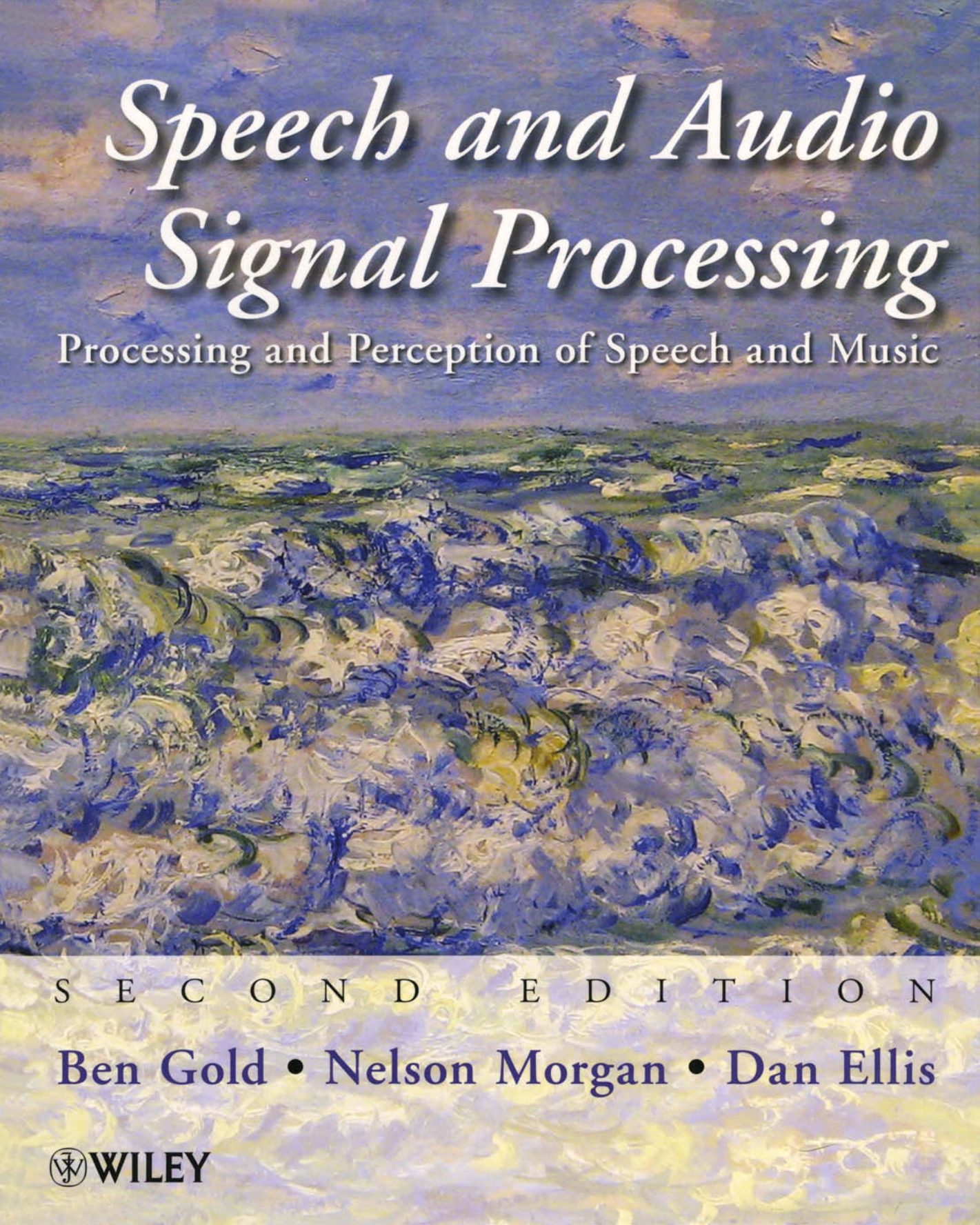


The background of the book cover is an impressionistic painting of a landscape. The upper portion shows a sky with soft, blended colors of blue, purple, and white. Below the sky, the landscape is depicted with thick, visible brushstrokes in shades of blue, green, and yellow, suggesting a field or a body of water with reeds or grass. The overall style is reminiscent of J.M.W. Turner's work.

Speech and Audio Signal Processing

Processing and Perception of Speech and Music

S E C O N D E D I T I O N

Ben Gold • Nelson Morgan • Dan Ellis

 **WILEY**

This page intentionally left blank

*SPEECH AND AUDIO
SIGNAL PROCESSING*

This page intentionally left blank

SPEECH AND AUDIO SIGNAL PROCESSING

Processing and Perception
of Speech and Music

Second Edition

BEN GOLD

*Massachusetts Institute of Technology
Lincoln Laboratory*

NELSON MORGAN

*International Computer Science Institute
and University of California at Berkeley*

DAN ELLIS

*Columbia University
and International Computer Science Institute*

with contributions from:

Hervé Bourlard
Eric Fosler-Lussier
Gerald Friedland
Jeff Gilbert
Simon King
David van Leeuwen
Michael Seltzer
Steven Wegman



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2011 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 978-0-470-19536-9

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

*This book is dedicated to
our families
and our students*

This page intentionally left blank

CONTENTS

PREFACE TO THE 2011 EDITION xxi

- 0.1 Why We Created a New Edition **xxi**
- 0.2 What is New **xxi**
- 0.3 A Final Thought **xxii**

CHAPTER 1 *INTRODUCTION* 1

- 1.1 Why We Wrote This Book 1
- 1.2 How to Use This Book 2
- 1.3 A Confession 4
- 1.4 Acknowledgments 5

PART I

HISTORICAL BACKGROUND

CHAPTER 2 *SYNTHETIC AUDIO: A BRIEF HISTORY* 9

- 2.1 Von Kempelen 9
- 2.2 The Voder 9
- 2.3 Teaching the Operator to Make the Voder “Talk” 11
- 2.4 Speech Synthesis After the Voder 13
- 2.5 Music Machines 13
- 2.6 Exercises 17

CHAPTER 3 *SPEECH ANALYSIS AND SYNTHESIS OVERVIEW* 21

- 3.1 Background 21
 - 3.1.1 Transmission of Acoustic Signals 21
 - 3.1.2 Acoustical Telegraphy before Morse Code 22
 - 3.1.3 The Telephone 23
 - 3.1.4 The Channel Vocoder and Bandwidth Compression 23
- 3.2 Voice-coding concepts 25
- 3.3 Homer Dudley (1898–1981) 29
- 3.4 Exercises 36
- 3.5 Appendix: Hearing of the Fall of Troy 37

CHAPTER 4 *BRIEF HISTORY OF AUTOMATIC SPEECH RECOGNITION* **40**

-
- 4.1 Radio Rex **40**
 - 4.2 Digit Recognition **42**
 - 4.3 Speech Recognition in the 1950s **43**
 - 4.4 The 1960s **43**
 - 4.4.1 Short-Term Spectral Analysis **45**
 - 4.4.2 Pattern Matching **45**
 - 4.5 1971–1976 ARPA Project **46**
 - 4.6 Achieved by 1976 **46**
 - 4.7 The 1980s in Automatic Speech Recognition **47**
 - 4.7.1 Large Corpora Collection **47**
 - 4.7.2 Front Ends **48**
 - 4.7.3 Hidden Markov Models **48**
 - 4.7.4 The Second (D)ARPA Speech-Recognition Program **49**
 - 4.7.5 The Return of Neural Nets **50**
 - 4.7.6 Knowledge-Based Approaches **50**
 - 4.8 More Recent Work **51**
 - 4.9 Some Lessons **53**
 - 4.10 Exercises **54**

CHAPTER 5 *SPEECH-RECOGNITION OVERVIEW* **59**

-
- 5.1 Why Study Automatic Speech Recognition? **59**
 - 5.2 Why is Automatic Speech Recognition Hard? **60**
 - 5.3 Automatic Speech Recognition Dimensions **62**
 - 5.3.1 Task Parameters **62**
 - 5.3.2 Sample Domain: Letters of the Alphabet **63**
 - 5.4 Components of Automatic Speech Recognition **64**
 - 5.5 Final Comments **67**
 - 5.6 Exercises **69**

PART II*MATHEMATICAL BACKGROUND***CHAPTER 6** *DIGITAL SIGNAL PROCESSING* **73**

-
- 6.1 Introduction **73**
 - 6.2 The z Transform **73**
 - 6.3 Inverse z Transform **74**
 - 6.4 Convolution **75**
 - 6.5 Sampling **76**
 - 6.6 Linear Difference Equations **77**
 - 6.7 First-Order Linear Difference Equations **78**

6.8	Resonance	79
6.9	Concluding Comments	83
6.10	Exercises	84

CHAPTER 7	<i>DIGITAL FILTERS AND DISCRETE FOURIER TRANSFORM</i>	87
------------------	--	-----------

7.1	Introduction	87
7.2	Filtering Concepts	88
7.3	Transformations for Digital Filter Design	92
7.4	Digital Filter Design with Bilinear Transformation	93
7.5	The Discrete Fourier Transform	94
7.6	Fast Fourier Transform Methods	98
7.7	Relation Between the DFT and Digital Filters	100
7.8	Exercises	101

CHAPTER 8	<i>PATTERN CLASSIFICATION</i>	105
------------------	--------------------------------------	------------

8.1	Introduction	105
8.2	Feature Extraction	107
8.2.1	Some Opinions	108
8.3	Pattern-Classification Methods	109
8.3.1	Minimum Distance Classifiers	109
8.3.2	Discriminant Functions	111
8.3.3	Generalized Discriminators	112
8.4	Support Vector Machines	115
8.5	Unsupervised Clustering	117
8.6	Conclusions	118
8.7	Exercises	118
8.8	Appendix: Multilayer Perceptron Training	119
8.8.1	Definitions	119
8.8.2	Derivation	120

CHAPTER 9	<i>STATISTICAL PATTERN CLASSIFICATION</i>	124
------------------	--	------------

9.1	Introduction	124
9.2	A Few Definitions	124
9.3	Class-Related Probability Functions	125
9.4	Minimum Error Classification	126
9.5	Likelihood-Based MAP Classification	127
9.6	Approximating a Bayes Classifier	128
9.7	Statistically Based Linear Discriminants	130
9.7.1	Discussion	131
9.8	Iterative Training: The EM Algorithm	131
9.8.1	Discussion	136
9.9	Exercises	137

PART III
ACOUSTICS

CHAPTER 10 *WAVE BASICS* **141**

- 10.1 Introduction **141**
- 10.2 The Wave Equation for the Vibrating String **142**
- 10.3 Discrete-Time Traveling Waves **143**
- 10.4 Boundary Conditions and Discrete Traveling Waves **144**
- 10.5 Standing Waves **144**
- 10.6 Discrete-Time Models of Acoustic Tubes **146**
- 10.7 Acoustic Tube Resonances **147**
- 10.8 Relation of Tube Resonances to Formant Frequencies **148**
- 10.9 Exercises **150**

CHAPTER 11 *ACOUSTIC TUBE MODELING OF SPEECH PRODUCTION* **152**

- 11.1 Introduction **152**
- 11.2 Acoustic Tube Models of English Phonemes **152**
- 11.3 Excitation Mechanisms in Speech Production **156**
- 11.4 Exercises **157**

CHAPTER 12 *MUSICAL INSTRUMENT ACOUSTICS* **158**

- 12.1 Introduction **158**
- 12.2 Sequence of Steps in a Plucked or Bowed String Instrument **159**
- 12.3 Vibrations of the Bowed String **159**
- 12.4 Frequency-Response Measurements of the Bridge of a Violin **160**
- 12.5 Vibrations of the Body of String Instruments **163**
- 12.6 Radiation Pattern of Bowed String Instruments **167**
- 12.7 Some Considerations in Piano Design **169**
- 12.8 The Trumpet, Trombone, French Horn, and Tuba **175**
- 12.9 Exercises **177**

CHAPTER 13 *ROOM ACOUSTICS* **179**

- 13.1 Introduction **179**
- 13.2 Sound Waves **179**
 - 13.2.1 One-Dimensional Wave Equation **180**
 - 13.2.2 Spherical Wave Equation **180**
 - 13.2.3 Intensity **181**
 - 13.2.4 Decibel Sound Levels **182**
 - 13.2.5 Typical Power Sources **182**

- 13.3 Sound Waves in Rooms 183
 - 13.3.1 Acoustic Reverberation 184
 - 13.3.2 Early Reflections 187
- 13.4 Room Acoustics as a Component in Speech Systems 188
- 13.5 Exercises 189

PART IV

AUDITORY PERCEPTION

CHAPTER 14 EAR PHYSIOLOGY 193

- 14.1 Introduction 193
- 14.2 Anatomical Pathways From the Ear to the Perception of Sound 193
- 14.3 The Peripheral Auditory System 195
- 14.4 Hair Cell and Auditory Nerve Functions 196
- 14.5 Properties of the Auditory Nerve 198
- 14.6 Summary and Block Diagram of the Peripheral Auditory System 205
- 14.7 Exercises 207

CHAPTER 15 PSYCHOACOUSTICS 209

- 15.1 Introduction 209
- 15.2 Sound-Pressure Level and Loudness 210
- 15.3 Frequency Analysis and Critical Bands 212
- 15.4 Masking 214
- 15.5 Summary 216
- 15.6 Exercises 217

CHAPTER 16 MODELS OF PITCH PERCEPTION 218

- 16.1 Introduction 218
- 16.2 Historical Review of Pitch-Perception Models 218
- 16.3 Physiological Exploration of Place Versus Periodicity 223
- 16.4 Results from Psychoacoustic Testing and Models 224
- 16.5 Summary 228
- 16.6 Exercises 230

CHAPTER 17 SPEECH PERCEPTION 232

- 17.1 Introduction 232
- 17.2 Vowel Perception: Psychoacoustics and Physiology 232
- 17.3 The Confusion Matrix 235
- 17.4 Perceptual Cues for Plosives 238
- 17.5 Physiological Studies of Two Voiced Plosives 239

- 17.6 Motor Theories of Speech Perception 241
- 17.7 Neural Firing Patterns for Connected Speech Stimuli 243
- 17.8 Concluding Thoughts 244
- 17.9 Exercises 247

CHAPTER 18 *HUMAN SPEECH RECOGNITION* 250

- 18.1 Introduction 250
- 18.2 The Articulation Index and Human Recognition 250
 - 18.2.1 The Big Idea 250
 - 18.2.2 The Experiments 251
 - 18.2.3 Discussion 252
- 18.3 Comparisons Between Human and Machine Speech Recognizers 253
- 18.4 Concluding Thoughts 256
- 18.5 Exercises 258

PART V

SPEECH FEATURES

CHAPTER 19 *THE AUDITORY SYSTEM AS A FILTER BANK* 263

- 19.1 Introduction 263
- 19.2 Review of Fletcher's Critical Band Experiments 263
- 19.3 Threshold Measurements and Filter Shapes 265
- 19.4 Gamma-Tone Filters, Roex Filters, and Auditory Models 270
- 19.5 Other Considerations in Filter-Bank Design 272
- 19.6 Speech Spectrum Analysis Using the FFT 274
- 19.7 Conclusions 275
- 19.8 Exercises 275

CHAPTER 20 *THE CEPSTRUM AS A SPECTRAL ANALYZER* 277

- 20.1 Introduction 277
- 20.2 A Historical Note 277
- 20.3 The Real Cepstrum 278
- 20.4 The Complex Cepstrum 279
- 20.5 Application of Cepstral Analysis to Speech Signals 281
- 20.6 Concluding Thoughts 283
- 20.7 Exercises 284

CHAPTER 21 *LINEAR PREDICTION* 286

- 21.1 Introduction 286
- 21.2 The Predictive Model 286

21.3	Properties of the Representation	290
21.4	Getting the Coefficients	292
21.5	Related Representations	294
21.6	Concluding Discussion	295
21.7	Exercises	297

PART VI

AUTOMATIC SPEECH RECOGNITION

CHAPTER 22 *FEATURE EXTRACTION FOR ASR* 301

22.1	Introduction	301
22.2	Common Feature Vectors	301
22.3	Dynamic Features	306
22.4	Strategies for Robustness	307
22.4.1	Robustness to Convolutional Error	307
22.4.2	Robustness to Room Reverberation	309
22.4.3	Robustness to Additive Noise	311
22.4.4	Caveats	313
22.5	Auditory Models	313
22.6	Multichannel Input	314
22.7	Discriminant Features	315
22.8	Discussion	315
22.9	Exercises	316

CHAPTER 23 *LINGUISTIC CATEGORIES FOR SPEECH RECOGNITION* 319

23.1	Introduction	319
23.2	Phones and Phonemes	319
23.2.1	Overview	319
23.2.2	What Makes a Phone?	320
23.2.3	What Makes a Phoneme?	321
23.3	Phonetic and Phonemic Alphabets	321
23.4	Articulatory Features	322
23.4.1	Consonants	322
23.4.2	Vowels	326
23.4.3	Why Use Features?	327
23.5	Subword Units as Categories for ASR	327
23.6	Phonological Models for ASR	329
23.6.1	Phonological rules	329
23.6.2	Pronunciation rule induction	329
23.7	Context-Dependent Phones	330
23.8	Other Subword Units	331
23.8.1	Properties in Fluent Speech	332

- 23.9 Phrases 332
- 23.10 Some Issues in Phonological Modeling 332
- 23.11 Exercises 334

CHAPTER 24 *DETERMINISTIC SEQUENCE RECOGNITION FOR ASR* 337

- 24.1 Introduction 337
- 24.2 Isolated Word Recognition 338
 - 24.2.1 Linear Time Warp 339
 - 24.2.2 Dynamic Time Warp 340
 - 24.2.3 Distances 344
 - 24.2.4 End-Point Detection 344
- 24.3 Connected Word Recognition 346
- 24.4 Segmental Approaches 347
- 24.5 Discussion 348
- 24.6 Exercises 349

CHAPTER 25 *STATISTICAL SEQUENCE RECOGNITION* 350

- 25.1 Introduction 350
- 25.2 Stating the Problem 351
- 25.3 Parameterization and Probability Estimation 353
 - 25.3.1 Markov Models 354
 - 25.3.2 Hidden Markov Model 356
 - 25.3.3 HMMs for Speech Recognition 357
 - 25.3.4 Estimation of $P(X|M)$ 358
- 25.4 Conclusion 362
- 25.5 Exercises 363

CHAPTER 26 *STATISTICAL MODEL TRAINING* 364

- 26.1 Introduction 364
- 26.2 HMM Training 365
- 26.3 Forward–Backward Training 368
- 26.4 Optimal Parameters for Emission Probability Estimators 371
 - 26.4.1 Gaussian Density Functions 371
 - 26.4.2 Example: Training with Discrete Densities 372
- 26.5 Viterbi Training 373
 - 26.5.1 Example: Training with Gaussian Density Functions 375
 - 26.5.2 Example: Training with Discrete Densities 375
- 26.6 Local Acoustic Probability Estimators for ASR 376
 - 26.6.1 Discrete Probabilities 376
 - 26.6.2 Gaussian Densities 377
 - 26.6.3 Tied Mixtures of Gaussians 377
 - 26.6.4 Independent Mixtures of Gaussians 377

26.6.5	Neural Networks	377
26.7	Initialization	378
26.8	Smoothing	378
26.9	Conclusions	379
26.10	Exercises	379

CHAPTER 27 *DISCRIMINANT ACOUSTIC PROBABILITY ESTIMATION* **381**

27.1	Introduction	381
27.2	Discriminant Training	382
27.2.1	Maximum Mutual Information	383
27.2.2	Corrective Training	383
27.2.3	Generalized Probabilistic Descent	384
27.2.4	Direct Estimation of Posteriors	385
27.3	HMM-ANN Based ASR	388
27.3.1	MLP Architecture	388
27.3.2	MLP Training	388
27.3.3	Embedded Training	389
27.4	Other Applications of ANNs to ASR	390
27.5	Exercises	391
27.6	Appendix: Posterior Probability Proof	391

CHAPTER 28 *ACOUSTIC MODEL TRAINING: FURTHER TOPICS* **394**

28.1	Introduction	394
28.2	Adaptation	394
28.2.1	MAP and MLLR	394
28.2.2	Speaker Adaptive Training	399
28.2.3	Vocal tract length normalization	401
28.3	Lattice-Based MMI and MPE	402
28.3.1	Details of mean estimation using lattice-based MMI and MPE	405
28.4	Conclusion	412
28.5	Exercises	413

CHAPTER 29 *SPEECH RECOGNITION AND UNDERSTANDING* **416**

29.1	Introduction	416
29.2	Phonological Models	417
29.3	Language Models	419
29.3.1	n -Gram Statistics	421
29.3.2	Smoothing	422
29.4	Decoding With Acoustic and Language Models	423
29.5	A Complete System	424
29.6	Accepting Realistic Input	426
29.7	Concluding Comments	427

PART VII

*SYNTHESIS AND CODING***CHAPTER 30** *SPEECH SYNTHESIS* **431**

-
- 30.1 Introduction **431**
 - 30.2 Concatenative Methods **433**
 - 30.2.1 Database **433**
 - 30.2.2 Unit selection **434**
 - 30.2.3 Concatenation and optional modification **435**
 - 30.3 Statistical Parametric Methods **436**
 - 30.3.1 Vocoding: from waveforms to features and back **436**
 - 30.3.2 Statistical modeling for speech generation **438**
 - 30.3.3 Advanced techniques **440**
 - 30.4 A Historical Perspective **441**
 - 30.5 Speculation **443**
 - 30.5.1 Physical models **444**
 - 30.5.2 Sub-word units and the role of linguistic knowledge **445**
 - 30.5.3 Prosody matters **445**
 - 30.6 Tools and Evaluation **446**
 - 30.6.1 Further reading **447**
 - 30.7 Exercises **447**
 - 30.8 Appendix: Synthesizer Examples **448**
 - 30.8.1 The Klatt Recordings **448**
 - 30.8.2 Development of Speech Synthesizers **448**
 - 30.8.3 Segmental Synthesis by Rule **449**
 - 30.8.4 Synthesis By Rule of Segments and Sentence Prosody **449**
 - 30.8.5 Fully Automatic Text-To-Speech Conversion: Formants and diphones **450**
 - 30.8.6 The van Santen Recordings **451**
 - 30.8.7 Fully Automatic Text-To-Speech Conversion: Unit selection and HMMs **451**

CHAPTER 31 *PITCH DETECTION* **455**

-
- 31.1 Introduction **455**
 - 31.2 A Note on Nomenclature **455**
 - 31.3 Pitch Detection, Perception and Articulation **456**
 - 31.4 The Voicing Decision **457**
 - 31.5 Some Difficulties in Pitch Detection **458**
 - 31.6 Signal Processing to Improve Pitch Detection **458**
 - 31.7 Pattern-Recognition Methods for Pitch Detection **462**

- 31.8 Smoothing to Fix Errors in Pitch Estimation 467
- 31.9 Normalizing the Autocorrelation Function 469
- 31.10 Exercises 471

CHAPTER 32 *VOCODERS* **473**

- 32.1 Introduction 473
- 32.2 Standards for Digital Speech Coding 473
- 32.3 Design Considerations in Channel Vocoder Filter Banks 473
- 32.4 Energy Measurements in a Channel Vocoder 476
- 32.5 A Vocoder Design for Spectral Envelope Estimation 478
- 32.6 Bit Saving in Channel Vocoders 478
- 32.7 Design of the Excitation Parameters for a Channel Vocoder 482
- 32.8 LPC Vocoders 484
- 32.9 Cepstral Vocoders 484
- 32.10 Design Comparisons 485
- 32.11 Vocoder Standardization 489
- 32.12 Exercises 490

CHAPTER 33 *LOW-RATE VOCODERS* **493**

- 33.1 Introduction 493
- 33.2 The Frame-Fill Concept 494
- 33.3 Pattern Matching or Vector Quantization 496
- 33.4 The Kang–Coulter 600-bps Vocoder 497
- 33.5 Segmentation Methods for Bandwidth Reduction 498
- 33.6 Exercises 503

CHAPTER 34 *MEDIUM-RATE AND HIGH-RATE VOCODERS* **505**

- 34.1 Introduction 505
- 34.2 Voice Excitation and Spectral Flattening 505
- 34.3 Voice-Excited Channel Vocoder 506
- 34.4 Voice-Excited and Error-Signal-Excited LPC Vocoders 508
- 34.5 Waveform Coding with Predictive Methods 510
- 34.6 Adaptive Predictive Coding of Speech 512
- 34.7 Subband Coding 513
- 34.8 Multipulse LPC Vocoders 514
- 34.9 Code-Excited Linear Predictive Coding 516
 - 34.9.1 Basic CELP 516
 - 34.9.2 Modifications to CELP 518
 - 34.9.3 Non-Gaussian Codebook Sequences 518
 - 34.9.4 Low-Delay CELP 519
- 34.10 Reducing Codebook Search Time in CELP 520
 - 34.10.1 Filter Simplification 520

34.10.2	Speeding Up the Search	522
34.10.3	Multiresolution Codebook Search	523
34.10.4	Partial Sequence Elimination	524
34.10.5	Tree-Structured Delta Codebooks	524
34.10.6	Adaptive Codebooks	525
34.10.7	Linear Combination Codebooks	526
34.10.8	Vector Sum Excited Linear Prediction	527
34.11	Conclusions	527
34.12	Exercises	528

CHAPTER 35 *PERCEPTUAL AUDIO CODING* 531

35.1	Transparent Audio Coding	531
35.2	Perceptual Masking	533
35.2.1	Psychoacoustic phenomena	533
35.2.2	Computational models	535
35.3	Noise Shaping	538
35.3.1	Subband analysis	539
35.3.2	Temporal noise shaping	542
35.4	Some Example Coding Schemes	546
35.4.1	MPEG-1 Audio layers I and II	546
35.4.2	MPEG-1 Audio Layer III (MP3)	546
35.4.3	MPEG-2 Advanced Audio Codec (AAC)	547
35.5	Summary	548
35.6	Exercises	549

PART VIII

OTHER APPLICATIONS

CHAPTER 36 *SOME ASPECTS OF COMPUTER MUSIC SYNTHESIS* 553

36.1	Introduction	553
36.2	Some Examples of Acoustically Generated Musical Sounds	553
36.3	Music Synthesis Concepts	555
36.4	Analysis-Based Synthesis	557
36.5	Other Techniques for Music Synthesis	560
36.6	Reverberation	562
36.7	Several Examples of Synthesis	563
36.8	Exercises	565

CHAPTER 37 *MUSIC SIGNAL ANALYSIS* 567

37.1	The Information in Music Audio	567
37.2	Music Transcription	568

37.3	Note Transcription	569
37.4	Score Alignment	571
37.5	Chord Transcription	574
37.6	Structure Detection	576
37.7	Conclusion	577
37.8	Exercises	578

CHAPTER 38 *MUSIC RETRIEVAL* 581

38.1	The Music Retrieval Problem	581
38.2	Music Fingerprinting	582
38.3	Query by Humming	584
38.4	Cover Song Matching	587
38.5	Music Classification and Autotagging	589
38.6	Music Similarity	591
38.7	Conclusions	592
38.8	Exercises	592

CHAPTER 39 *SOURCE SEPARATION* 595

39.1	Sources and Mixtures	595
39.2	Evaluating Source Separation	596
39.3	Multi-Channel Approaches	598
39.4	Beamforming with Microphone Arrays	599
39.4.1	A multi-channel signal model	601
39.4.2	Time-invariant Beamformers	602
39.4.3	Adaptive beamformers	604
39.4.4	Alternative Objective Criteria	605
39.5	Independent Component Analysis	605
39.6	Computational Auditory Scene Analysis	607
39.7	Model-Based Separation	610
39.8	Conclusions	613
39.9	Exercises	614

CHAPTER 40 *SPEECH TRANSFORMATIONS* 617

40.1	Introduction	617
40.2	Time-Scale Modification	617
40.3	Transformation Without Explicit Pitch Detection	620
40.4	Transformations in Analysis–Synthesis Systems	621
40.5	Speech Modifications in the Phase Vocoder	623
40.6	Speech Transformations Without Pitch Extraction	625
40.7	The Sine Transform Coder as a Transformation Algorithm	628
40.8	Voice Modification to Emulate a Target Voice	629
40.9	Exercises	630

CHAPTER 41 *SPEAKER VERIFICATION* **633**

- 41.1 Introduction **633**
- 41.2 General Design of a Speaker Recognition System **634**
- 41.3 Example System Components **635**
 - 41.3.1 Features **635**
 - 41.3.2 Models **635**
 - 41.3.3 Score normalization **637**
 - 41.3.4 Fusion and calibration **638**
- 41.4 Evaluation **638**
- 41.5 Modern Research Challenges **641**
- 41.6 Exercises **641**

CHAPTER 42 *SPEAKER DIARIZATION* **644**

- 42.1 Introduction **644**
- 42.2 General Design of a Speaker Diarization System **645**
- 42.3 Example System Components **647**
 - 42.3.1 Features **647**
 - 42.3.2 Segmentation and clustering **647**
 - 42.3.3 Acoustic beamforming **649**
 - 42.3.4 Speech activity detection **650**
- 42.4 Research Challenges **651**
 - 42.4.1 Overlap resolution **651**
 - 42.4.2 Multimodal diarization **651**
 - 42.4.3 Further challenges **652**
- 42.5 Exercises **652**

PREFACE TO THE 2011 EDITION

0.1 WHY WE CREATED A NEW EDITION

Technology moves at a dizzying pace; however, progress can actually seem quite slow in any area that we are deeply involved in. Conference proceedings are filled with incremental advances over previous methods, and entirely novel (and successful) approaches to speech and audio processing are rare. But a lot can happen in a decade, and it has. In addition to quite new methods, there are also many ideas that had not really been refined enough to show progress in the 1990s, but which now are in common use. For instance, Maximum Mutual Information methods, which were developed for ASR many years ago and were briefly described in the previous edition of this book, was significantly refined in the last decade, and the newer versions of this approach are now widely used. Consequently, we devoted new sections of this revision to MMI (and related methods like MPE).

These advances might have been sufficient to warrant an update of our textbook, but there were other reasons as well. A decade of teaching with the book has revealed a number of bugs and deficiencies, and a new edition affords us the opportunity to correct them. For instance, the previous version had nothing about sound source separation, an area that has received considerable attention in the last decade. Approaches to the coding, transcription, and retrieval of music are also now significant areas of audio signal processing, and were not originally covered in the book.

Last, and not least, the new edition has the benefit of a fresh look at the overall subject from our new co-author, Professor Dan Ellis from Columbia University. This hand-off is a key step in keeping the text current.

As with the previous edition, we've attempted to keep the overall style consistent, focusing on what we think is essential, and leaving many details for other publications. We hope that this choice has helped to make the text useful for many readers.

0.2 WHAT IS NEW

As noted above, we have edited and modified many of the chapters, but we also have added entirely new ones. These are:

- Acoustic model training: further topics – MAP and MLLR adaptation methods, and on MMI and MPE discriminant training (Chapter 28, by new contributor Steven Wegmann of Cisco and ICSI).

- Perceptual Audio Coding – MPEG audio and the related psychoacoustics (Chapter 35, by Dan Ellis).
- Music Signal Analysis – automatic transcription of music (Chapter 37, by Dan Ellis).
- Music Retrieval – music retrieval, including cover song detection (Chapter 38, by Dan Ellis).
- Source Separation – methods to separate different signals, including CASA and multi-microphone methods (Chapter 39, by Dan Ellis, with a section on microphone arrays by Michael Seltzer of Microsoft Research).
- Speaker Diarization – determining who spoke when (Chapter 42, by new contributor Gerald Friedland of ICSI).

Two other chapters have essentially been entirely rewritten: Speech Synthesis (Chapter 30, by Simon King from Edinburgh University), and Speaker Verification (Chapter 41, by David van Leeuwen from TNO). Also, Eric Fosler (of Ohio State University) has extensively revised his chapter on Linguistic Categories for Speech Recognition (Chapter 23).

Many other chapters have also undergone significant revisions; for instance, there are a number of significant updates to the chapters on ASR history (Chapter 4) and on feature extraction for ASR (Chapter 22), and a brief description of the Support Vector Machine (SVM) has been added to the deterministic pattern classification chapter (Chapter 8) in recognition of its increased importance. Finally, the Introduction has been modified to reflect the new distribution of chapters.

0.3 A FINAL THOUGHT

Ben Gold was the key inspiration and co-author for the first edition; there clearly would have been no book without him. He also was an inspiration and role model for me (Morgan) personally. It saddens me that he cannot be here for the new edition, but I know that his generous spirit would have welcomed the new contributions from Dan Ellis and others.

INTRODUCTION

We are confronted with insurmountable opportunities.

—Walt Kelly

1.1 WHY WE WROTE THIS BOOK

Speech and music are the most basic means of adult human communication. As technology advances and increasingly sophisticated tools become available to use with speech and music signals, scientists can study these sounds more effectively and invent new ways of applying them for the benefit of humankind. Such research has led to the development of speech and music synthesizers, speech transmission systems, and automatic speech recognition (ASR) systems. Hand in hand with this progress has come an enhanced understanding of how people produce and perceive speech and music. In fact, the processing of speech and music by devices and the perception of these sounds by humans are areas that inherently interact with and enhance each other.

Despite significant progress in this field, there is still much that is not well understood. Speech and music technology could be greatly improved. For instance, in the presence of unexpected acoustic variability, ASR systems often perform much worse than human listeners (still!). Speech that is synthesized from arbitrary text still sounds artificial. Speech-coding techniques remain far from optimal, and the goal of transparent transmission of speech and music with minimal bandwidth is still distant. All fields associated with the processing and perception of speech and music stand to benefit greatly from continued research efforts. Finally, the growing availability of computer applications incorporating audio (particularly over the Internet and in portable devices) has increased the need for an ever-wider group of engineers and computer scientists to understand audio signal processing. For all of these reasons, as well as our own need to standardize a text for our graduate course at UC Berkeley, we wrote this book; and for the reasons noted in the Preface, we have updated it for the current edition.

The notes on which this book is based proved beneficial to graduate students for close to a decade; during this time, of course, the material evolved, including a problem set for each chapter. The material includes coverage of the physiology and psychoacoustics of hearing as well as the results from research on pitch and speech perception, vocoding methods, and information on many aspects of ASR. To this end, the authors have made use of their own research in these fields, as well as the methods and results of many other

contributors. And as noted in the Preface, this edition includes contributions from new authors as well, in order to broaden the coverage and bring it up to date.

In many chapters, the material is written in a historical framework. In some cases, this is done for motivation's sake; the material is part of the historical record, and we hope that the reader will be interested. In other cases, the historical methods provide a convenient introduction to a topic, since they often are simpler versions of more current approaches. Overall, we have tried to take a long-term perspective on technology developments, which in our view requires incorporating a historical context. The fact that otherwise excellent books on this topic have typically avoided this perspective was one of our major motivations for writing this book.

1.2 HOW TO USE THIS BOOK

This text covers a large number of topics in speech and audio signal processing. While we felt that such a wide range was necessary, we also needed to present a level of detail that is appropriate for a graduate text. Therefore, we have elected to focus on basic material with advanced discussion in selected subtopics. We have assumed that readers have prior experience with core mathematical concepts such as difference equations or probability density functions, but we do not assume that the reader is an expert in their use. Consequently, we will often provide a brief and selected introduction to these concepts to refresh the memories of students who have studied the background material at one time but who have not used it recently. The background topics are selected with a particular focus, namely, to be useful to both students and working professionals in the fields of ASR and speaker recognition, speech bandwidth compression, speech analysis and synthesis, and music analysis and synthesis. Topics from the areas of digital signal processing, pattern recognition, and ear physiology and psychoacoustics are chosen so as to be helpful in understanding the basic approaches for speech and audio applications.

The remainder of this book comprises 41 chapters, grouped into eight sections. Each section or part consists of three to seven chapters that are conceptually linked. Each part begins with a short description of its contents and purpose. These parts are as follows:

- I. Historical Background.** In Chapters 2 through 5 we lay the groundwork for key concepts to be explored later in the book, providing a top-level summary of speech and music processing from a historical perspective. Topics include speech and music analysis, synthesis, and speech recognition.
- II. Mathematical Background.** The basic elements of digital signal processing (Chapters 6 and 7) and pattern recognition (Chapters 8 and 9) comprise the core engineering mathematics needed to understand the application areas described in this book.
- III. Acoustics.** The topics in this section (Chapters 10–13) range from acoustic wave theory to simple models for acoustics in human vocal tracts, tubes, strings, and rooms. All of these aspects of acoustics are significant for an understanding of speech and audio signal processing.

- IV. Auditory Perception.** This section (Chapters 14–18) begins with descriptions of how the outer ear, middle ear, and inner ear work; most of the available information comes from experiments on small mammals, such as cats. Insights into human hearing are derived from experimental psychoacoustics. These fundamentals then lead to the study of human pitch perception as applied to speech and music, as well as to studies of human speech perception and recognition. Some of these topics are further developed in Chapters 34 and 35 in the context of perceptual audio coding.
- V. Speech Features.** Systems for ASR and vocoding have nearly always incorporated filter banks, cepstral analysis, linear predictive coding, or some combination of these basic methods. Each of these approaches has been given a full chapter (19–21).
- VI. Automatic Speech Recognition.** Eight chapters (22–29) are devoted to this study of ASR. Topics range from feature extraction to statistical and deterministic sequence analysis, with coverage of both standard and discriminant training of hidden Markov models (including neural network approaches). A new chapter (Chapter 28) updates the book to include now-standard adaptation techniques, as well as further explanation of discriminant training techniques that are commonly used. Part VI concludes with an overview of a complete ASR system.
- VII. Synthesis and Coding.** Speech synthesis (culminating in text-to-speech systems) is first presented in Chapter 30, a chapter that has largely been rewritten to emphasize concatenative and HMM-based techniques that have become dominant in recent years. Chapter 31 is devoted to pitch detection, which applies to both speech and music devices. Many aspects of vocoding systems are then described in Chapters 32–34, ranging from very-high-quality systems working at relatively high bit rates to extremely low-rate systems. Finally, Chapter 35 provides a description of perceptual audio coding, now used for consumer music systems.
- VIII. Other Applications.** In Chapters 36–42 we present several application areas that were not covered in the bulk of the book. Chapter 36 is a review of major issues in music synthesis. Chapter 37 introduces the transcription of music through several kinds of signal analysis. Chapter 38 is focused on methods for identifying and selecting musical selections. Chapter 39 introduces the topic of source separation, which ultimately could be the critical step in bringing many other applications to a human level of performance, since most desired sounds in the real world exist in the context of other sounds occurring simultaneously. Modifications of the time scale, pitch, and spectral envelope can transform speech and music in ways that are increasingly finding common applications (Chapter 40).
- Chapter 41 is an overview of speaker recognition, with an emphasis on speaker verification. With increasing access to electronic information and expansion of electronic commerce, verification of the identity of a system user is becoming increasingly important. This chapter has largely been rewritten to reflect the significant progress that has occurred in this field since 1999. A related area, speaker diarization, is the topic of the final chapter (42). An area of significant commercial and public interest is the labeling of multiparty conversations (such as a technical meeting) with what is sometimes called a rich transcription, which includes not only the

sequence of words, but also other automatic annotations such as the attribution of which speaker is speaking when; this latter capability is often referred to as speaker diarization.

Readers with sufficient background may choose to focus on the application areas described in Parts V–VIII, as the first four parts primarily give preparatory material. However, in our experience, readers at a graduate or senior undergraduate level in electrical engineering or computer science will benefit from the earlier parts as well. In teaching this course, we have also found the problem sets to be helpful in clarifying understanding, and we suspect that they would have similar value for industrial researchers. Another useful study aid is provided by a collection of audio examples that we have used in our course. These examples have been made freely available via the book's World-Wide Web site which can be found at <http://catalog.wiley.com/>. This Web site may also be augmented over time to include links to errata and addenda for the book.

Other books on a similar topic but with a different emphasis can also be used to complement the material here; in particular, we recommend [9] or [11]; a more recent book with significant detail on current methods is [4].

Additionally, a more complete exposition on the background material introduced in Parts II–IV can be found in such texts as the following:

- [8] for digital signal processing
- [1] or [3] for pattern recognition (note that this is the revised edition of the classic [2])
- [6] for acoustics
- [10] for auditory physiology
- [7] for psychoacoustics

Finally, an excellent book already in its second edition is [5], which focuses much more on the language-related aspects of speech processing.

1.3 A CONFESSION

The authors have chosen to spend much of their lives studying speech and audio signals and systems. Although we would like to say that we have done this to benefit society, much of the reason for our vocational path is a combination of happenstance and hedonism; in other words, dumb luck and a desire to have fun. We have enjoyed ourselves in this work, and we continue to do so. Speech and audio processing has become a fulfilling obsession for us, and we hope that some of our readers will adopt and enjoy this obsession too.

1.4 ACKNOWLEDGMENTS

Many other people contributed to this book. Students in our graduate class at Berkeley contributed greatly over the years, both by their need for a text and by their original scribe notes that inspired us to write the book. Two (former) students in particular, Eric Fosler-Lussier and Jeff Gilbert, ultimately wrote material that was the basis of Chapters 22 and 33, respectively. Hervé Bourlard came through very quickly with the original core of the Speaker Verification chapter when we realized that we had no material on speaker identification or verification, and David van Leeuwen provided the updates for the current version. Simon King wrote a new chapter on Speech Synthesis, and Alan Black provided useful comments and criticism. Steven Wegmann wrote new material on adaptation and discriminant training. John Lazzaro provided useful comments on the new perceptual audio and music chapters, and Mike Seltzer added important material on microphone arrays for the source separation chapter. Finally Martin Cooke helped with his remarks on our draft of the CASA description.

For the original version, Su-Lin Wu was an extremely helpful critic, both on the speech recognition sections and on some of the psychoacoustics material. Anita Bounds-Morgan provided very useful editorial assistance. The anonymous reviewers that our publisher used at that time were also quite helpful.

We certainly appreciate the indulgence of the International Computer Science Institute and Lincoln Laboratory for permitting us to develop the original manuscript on the equipment of these two labs, and for Columbia University for similarly providing the necessary resources for Dan Ellis's extensive efforts to update the current version. Devra Polack and Elizabeth Weinstein of ICSI also provided a range of secretarial and artistic support that was very important to the success of the earlier project. We also thank Bill Zobrist of Wiley for his interest in the original book, and George Telecki for his support for our developing the second edition.

Finally, we extend our special thanks to our wives, Sylvia, Anita, and Sarah, for putting up with our intrusion on family time as a result of all the days and evenings that were spent on this book.

BIBLIOGRAPHY

1. Bishop, C., *Neural Networks for Pattern Recognition*, Oxford Univ. Press, London/New York, 1996.
2. Duda, D., and Hart, P., *Pattern Classification and Scene Analysis*, Wiley-Interscience, New York, 1973.
3. Duda, D., Hart, P., and Stork, D., *Pattern Classification (2nd Ed.)*, Wiley-Interscience, New York, 2001.
4. Huang, X., Acero, A., and Hon, H.-W., *Spoken Language Processing*, Prentice-Hall, Englewood Cliffs, N.J., 2001.
5. Jurafsky, D., and Martin, J., *Speech and Language Processing*, Prentice-Hall, Upper Saddle River, N.J., 2009.

6. Kinsler, L., and Frey, A., *Fundamentals of Acoustics*, Wiley, New York, 1962.
7. Moore, B. C. J., *An Introduction to the Psychology of Hearing*, 5th ed. Academic Press, New York/London, 2003.
8. Oppenheim, A., and Schaffer, R., *Discrete-Time Signal Processing (3rd Ed.*, Prentice–Hall, Englewood Cliffs, N.J., 2009.
9. O’Shaughnessy, D., *Speech Communication*, Addison–Wesley, Reading, Mass., 1987.
10. Pickles, J., *An Introduction to the Physiology of Hearing*, Academic Press, New York, 1982.
11. Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice–Hall, Englewood Cliffs, N.J., 1993.