 0$. Thus,
(3.32), is given by
a γ_0 , the signal is

en by $E_d = A^2T$.
uation (3.63) as

$$(3.73)$$

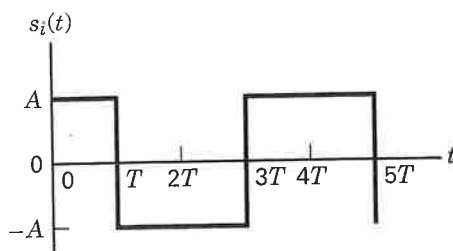
gy per bit is $E_b =$
relationship was

en in Figure 3.12b
its, the multiplier
its of $r(t) s_i(t)$ out
an integrator cir-
e integrator trans-
ansfer function of
al into the product
of energy), we
to received signal

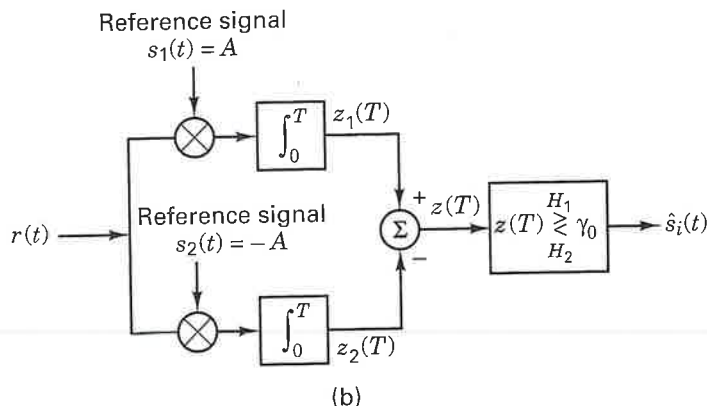
podal signaling—

$$(3.74)$$

etection Chap. 3



(a)



(b)

Figure 3.13 Detection of bipolar baseband signaling. (a) Bipolar signaling example. (b) Correlator detector.

As defined earlier, the term “antipodal” refers to binary signals that are mirror images of one another; that is, $s_1(t) = -s_2(t)$. A correlator receiver for such antipodal waveforms can be configured as shown in Figure 3.13b. One correlator multiplies and integrates the incoming signal $r(t)$ with the prototype signal $s_1(t)$; the second correlator multiplies and integrates $r(t)$ with $s_2(t)$.

Figure 3.13b captures the essence of a digital receiver’s main function. That is, during each symbol interval, a noisy input signal is sent down multiple “avenues” in an effort to correlate it with each of the possible candidates. The receiver then seeks the largest output voltage (the best match) to make a detection. For the binary example, there are just two possible candidates. For a 4-ary example, there would be four candidates, and so forth. In Figure 3.13b, the correlator outputs are designated $z_i(T)$ ($i = 1, 2$). The test statistic, formed from the difference of the correlator outputs, is

$$z(T) = z_1(T) - z_2(T) \quad (3.75)$$

and the decision is made using the threshold shown in Equation (3.32). For antipodal signals, $a_1 = -a_2$; therefore, $\gamma_0 = 0$. Thus, if the test statistic $z(T)$ is positive, the signal is declared to be $s_1(T)$; and if it is negative, it is declared to be $s_2(T)$.

From Equation (3.62), the energy-difference signal is $E_d = (2A)^2T$. Then, the bit-error performance at the output can be obtained from Equation (3.63) as

$$P_B = Q\left(\sqrt{\frac{E_d}{2N_0}}\right) = Q\left(\sqrt{\frac{2A^2T}{N_0}}\right) = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \quad (3.76)$$

where the average energy per bit is given by $E_b = A^2T$. Equation (3.76) corroborates Equation (3.70) where this relationship was established for antipodal signaling in a more general way.

3.2.5.3 Signaling Described with Basis Functions

Instead of designating $s_i(t)$ as the reference signals in the correlator of Figure 3.13b, we can use the concept of *basis functions* described in Section 3.1.3. Binary signaling with unipolar or bipolar pulses provides particularly simple examples for doing this, because the entire signaling space can be described by just one basis function. If we **normalize the space by choosing $K_j = 1$** in Equation (3.9), then it should be clear that the basis function $\psi_1(t)$ must be equal to $\sqrt{1/T}$.

For unipolar pulse signaling, we could then write

$$s_1(t) = a_{11}\psi_1(t) = A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = A$$

and

$$s_2(t) = a_{21}\psi_1(t) = 0 \times \left(\sqrt{\frac{1}{T}}\right) = 0$$

where the coefficients a_{11} and a_{21} equal $A\sqrt{T}$ and 0, respectively.

For bipolar pulse signaling, we would write

$$s_1(t) = a_{11}\psi_1(t) = A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = A$$

and

$$s_2(t) = a_{21}\psi_1(t) = -A\sqrt{T} \times \left(\sqrt{\frac{1}{T}}\right) = -A$$

where the coefficients a_{11} and a_{21} equal $A\sqrt{T}$ and $-A\sqrt{T}$, respectively. For the case of antipodal pulses, we can envision the correlator receiver taking the form of Figure 3.12b with the reference signal equal to $\sqrt{1/T}$. Then, for the case of $s_1(t) = A$ being sent, we can write

$$a_1(T) = \mathbf{E}\{z(T)|s_1(t)\} = \mathbf{E}\left\{\int_0^T \frac{A}{\sqrt{T}} + \frac{n(t)}{\sqrt{T}} dt\right\} = A\sqrt{T}$$

This follows because $\mathbf{E}\{n(t)\} = 0$, and since, for antipodal signaling, $E_b = A^2T$, it follows that $a_1(T) = \sqrt{E_b}$. Similarly, for a received signal $r(t) = s_2(t) + n(t)$, it follows that $a_2(T) = -\sqrt{E_b}$. When reference signals are treated in this way, then the ex-

$A)^2T$. Then, the
(3.63) as

(3.76)

(3.76) corroborates
antipodal signal-

relator of Figure
on 3.1.3. Binary
examples for
just one basis
on (3.9), then it

ectively. For the
taking the form of
the case of $s_1(t) =$

$$1 \sqrt{T}$$

$E_b = A^2T$, it fol-
+ $n(t)$, it follows
way, then the ex-

ection Chap. 3

pected value of $z(T)$ has a magnitude of $\sqrt{E_b}$, which has units of normalized volts proportional to received energy. This basis-function treatment of the correlator yields a convenient value of $z(T)$ that is consistent with units of volts out of multipliers and integrators. We therefore repeat an important point: At the output of the sampler (the predetection point), the test statistic $z(T)$ is a voltage signal that is proportional to received signal energy.

Figure 3.14 illustrates curves of P_B versus E_b/N_0 for bipolar and unipolar signaling. There are only two fair ways to compare such curves. By drawing a vertical line at some given E_b/N_0 , say, 10 dB, we see that the unipolar signaling yields P_B in the order of 10^{-3} , but that the bipolar signaling yields P_B in the order of 10^{-6} . The lower curve is the better performing one. Also, by drawing a horizontal line at some required P_B , say 10^{-5} , we see that with unipolar signaling each received bit would require an E_b/N_0 of about 12.5 dB, but with bipolar signaling, we could get away with requiring each received bit to have an E_b/N_0 of only about 9.5 dB. Of course, the smaller requirement of E_b/N_0 is better (using less power, smaller batteries). In general, the better performing curves are the ones closest to the axes, lower and leftmost. In examining the two curves in Figure 3.14, we can see a 3-dB error-performance improvement for bipolar compared with unipolar signaling. This

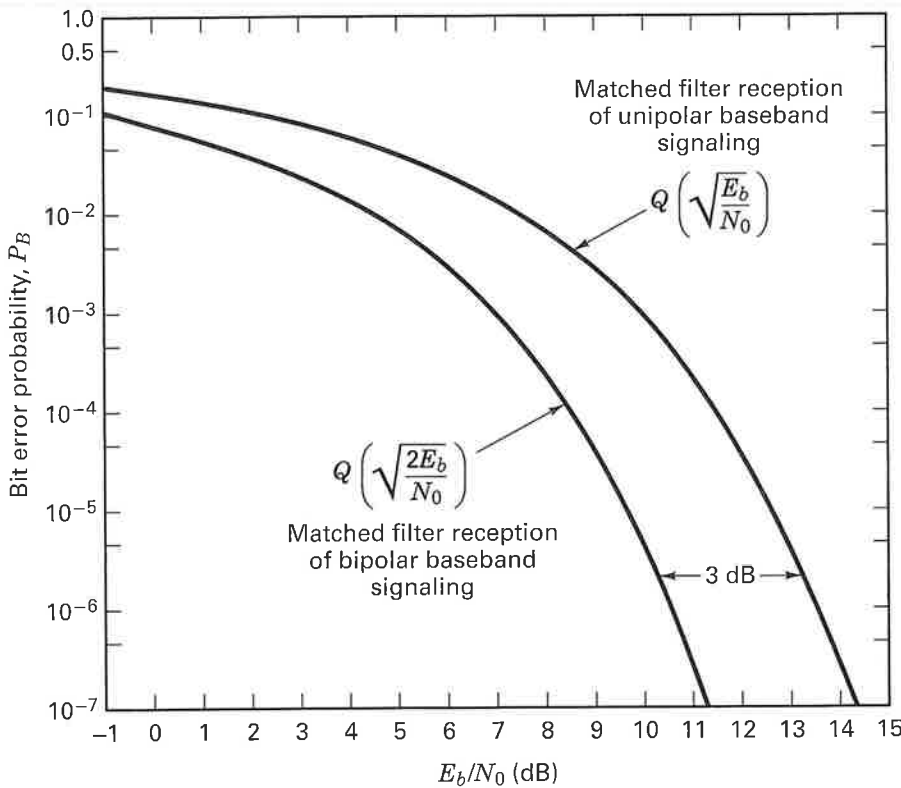


Figure 3.14 Bit error performance of unipolar and bipolar signaling.

difference could have been predicted by the factor-of-2 difference in the coefficient of E_b/N_0 in Equation (3.70) compared with (3.71). In Chapter 4, it is shown that, with MF detection, *bandpass* antipodal signaling (e.g., binary phase-shift keying) has the same P_B performance as *baseband* antipodal signaling (e.g., bipolar pulses). It is also shown that, with MF detection, *bandpass* orthogonal signaling (e.g., orthogonal frequency-shift keying) has the same P_B performance as *baseband* orthogonal signaling (e.g., unipolar pulses).

3.3 INTERSYMBOL INTERFERENCE

Figure 3.15a introduces the filtering aspects of a typical digital communication system. There are various filters (and reactive circuit elements such as inductors and capacitors) throughout the system—in the transmitter, in the receiver, and in the channel. At the transmitter, the information symbols, characterized as impulses or voltage levels, modulate pulses that are then filtered to comply with some bandwidth constraint. For baseband systems, the channel (a cable) has distributed reactances that distort the pulses. Some bandpass systems, such as wireless systems, are characterized by fading channels (see Chapter 15), that behave like undesirable filters manifesting signal distortion. When the receiving filter is configured to compensate for the distortion caused by *both* the transmitter and the channel, it is often referred to as an *equalizing filter* or a *receiving/equalizing filter*. Figure 3.15b illustrates a convenient model for the system, lumping all the filtering effects into one overall equivalent system transfer function

$$H(f) = H_t(f) H_c(f) H_r(f) \quad (3.77)$$

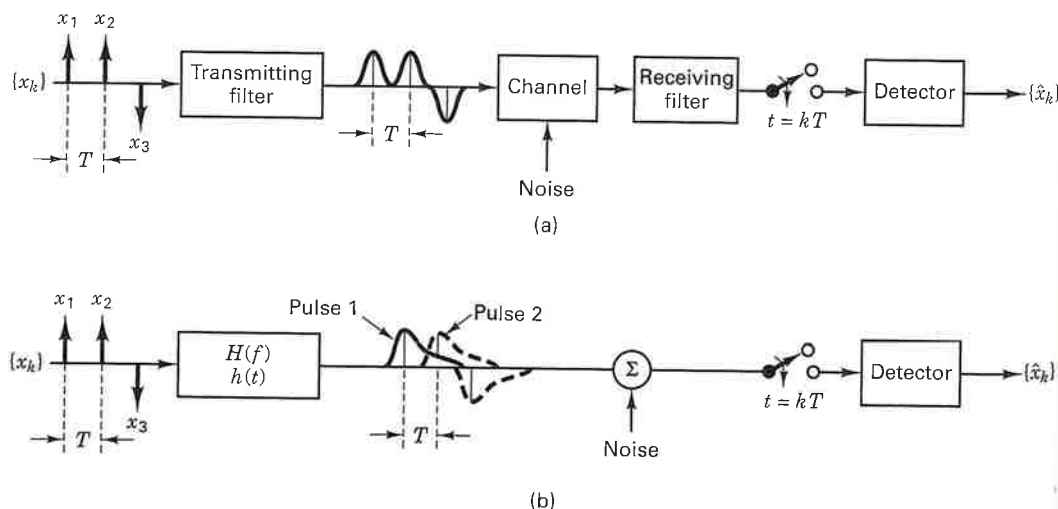
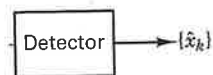


Figure 3.15 Intersymbol interference in the detection process. (a) Typical baseband digital system. (b) Equivalent model.

the coefficient
it is shown that,
se-shift keying)
bipolar pulses).
naling (e.g., or-
aseband orthog-

communication sys-
as inductors and
river, and in the
d as impulses or
with some band-
distributed reac-
less systems, are
e undesirable fil-
figured to com-
annel, it is often
figure 3.15b illus-
g effects into one

(3.77)



cal baseband

tection Chap. 3

where $H_t(f)$ characterizes the transmitting filter, $H_c(f)$ the filtering within the channel, and $H_r(f)$ the receiving/equalizing filter. The characteristic $H(f)$, then, represents the composite system transfer function due to all the filtering at various locations throughout the transmitter/channel/receiver chain. In a binary system with a common PCM waveform, such as NRZ-L, the detector makes a symbol decision by comparing a sample of the received pulse to a threshold; for example, the detector in Figure 3.15 decides that a binary one was sent if the received pulse is positive, and that a binary zero was sent, if the received pulse is negative. Due to the effects of system filtering, the received pulses can overlap one another as shown in Figure 3.15b. The tail of a pulse can "smear" into adjacent symbol intervals, thereby interfering with the detection process and degrading the error performance; such interference is termed *intersymbol interference* (ISI). Even in the absence of noise, the effects of filtering and channel-induced distortion lead to ISI. Sometimes $H_c(f)$ is specified, and the problem remains to determine $H_t(f)$ and $H_r(f)$, such that the ISI is minimized at the output of $H_r(f)$.

Nyquist [6] investigated the problem of specifying a received pulse shape so that no ISI occurs at the detector. He showed that the theoretical minimum system bandwidth needed in order to detect R_s symbols/s, without ISI, is $R_s/2$ hertz. This occurs when the system transfer function $H(f)$ is made rectangular, as shown in Figure 3.16a. For baseband systems, when $H(f)$ is such a filter with single-sided bandwidth $1/2T$ (the *ideal Nyquist filter*), its impulse response, the inverse Fourier transform of $H(f)$ (from Table A.1) is of the form $h(t) = \text{sinc}(t/T)$, shown in Figure 3.16b. This sinc (t/T) -shaped pulse is called the *ideal Nyquist pulse*; its multiple lobes comprise a mainlobe and sidelobes called pre- and post-mainlobe *tails* that are infinitely long. Nyquist established that if each pulse of a received sequence is of the form sinc (t/T) , the pulses can be detected without ISI. Figure 3.16b illustrates how ISI is avoided. There are two successive pulses, $h(t)$ and $h(t - T)$. Even though $h(t)$ has long tails, the figure shows a tail passing through zero amplitude at the instant $(t = T)$ when $h(t - T)$ is to be sampled, and likewise all tails pass through zero amplitude when any other pulse of the sequence $h(t - kT)$, $k = \pm 1, \pm 2, \dots$ is to be sampled. Therefore, assuming that the sample timing is perfect, there will be no ISI degradation introduced. For baseband systems, the bandwidth

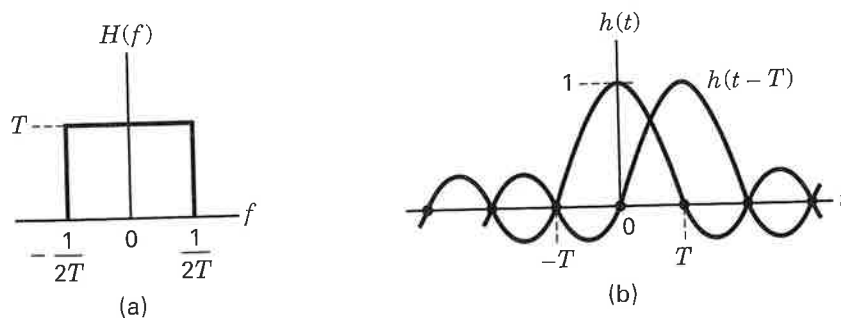


Figure 3.16 Nyquist channel for zero ISI. (a) Rectangular system transfer function $H(f)$. (b) Received pulse shape $h(t) = \text{sinc}(t/T)$.

required to detect $1/T$ such pulses (symbols) per second is equal to $1/2T$; in other words, a system with bandwidth $W = 1/2T = R_s/2$ hertz can support a maximum transmission rate of $2W = 1/T = R_s$ symbols/s (*Nyquist bandwidth constraint*) without ISI. Thus, for ideal Nyquist filtering (and zero ISI), the maximum possible symbol transmission rate per hertz, called the *symbol-rate packing*, is 2 symbols/s/Hz. It should be clear from the rectangular-shaped transfer function of the ideal Nyquist filter and the infinite length of its corresponding pulse, that such ideal filters are not realizable; they can only be approximately realized.

The names “Nyquist filter” and “Nyquist pulse” are often used to describe the general class of filtering and pulse-shaping that satisfy zero ISI at the sampling points. A Nyquist filter is one whose frequency transfer function can be represented by a rectangular function convolved with any real even-symmetric frequency function. A Nyquist pulse is one whose shape can be represented by a sinc (t/T) function multiplied by another time function. Hence, there are a countless number of Nyquist filters and corresponding pulse shapes. Amongst the class of Nyquist filters, the most popular ones are the raised cosine and the root-raised cosine, treated below.

A fundamental parameter for communication systems is *bandwidth efficiency*, R/W , whose units are bits/s/Hz. As the units imply, R/W represents a measure of data throughput per hertz of bandwidth and thus measures how efficiently any signaling technique utilizes the bandwidth resource. Since the Nyquist bandwidth constraint dictates that the theoretical maximum symbol-rate packing without ISI is 2 symbols/s/Hz, one might ask what it says about the maximum number of bits/s/Hz. It says nothing about bits, directly; the constraint deals only with pulses or symbols, and the ability to detect their amplitude values without distortion from other pulses. To find R/W for any signaling scheme, one must know how many bits each symbol represents, which is a separate issue. Consider an M -ary PAM signaling set. Each symbol (comprising k bits) is represented by one of M -pulse amplitudes. For $k = 6$ bits per symbol, the symbol set size is $M = 2^k = 64$ amplitudes. Thus with 64-ary PAM, the theoretical maximum bandwidth efficiency that is possible without ISI is 12 bits/s/Hz. (Bandwidth efficiency is treated in greater detail in Chapter 9.)

3.3.1 Pulse Shaping to Reduce ISI

3.3.1.1 Goals and Trade-offs

The more compact we make the signaling spectrum, the higher is the allowable data rate or the greater is the number of users that can simultaneously be served. This has important implications to communication service providers, since greater utilization of the available bandwidth translates into greater revenue. For most communication systems (with the exception of spread-spectrum systems, covered in Chapter 12), our goal is to reduce the required system bandwidth as much as possible. Nyquist has provided us with a basic limitation to such bandwidth reduction. What would happen if we tried to force a system to operate at smaller bandwidths than the constraint dictates? The pulses would become spread in time,

which would degrade the system's error performance due to increased ISI. A prudent goal is to compress the bandwidth of the data impulses to some reasonably small bandwidth greater than the Nyquist minimum. This is accomplished by pulse-shaping with a Nyquist filter. If the band edge of the filter is steep, approaching the rectangle in Figure 3.16a, then the signaling spectrum can be made most compact. However, such a filter has an impulse response duration approaching infinity, as indicated in Figure 3.16b. Each pulse extends into every pulse in the entire sequence. Long time responses exhibit large-amplitude tails nearest the main lobe of each pulse. Such tails are undesirable because, as shown in Figure 3.16b, they contribute zero ISI *only* when the sampling is performed *at exactly* the correct sampling time; when the tails are large, small timing errors will result in ISI. Therefore, although a compact spectrum provides optimum bandwidth utilization, it is very susceptible to ISI degradation induced by timing errors.

3.3.1.2 The Raised-Cosine Filter

Earlier, it was stated that the receiving filter is often referred to as an *equalizing filter*, when it is configured to compensate for the distortion caused by both the transmitter and the channel. In other words, the configuration of this filter is chosen so as to optimize the composite system frequency transfer function $H(f)$, shown in Equation (3.77). One frequently used $H(f)$ transfer function belonging to the Nyquist class (zero ISI at the sampling times) is called the *raised-cosine filter*. It can be expressed as

$$H(f) = \begin{cases} 1 & \text{for } |f| < 2W_0 - W \\ \cos^2 \left(\frac{\pi}{4} \frac{|f| + W - 2W_0}{W - W_0} \right) & \text{for } 2W_0 - W < |f| < W \\ 0 & \text{for } |f| > W \end{cases} \quad (3.78)$$

where W is the absolute bandwidth and $W_0 = 1/2T$ represents the minimum Nyquist bandwidth for the rectangular spectrum and the -6-dB bandwidth (or half-amplitude point) for the raised-cosine spectrum. The difference $W - W_0$ is termed the "excess bandwidth," which means additional bandwidth beyond the Nyquist minimum (i.e., for the rectangular spectrum, W is equal to W_0). The *roll-off factor* is defined to be $r = (W - W_0)/W_0$, where $0 \leq r \leq 1$. It represents the excess bandwidth divided by the filter -6-dB bandwidth (i.e., the fractional excess bandwidth). For a given W_0 , the roll-off r specifies the required excess bandwidth as a fraction of W_0 and characterizes the steepness of the filter roll off. The raised-cosine characteristic is illustrated in Figure 3.17a for roll-off values of $r = 0$, $r = 0.5$, and $r = 1$. The $r = 0$ roll-off is the Nyquist minimum-bandwidth case. Note that when $r = 1$, the required excess bandwidth is 100%, and the tails are quite small. A system with such an overall spectral characteristic can provide a symbol rate of R_s symbols/s using a bandwidth of R_s hertz (twice the Nyquist minimum bandwidth), thus yielding a symbol-rate packing of 1 symbol/s/Hz. The corresponding impulse response for the $H(f)$ of Equation (3.78) is

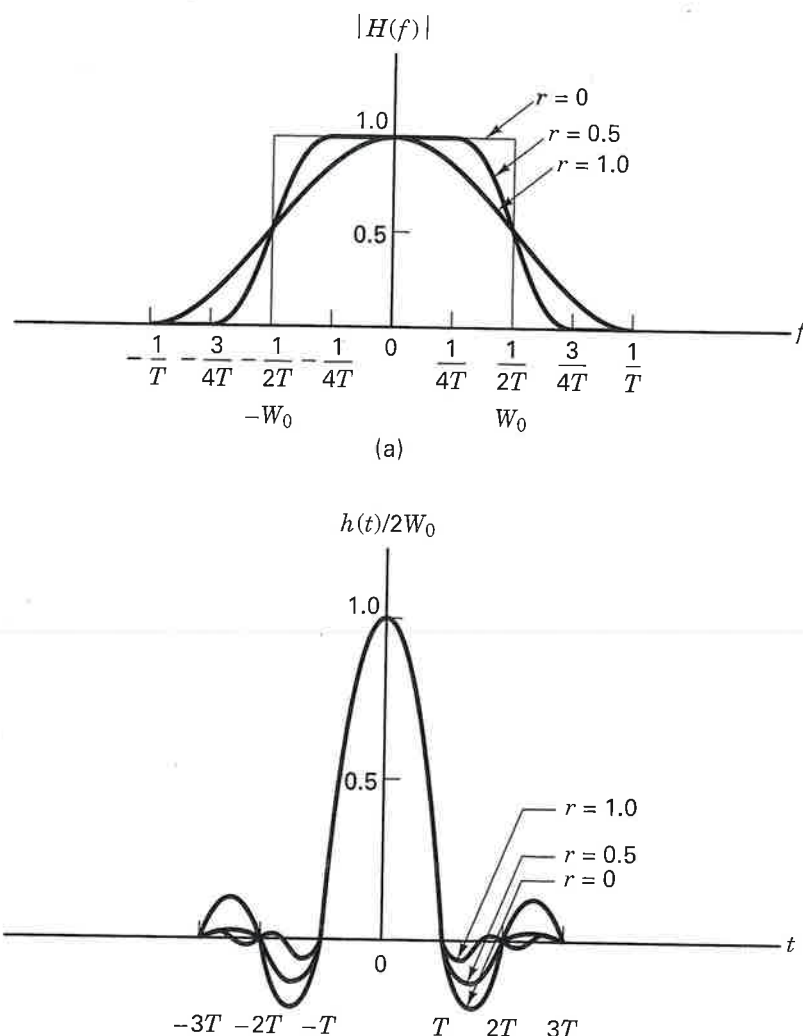


Figure 3.17 Raised-cosine filter characteristics. (a) System transfer function. (b) System impulse response.

$$h(t) = 2W_0(\text{sinc } 2W_0 t) \frac{\cos [2\pi(W - W_0)t]}{1 - [4(W - W_0)t]^2} \quad (3.79)$$

and is plotted in Figure 3.17b for $r=0$, $r=0.5$, and $r=1$. The tails have zero value at each pulse-sampling time, regardless of the roll-off value.

We can only approximately implement a filter described by Equation (3.78) and a pulse shape described by Equation (3.79), since, strictly speaking, the raised-cosine spectrum is not physically realizable (for the same reason that the ideal Nyquist filter is not realizable). A realizable filter must have an impulse response of finite duration and exhibit a zero output prior to the pulse turn-on time (see Sec-

tion 1.7.2), which is not the case for the family of raised-cosine characteristics. These unrealizable filters are *noncausal* (the filter impulse response has infinite duration, and the filtered pulse begins at time $t = -\infty$). A pulse-shaping filter should satisfy two requirements. It should provide the desired roll-off, and it should be realizable (the impulse response needs to be truncated to a finite length).

Starting with the Nyquist bandwidth constraint that the minimum required system bandwidth W for a symbol rate of R_s symbols/s without ISI is $R_s/2$ hertz, a more general relationship between required bandwidth and symbol transmission rate involves the filter roll-off factor r and can be stated as

$$W = \frac{1}{2}(1 + r)R_s \quad (3.80)$$

Thus, with $r = 0$, Equation (3.80) describes the minimum required bandwidth for ideal Nyquist filtering. For $r > 0$, there is a bandwidth expansion beyond the Nyquist minimum; thus, for this case, R_s is now less than twice the bandwidth. If the demodulator outputs one sample per symbol, then the Nyquist sampling theorem has been violated, since we are left with too few samples to reconstruct the analog waveform unambiguously (aliasing is present). However, for digital communication systems, we are not interested in reconstructing the analog waveform. Since the family of raised-cosine filters is characterized by zero ISI at the times that the symbols are sampled, we can still achieve unambiguous detection.

Bandpass-modulated signals (see Chapter 4), such as amplitude shift keying (ASK) and phase-shift keying (PSK), require twice the transmission bandwidth of the equivalent baseband signals. (See Section 1.7.1.) Such frequency-translated signals, occupying twice their baseband bandwidth, are often called double-sideband (DSB) signals. Therefore, for ASK- and PSK-modulated signals, the relationship between the required DSB bandwidth W_{DSB} and the symbol transmission rate R_s is

$$W_{\text{DSB}} = (1 + r)R_s \quad (3.81)$$

Recall that the raised-cosine frequency transfer function describes the composite $H(f)$ that is the "full round trip" from the inception of the message (as an impulse) at the transmitter, through the channel, and through the receiving filter. The filtering at the receiver is the compensating portion of the overall transfer function to help bring about zero ISI with an overall transfer function, such as the raised cosine. Often this is accomplished by choosing (matching) the receiving filter and the transmitting filter so that each has a transfer function known as a root-raised cosine (square root of the raised cosine). Neglecting any channel-induced ISI, the product of these root-raised-cosine functions yields the composite raised-cosine system transfer function. Whenever a separate equalizing filter is introduced to mitigate the effects of channel-induced ISI, the receiving and equalizing filters together should be configured to compensate for the distortion caused by both the transmitter and the channel so as to yield an overall system transfer function characterized by zero ISI.

Let's review the trade-off that faces us in specifying pulse-shaping filters. The larger the filter roll-off, the shorter will be the pulse tails (which implies smaller tail

amplitudes). Small tails exhibit less sensitivity to timing errors and thus make for small degradation due to ISI. Notice in Figure 3.17b, for $r = 1$, that timing errors can still result in some ISI degradation. However, the problem is not as serious as it is for the case in which $r = 0$, because the tails of the $h(t)$ waveform are of much smaller amplitude for $r = 1$ than they are for $r = 0$. The cost is more excess bandwidth. On the other hand, the smaller the filter roll-off, the smaller will be the excess bandwidth, thereby allowing us to increase the signaling rate or the number of users that can simultaneously use the system. The cost is longer pulse tails, larger pulse amplitudes, and thus, greater sensitivity to timing errors.

3.3.2 Two Types of Error-Performance Degradation

The effects of error-performance degradation in digital communications can be partitioned into two categories. The first is due to a decrease in received signal power or an increase in noise or interference power, giving rise to a loss in signal-to-noise ratio or E_b/N_0 . The second is due to signal distortion, such as might be caused by intersymbol interference (ISI). Let us demonstrate how different are the effects of these two degradation types.

Suppose that we require a communication system with a bit-error probability P_B versus E_b/N_0 characteristic corresponding to the solid-line curve plotted in Figure 3.18a. Suppose that after the system is configured and measurements are taken, we find, to our disappointment, that the performance does not follow the theoretical curve, but in fact follows the dashed line plot. A loss in E_b/N_0 has come about because of some signal losses or an increased level of noise or interference. For a

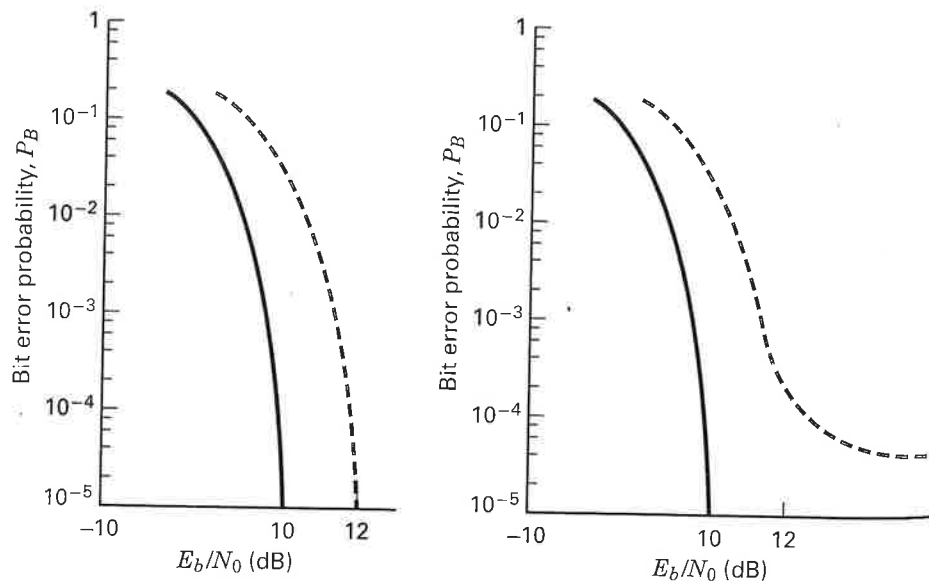
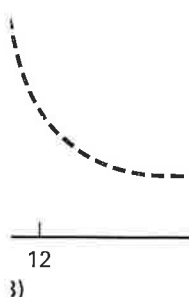


Figure 3.18 (a) Loss in E_b/N_0 . (b) Irreducible P_B caused by distortion.

and thus make for
that timing errors
not as serious as it
form are of much
it is more excess
smaller will be the
te or the number
pulse tails, larger

communications can be
received signal
a loss in signal-
much as might be
different are the

error probability
e plotted in Fig-
ments are taken,
low the theoreti-
has come about
interference. For a



ortion.

ction Chap. 3

desired bit-error probability of 10^{-5} , the theoretical required E_b/N_0 is 10 dB. Since our system performance falls short of our goal, we can see from the dashed-line curve that, for the same bit-error probability of 10^{-5} , the required E_b/N_0 is now 12 dB. If there were no way to remedy this problem, how much more E_b/N_0 would have to be provided in order to meet the required bit-error probability? The answer is 2 dB, of course. It might be a serious problem—especially if the system is power-limited, and it is difficult to come up with the additional 2 dB. But that loss in E_b/N_0 is not so terrible when compared with the possible effects of degradation caused by a distortion mechanism.

In Figure 3.18b, again imagine that we do not meet the desired performance of the solid-line curve. But instead of suffering a simple loss in signal-to-noise ratio, there is a degradation effect brought about by ISI (plotted with the dashed line). If there were no way to remedy this problem, how much more E_b/N_0 would be required in order to meet the desired bit-error probability? It would require an infinite amount—or, in other words, there is no amount of E_b/N_0 that will ameliorate this problem. More E_b/N_0 cannot help when the curve manifests such an irreducible P_B (assuming that the bottoming-out point is located above the system's required P_B). Undoubtedly, every P_B -versus- E_b/N_0 curve bottoms out somewhere, but if the bottoming-out point is well below the region of interest, it will be of no consequence.

More E_b/N_0 may not help the ISI problem (it won't help at all if the P_B curve has reached an irreducible level). This can be inferred by looking at the overlapped pulses in Figure 3.15b; if we increase the E_b/N_0 , the ratio of that overlap does not change. The pulses are subject to the same distortion. What, then, is the usual cure for the degradation effects of ISI? The cure is found in a technique called equalization. (See Section 3.4.) Since the distortion effects of ISI are caused by filtering in the transmitter and the channel, equalization can be thought of as the process that reverses such nonoptimum filtering effects.

Example 3.3 Bandwidth Requirements

- Find the minimum required bandwidth for the baseband transmission of a four-level PAM pulse sequence having a data rate of $R = 2400$ bits/s if the system transfer characteristic consists of a raised-cosine spectrum with 100% excess bandwidth ($r = 1$).
- The same 4-ary PAM sequence is modulated onto a carrier wave, so that the baseband spectrum is shifted and centered at frequency f_0 . Find the minimum required DSB bandwidth for transmitting the modulated PAM sequence. Assume that the system transfer characteristic is the same as in part (a).

Solution

- $M = 2^k$; since $M = 4$ levels, $k = 2$.

$$\text{Symbol or pulse rate } R_s = \frac{R}{k} = \frac{2400}{2} = 1200 \text{ symbols/s;}$$

$$\text{Minimum bandwidth } W = \frac{1}{2}(1 + r)R_s = \frac{1}{2}(2)(1200) = 1200 \text{ Hz.}$$

Figure 3.19a illustrates the baseband PAM received pulse in the time domain—an approximation to the $h(t)$ in Equation (3.79). Figure 3.19b illustrates the Fourier

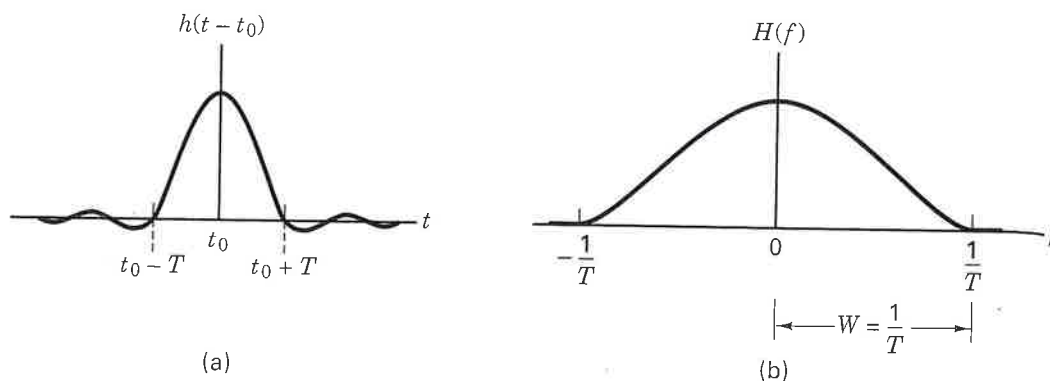


Figure 3.19 (a) Shaped pulse. (b) Baseband raised cosine spectrum.

transform of $h(t)$ —the raised cosine spectrum. Notice that the required bandwidth, W , starts at zero frequency and extends to $f = 1/T$; it is twice the size of the Nyquist theoretical minimum bandwidth.

(b) As in part (a),

$$R_s = 1200 \text{ symbols/s;}$$

$$W_{\text{DSB}} = (1 + r)R_s = 2(1200) = 2400 \text{ Hz.}$$

Figure 3.20a illustrates the modulated PAM received pulse. This waveform can be viewed as the product of a high-frequency sinusoidal carrier wave and a waveform with the pulse shape of Figure 3.19a. The single-sided spectral plot in Figure 3.20b illustrates that the modulated bandwidth is

$$W_{\text{DSB}} = \left(f_0 + \frac{1}{T}\right) - \left(f_0 - \frac{1}{T}\right) = \frac{2}{T}.$$

When the spectrum of Figure 3.19b is shifted up in frequency, the negative and positive halves of the baseband spectrum are shifted up in frequency, thereby dou-

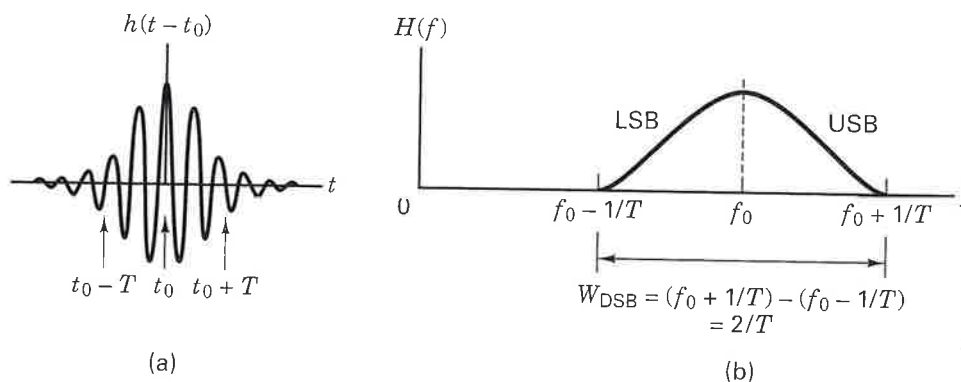


Figure 3.20 (a) Modulated shaped pulse. (b) DSB-modulated raised cosine spectrum.

bling the required transmission bandwidth. As the name implies, the DSB signal has two sidebands: the upper sideband (USB), derived from the baseband positive half, and the lower sideband (LSB), derived from the baseband negative half.

Example 3.4 Digital Telephone Circuits

Compare the system bandwidth requirements for a terrestrial 3-kHz analog telephone voice channel with that of a digital one. For the digital channel, the voice is formatted as a PCM bit stream, where the sampling rate for the analog-to-digital (A/D) conversion is 8000 samples/s and each voice sample is quantized to one of 256 levels. The bit stream is then transmitted using a PCM waveform and received with zero ISI.

Solution

The result of the sampling and quantization process yields PCM words such that each word (representing one sample) has one of $L = 256$ different levels. If each sample were sent as a 256-ary PAM pulse (symbol), then from Equation (3.80) we can write that the required system bandwidth (without ISI) for sending R_s symbols/s would be

$$W \geq \frac{R_s}{2} \text{ hertz}$$

where the equality sign holds true only for ideal Nyquist filtering. Since the digital telephone system uses PCM (binary) waveforms, each PCM word is converted to $\ell = \log_2 L = \log_2 256 = 8$ bits. Therefore, the system bandwidth required to transmit voice using PCM is

$$\begin{aligned} W_{\text{PCM}} &\geq (\log_2 L) \frac{R_s}{2} \text{ hertz} \\ &\geq \frac{1}{2} (8 \text{ bits/symbol}) (8000 \text{ symbols/s}) = 32 \text{ kHz} \end{aligned}$$

The 3-kHz analog voice channel will generally require approximately 4-kHz of bandwidth, including some bandwidth separation between channels, called *guard bands*. Therefore, the PCM format, using 8-bit quantization and binary signaling with a PCM waveform, requires at least eight times the bandwidth required for the analog channel.

3.3.3 Demodulation/Detection of Shaped Pulses

3.3.3.1 Matched Filters versus Conventional Filters

Conventional filters screen out unwanted spectral components of a received signal while maintaining some measure of fidelity for signals occupying a selected span of the spectrum, called the *pass-band*. These filters are generally designed to provide approximately uniform gain, a linear phase-versus-frequency characteristic over the pass-band, and a specified minimum attenuation over the remaining spectrum, called the *stop-band(s)*. A matched filter has a different "design priority," namely that of maximizing the SNR of a known signal in the presence of AWGN. Conventional filters are applied to random signals defined only by their bandwidth, while matched filters are applied to *known signals* with random parameters (such as amplitude and arrival time). The matched filter can be considered to be a *template* that is matched to the known shape of the signal being processed. A conven-

tional filter tries to preserve the temporal or spectral structure of the signal of interest. On the other hand, a matched filter significantly modifies the temporal structure by gathering the signal energy matched to its template, and, at the end of each symbol time, presenting the result as a peak amplitude. In general, a digital communications receiver processes received signals with both kinds of filters. The task of the conventional filter is to isolate and extract a high-fidelity estimate of the signal for presentation to the matched filter. The matched filter gathers the received signal energy, and when its output is sampled (at $t = T$), a voltage proportional to that energy is produced for subsequent detection and post-detection processing.

3.3.3.2 Nyquist Pulse and Square-Root Nyquist Pulse

Consider a sequence of data impulses at a transmitter input compared with the corresponding output sequence from a raised-cosine transfer function (before sampling). In Figure 3.21, transmitted data is represented by impulse waveforms that occur at times $\tau_0, \tau_1, \dots$. Filtering spreads the input waveforms, and thus delays them in time. We use the notation, $t_0, t_1, \dots$, to denote received time. The impulse event that was transmitted at time τ_0 arrives at the receiver at time t_0 corresponding to the start of the output pulse event. The premainlobe tail of a demodulated pulse is referred to as its *precursor*. For a real system with a fixed system-time reference, causality dictates that $t_0 \geq \tau_0$, and the time difference between τ_0 and t_0 represents any propagation delay in the system. In this example, the time duration from the start of a demodulated pulse precursor until the appearance of its mainlobe or peak amplitude is $3T$ (three pulse-time durations). Each output pulse in the sequence is superimposed with other pulses; each pulse has an effect on the main lobes of three earlier and three later pulses. When a pulse is filtered (shaped) so that it occupies more than one symbol time, we define the pulse *support time* as the total number of symbol intervals over which the pulse persists. In Figure 3.21, the pulse support time consists of 6-symbol intervals (7 data points with 6 intervals between them).

The impulse response of a root-raised cosine filter, called the *square-root Nyquist pulse*, is shown in Figure 3.22a (normalized to a peak value of unity, with a filter rolloff of $r = 0.5$). The impulse response of the raised-cosine filter, called the

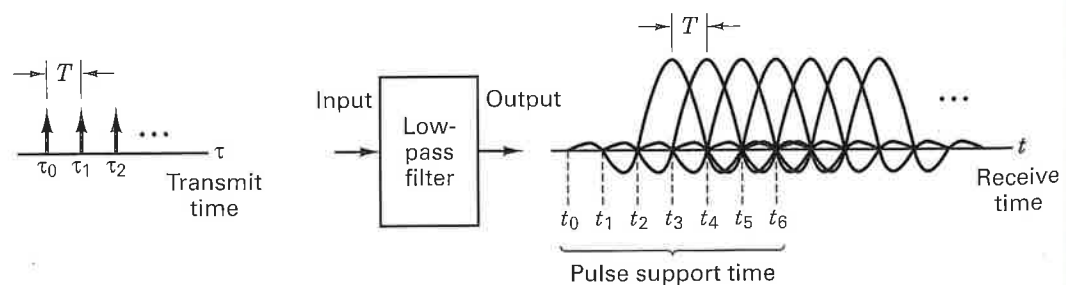


Figure 3.21 Filtered impulse sequence: output versus input.

ne signal of in-
the temporal
d, at the end of
neral, a digital
of filters. The
estimate of the
gathers the re-
oltage propor-
post-detection

compared with
unction (before
lse waveforms
and thus delays
e. The impulse
corresponding
odulated pulse
time reference,
and t_0 represents
ation from the
inlobe or peak
the sequence is
lobes of three
that it occupies
e total number
e pulse support
ween them).
he *square-root*
of unity, with a
ilter, called the

...
t
Receive
time

Nyquist pulse, is shown in Figure 3.22b (with the same normalization and filter rolloff). Inspecting these two pulse shapes, we see that they have a similar appearance, but the square-root Nyquist pulse makes slightly faster transitions, thus its spectrum (root-raised cosine) does not decay as rapidly as the spectrum (raised cosine) of the Nyquist pulse. Another subtle but important difference is that the square-root Nyquist pulse *does not* exhibit zero ISI (you can verify that the pulse tails in Figure 3.22a do not go through zero amplitude at the symbol times). However, if a root-raised cosine filter is used at both the transmitter and the receiver, the product of these transfer functions being a raised cosine, will give rise to an output having zero ISI.

It is interesting to see how the square-root Nyquist pulses appear at the output of a transmitter and how they appear after demodulation with a root-raised cosine MF. Figure 3.23a illustrates an example of sending a sequence of message symbols $\{+1 +1 -1 +3 +1 +3\}$ from a 4-ary set, where the members of the alphabet set are: $\{\pm 1, \pm 3\}$. Consider that the pulse modulation is 4-ary PAM, and that the pulses have been shaped with a root-raised cosine filter, having a roll-off value of 0.5. The analog waveform in this figure represents the transmitter output. Since the output waveform from any filter is delayed in time, then in Figure 3.23a, the input message symbols (shown as approximate impulses) have been delayed the same amount as the output waveform in order to align the message sequence with its corresponding filtered waveform (the square-root Nyquist shaped-pulse sequence). This is just a visual convenience so that the reader can compare the filter input with its output. It is, of course, only the output analog waveform that is transmitted (or modulated) onto a carrier wave.

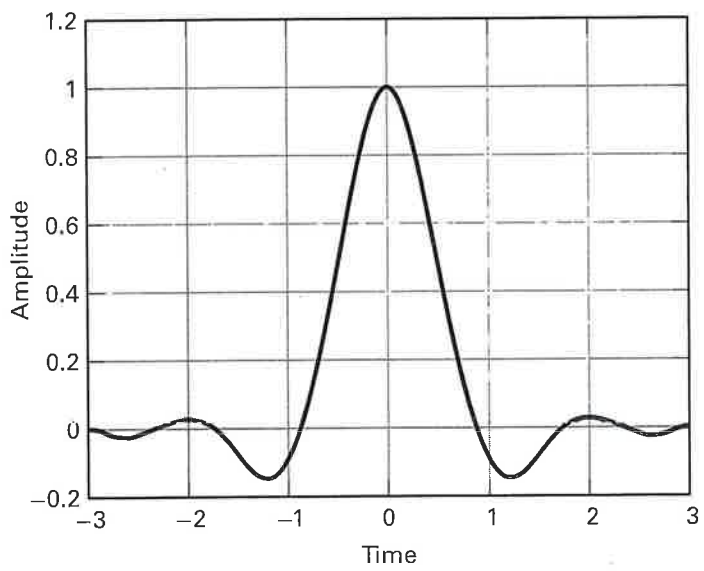


Figure 3.22a Square-root Nyquist pulse.

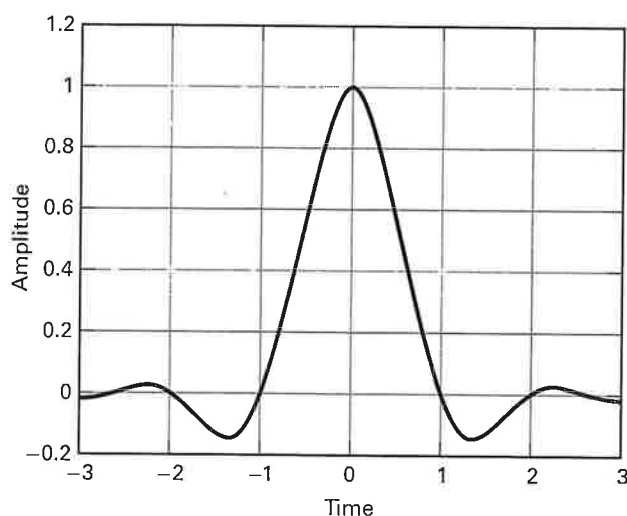


Figure 3.22b Nyquist pulse.

Figure 3.23b shows the same delayed message samples together with the output waveform from the root-raised cosine MF, yielding a raised-cosine transfer function for the overall system. Let us describe a simple test to determine if the filtered output (assuming no noise) contains ISI. It is only necessary to sample the filtered waveform at the times corresponding to the original input samples; if

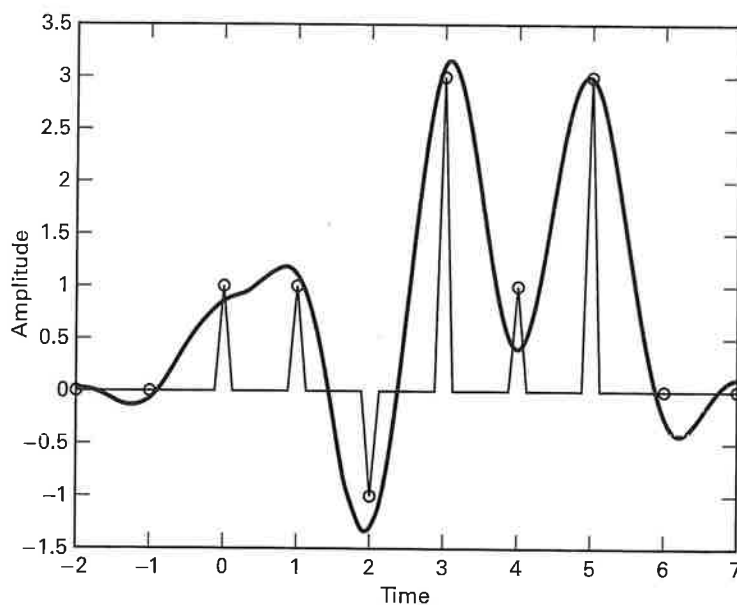


Figure 3.23a Square-root Nyquist-shaped M -ary waveform and delayed-input sample values.

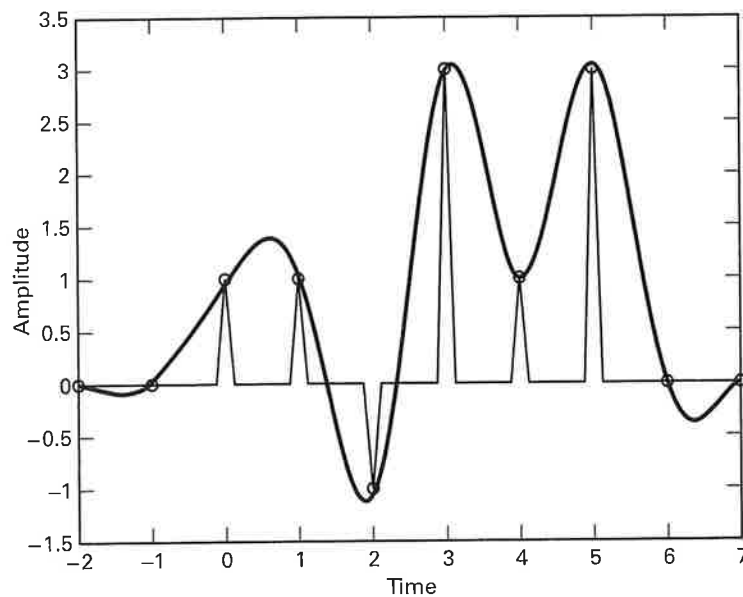


Figure 3.23b Output of raised-cosine matched filter and delayed-input sample values.

the resulting sample values are unchanged from those of the original message, then the filter output has zero ISI (at the sample times). When Figures 3.23a and 3.23b are compared with regard to ISI, it should be apparent that sampling the square-root Nyquist waveform of Figure 3.23a (transmitter output) *will not* yield the exact original samples; however, sampling the Nyquist waveform in Figure 3.23b (MF output) *will* yield the exact original samples. This supports the statement that a Nyquist filter yields zero ISI at the sample points, while any other filter does not do so.

3.4 EQUALIZATION

3.4.1 Channel Characterization

Many communication channels (e.g., telephone, wireless) can be characterized as band-limited linear filters with an impulse response $h_c(t)$ and a frequency response

$$H_c(f) = |H_c(f)|e^{j\theta_c(f)} \quad (3.82)$$

where $h_c(t)$ and $H_c(f)$ are Fourier transform pairs, $|H_c(f)|$ is the channel's amplitude response, and $\theta_c(f)$ is the channel's phase response. In order to achieve ideal (nondistorting) transmission characteristics over a channel, it was shown in Section 1.6.3, that within a signal's bandwidth W , $|H_c(f)|$ *must be constant*. Also, $\theta_c(f)$ *must be a linear function of frequency*, which is tantamount to saying that the delay must be constant for all spectral components of the signal. If $|H_c(f)|$ is not constant

within W , then the effect is amplitude distortion. If $\theta_c(f)$ is not a linear function of frequency within W , then the effect is phase distortion. For many channels that exhibit distortion of this type, such as fading channels, amplitude and phase distortion typically occur together. For a transmitted sequence of pulses, such distortion manifests itself as a signal dispersion or "smearing" so that any one pulse in the received demodulated sequence is not well defined. The overlap or smearing, known as *intersymbol interference* (ISI), described in Section 3.3, arises in most modulation systems; it is one of the major obstacles to reliable high-speed data transmission over bandlimited channels. In the broad sense, the name "equalization" refers to any signal processing or filtering technique that is designed to eliminate or reduce ISI.

In Figure 2.1, equalization is partitioned into two broad categories. The first category, *maximum-likelihood sequence estimation* (MLSE), entails making measurements of $h_c(t)$ and then providing a means for adjusting the receiver to the transmission environment. The goal of such adjustments is to enable the detector to make good estimates from the demodulated distorted pulse sequence. With an MLSE receiver, the distorted samples are not reshaped or directly compensated in any way; instead, the mitigating technique for the MLSE receiver is to adjust itself in such a way that it can better deal with the distorted samples. (An example of this method, known as Viterbi equalization, is treated in Section 15.7.1.) The second category, *equalization with filters*, uses filters to compensate the distorted pulses. In this second category, the detector is presented with a sequence of demodulated samples that the equalizer has modified or "cleaned up" from the effects of ISI. Equalizing with filters, the more popular approach and the one described in this section, lends itself to further partitioning. The filters can be described as to whether they are linear devices that contain only feedforward elements (*transversal equalizers*), or whether they are nonlinear devices that contain both feedforward and feedback elements (*decision feedback equalizers*). They can be grouped according to the automatic nature of their operation, which may be either *preset* or *adaptive*. They also can be grouped according to the filter's resolution or update rate. Are predetection samples provided only on symbol boundaries, that is, one sample per symbol? If so, the condition is known as *symbol spaced*. Are multiple samples provided for each symbol? If so, this condition is known as *fractionally spaced*.

We now modify Equation (3.77) by letting the receiving/equalizing filter be replaced by a separate receiving filter and equalizing filter, defined by frequency transfer functions $H_r(f)$ and $H_e(f)$, respectively. Also, let the overall system transfer function $H(f)$ be a raised-cosine filter, designated $H_{RC}(f)$. Thus, we write

$$H_{RC}(f) = H_t(f) H_c(f) H_r(f) H_e(f) \quad (3.83)$$

In practical systems, the channel's frequency transfer function $H_c(f)$ and its impulse response $h_c(t)$ are not known with sufficient precision to allow for a receiver design to yield zero ISI for all time. Usually, the transmit and receive filters are chosen to be matched so that

$$H_{RC}(f) = H_t(f) H_r(f) \quad (3.84)$$

In this way, $H_t(f)$ and $H_r(f)$ each have frequency transfer functions that are the square root of the raised cosine (root-raised cosine). Then, the equalizer transfer function needed to compensate for channel distortion is simply the inverse of the channel transfer function:

$$H_e(f) = \frac{1}{H_c(f)} = \frac{1}{|H_c(f)|} e^{-j\theta_c(f)} \quad (3.85)$$

Sometimes a system frequency transfer function manifesting ISI at the sampling points is purposely chosen (e.g., a Gaussian filter transfer function). The motivation for such a transfer function is to improve bandwidth efficiency, compared with using a raised-cosine filter. When such a design choice is made, the role of the equalizing filter is not only to compensate for the channel-induced ISI, but also to compensate for the ISI brought about by the transmitter and receiver filters [7].

3.4.2 Eye Pattern

An eye pattern is the display that results from measuring a system's response to baseband signals in a prescribed way. On the vertical plates of an oscilloscope we connect the receiver's response to a random pulse sequence. On the horizontal plates we connect a sawtooth wave at the signaling frequency. In other words, the horizontal time base of the oscilloscope is set equal to the symbol (pulse) duration. This setup superimposes the waveform in each signaling interval into a family of traces in a single interval $(0, T)$. Figure 3.24 illustrates the eye pattern that results for binary antipodal (bipolar pulse) signaling. Because the symbols stem from a random source, they are sometimes positive and sometimes negative, and the persistence of the cathode ray tube display allows us to see the resulting pattern shaped as an eye. The width of the opening indicates the time over which sampling for detection might be performed. Of course, the optimum sampling time corresponds to the maximum eye opening, yielding the greatest protection against noise.

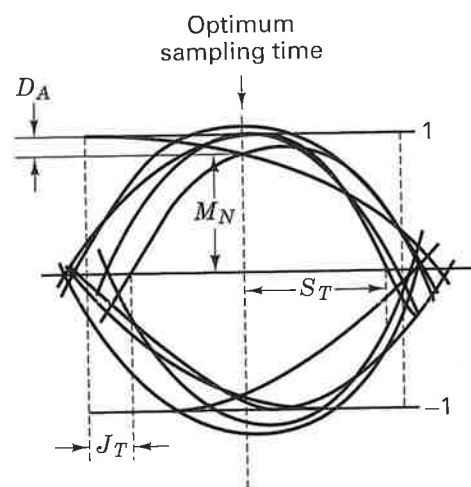


Figure 3.24 Eye Pattern.

If there were no filtering in the system—that is, if the bandwidth corresponding to the transmission of these data pulses were infinite—then the system response would yield ideal rectangular pulse shapes. In that case, the pattern would look like a box rather than an eye. In Figure 3.24, the range of amplitude differences labelled D_A is a measure of distortion caused by ISI, and the range of time differences of the zero crossings labelled J_T is a measure of the timing jitter. Measures of noise margin M_N and sensitivity-to-timing error S_T are also shown in the figure. In general, the most frequent use of the eye pattern is for qualitatively assessing the extent of the ISI. As the eye closes, ISI is increasing; as the eye opens, ISI is decreasing.

3.4.3 Equalizer Filter Types

3.4.3.1 Transversal Equalizer

A training sequence used for equalization is often chosen to be a noise-like sequence, “rich” in spectral content, which is needed to estimate the channel frequency response. In the simplest sense, training might consist of sending a single narrow pulse (approximating an ideal impulse) and thereby learning the impulse response of the channel. In practice, a pseudonoise (PN) signal is preferred over a single pulse for the training sequence because the PN signal has larger average power and hence larger SNR for the same peak transmitted power. For describing the transversal filter, consider that a single pulse was transmitted over a system designated to have a raised-cosine transfer function $H_{RC}(f) = H_t(f) H_r(f)$. Also consider that the channel induces ISI, so that the received demodulated pulse exhibits distortion, as shown in Figure 3.25, such that the pulse sidelobes do not go through zero at sample times adjacent to the mainlobe of the pulse. The distortion can be viewed as positive or negative echoes occurring both before and after the mainlobe. To achieve the desired raised-cosine transfer function, the equalizing filter should have a frequency response $H_e(f)$, as shown in Equation (3.85), such that the actual channel response when multiplied by $H_e(f)$ yields $H_{RC}(f)$. In other words, we would like the equalizing filter to generate a set of canceling echoes.

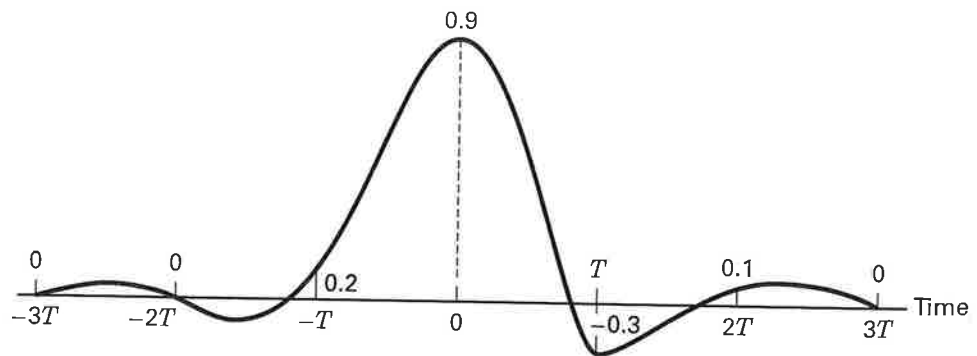


Figure 3.25 Received pulse exhibiting distortion.

responding to
m response
uld look like
ices labelled
rences of the
of noise mar-
. In general,
the extent of
easing.

a noise-like
channel fre-
ding a single
; the impulse
ferred over a
rger average
or describing
ver a system
) $H_r(f)$. Also
ted pulse ex-
es do not go
The distortion
and after the
equalizing fil-
85), such that
(f). In other
elting echoes.

Since we are interested in sampling the equalized waveform at only a few predetermined sampling times, then the design of an equalizing filter can be a straightforward task.

The transversal filter, depicted in Figure 3.26, is the most popular form of an easily adjustable equalizing filter consisting of a delay line with T -second taps (where T is the symbol duration). In such an equalizer, the current and past values of the received signal are linearly weighted with equalizer coefficients or tap weights $\{c_n\}$ and are then summed to produce the output. The main contribution is from a central tap, with the other taps contributing echoes of the main signal at symbol intervals on either side of the main signal. If it were possible for the filter to have an infinite number of taps, then the tap weights could be chosen to force the system impulse response to zero at all but one of the sampling times, thus making $H_e(f)$ correspond exactly to the inverse of the channel transfer function in Equation (3.85). Even though an infinite length filter is not realizable, one can still specify practical filters that approximate the ideal case.

In Figure 3.26, the outputs of the weighted taps are amplified, summed, and fed to a coefficient-adjustment device. The tap weights $\{c_n\}$ need to be chosen so as to subtract the effects of interference from symbols adjacent in time to the desired symbol. Consider that there are $(2N + 1)$ taps with weights $c_{-N}, c_{-N+1}, \dots, c_N$. Output samples $\{z(k)\}$ of the equalizer are then found by convolving the input samples $\{x(k)\}$ and tap weights $\{c_n\}$ as follows:

$$z(k) = \sum_{n=-N}^N x(k-n) c_n \quad k = -2N, \dots, 2N \quad n = -N, \dots, N \quad (3.86)$$

where $k = 0, \pm 1, \pm 2, \dots$ is a time index that is shown in parentheses. (Time may take on any range of values.) The index n is used two ways—as a time offset, and as a filter coefficient identifier (which is an address in the filter). When used in the latter

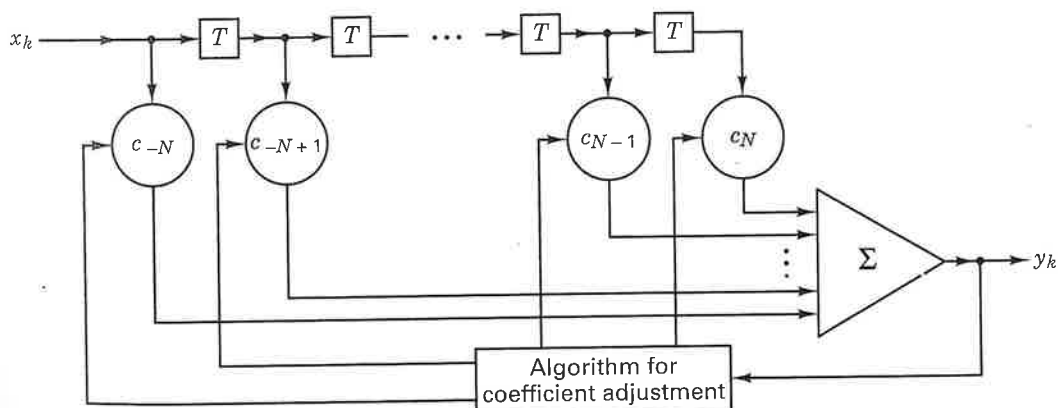


Figure 3.26 Transversal filter.

0
Time
3T

sense, it is shown as a subscript. By defining the vectors $\mathbf{z}$ and $\mathbf{c}$ and the matrix $\mathbf{x}$ as, respectively,

$$\mathbf{z} = \begin{bmatrix} z(-2N) \\ \vdots \\ z(0) \\ \vdots \\ z(2N) \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} c_{-N} \\ \vdots \\ c_0 \\ \vdots \\ c_N \end{bmatrix} \quad (3.87)$$

and

$$\mathbf{x} = \begin{bmatrix} x(-N) & 0 & 0 & \cdots & 0 & 0 \\ x(-N+1) & x(-N) & 0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x(N) & x(N-1) & x(N-2) & \cdots & x(-N+1) & x(-N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & x(N) & x(N-1) \\ 0 & 0 & 0 & \cdots & 0 & x(N) \end{bmatrix} \quad (3.88)$$

we can describe the relationship among $\{z(k)\}$, $\{x(k)\}$, and $\{c_n\}$ more compactly as

$$\mathbf{z} = \mathbf{x} \mathbf{c} \quad (3.89a)$$

And whenever the matrix $\mathbf{x}$ is square, with its rows and columns each having the same dimension as the number of elements in $\mathbf{c}$, we can find $\mathbf{c}$ by solving the following equation:

$$\mathbf{c} = \mathbf{x}^{-1} \mathbf{z} \quad (3.89b)$$

Notice that the size of the vector $\mathbf{z}$ and the number of rows in the matrix $\mathbf{x}$ may be chosen to be any value, because one might be interested in the ISI at sample points far removed from the mainlobe of the pulse in question. In Equations (3.86) through (3.88), the index k was arbitrarily chosen to allow for $4N+1$ sample points. The vectors $\mathbf{z}$ and $\mathbf{c}$ have dimensions $4N+1$ and $2N+1$, respectively, and the matrix $\mathbf{x}$ is nonsquare with dimensions $4N+1$ by $2N+1$. Such equations are referred to as an *overdetermined set* (i.e., there are more equations than unknowns). One can solve such a problem in a *deterministic way* known as the *zero-forcing solution*, or, in a *statistical way*, known as the *minimum mean-square error (MSE) solution*.

Zero-Forcing Solution. This solution starts by disposing of the top N and bottom N rows of the matrix $\mathbf{x}$ in Equation (3.88), thereby transforming $\mathbf{x}$ into a square matrix of dimension $2N+1$ by $2N+1$, transforming $\mathbf{z}$ into a vector of dimension $2N+1$, and yielding in Equation (3.89a), a deterministic set of $2N+1$ simultaneous equations. This zero-forcing solution minimizes the peak ISI distortion by selecting the $\{c_n\}$ weights so that the equalizer output is forced to zero at N sample points on either side of the desired pulse. In other words, the weights are chosen so that