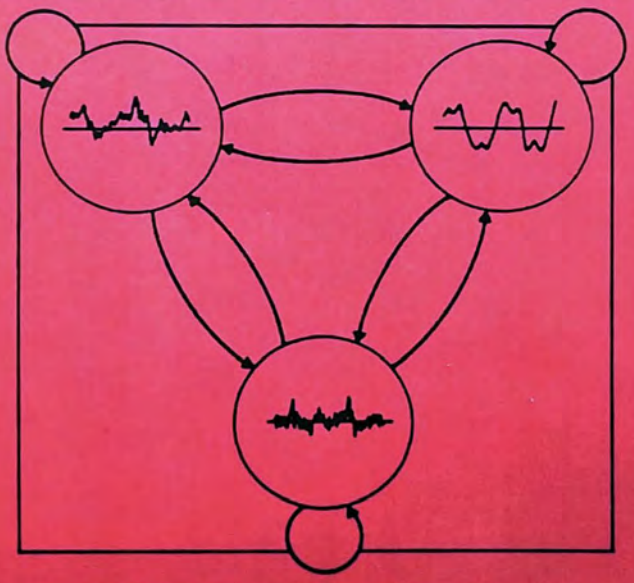
FUNDAMENTALS OF SPEECH RECOGNITION



LAWRENCE RABINER
BIING-HWANG JUANG

PRENTICE HALL SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR

Ex.1022 / Page 1 of 333
AMAZON.COM, INC.

FUNDAMENTALS OF SPEECH RECOGNITION

Lawrence Rabiner
Biing-Hwang Juang

Pearson Education

Prentice Hall P T R
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Rabiner, Lawrence R., 1943–
Fundamentals of speech recognition / Lawrence Rabiner, Biing-Hwang Juang.
p. cm.
Includes bibliographical references and index.
ISBN 0–13–015157–2
1. Automatic speech recognition. 2. Speech processing systems.
I. Juang, B. H. (Biing-Hwang) II. Title.
TK7895.S65R33 1993
006.4'54—dc20 92–34093
CIP

Editorial production and interior design: *bookworks*
Cover Designer: *Ben Santora*
Acquisitions Editor: *Karen Gettman*
Manufacturing Buyer: *Mary Elizabeth McCartney*

©1993 by AT&T. All rights reserved.

© **Published by Prentice Hall PTR**
Prentice-Hall, Inc.
A Pearson Education Company
Upper Saddle River, NJ 07458

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact:

Corporate Sales Department
PTR Prentice Hall

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 9 8

ISBN 0-13-015157-2

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

CONTENTS

LIST OF FIGURES xiii

LIST OF TABLES xxix

PREFACE xxxi

1 FUNDAMENTALS OF SPEECH RECOGNITION 1

1.1 Introduction 1
1.2 The Paradigm for Speech Recognition 3
1.3 Outline 3
1.4 A Brief History of Speech-Recognition Research 6

2 THE SPEECH SIGNAL: PRODUCTION, PERCEPTION, AND ACOUSTIC-PHONETIC CHARACTERIZATION 11

2.1 Introduction 11
2.1.1 The Process of Speech Production and Perception in Human Beings 11
2.2 The Speech-Production Process 14
2.3 Representing Speech in the Time and Frequency Domains 17
2.4 Speech Sounds and Features 20

Ex.1022 / Page 4 of 333
AMAZON.COM, INC.

2.4.1 The Vowels 21
2.4.2 Diphthongs 28
2.4.3 Semivowels 29
2.4.4 Nasal Consonants 30
2.4.5 Unvoiced Fricatives 31
2.4.6 Voiced Fricatives 32
2.4.7 Voiced and Unvoiced Stops 33
2.4.8 Review Exercises 37
2.5 Approaches to Automatic Speech Recognition by Machine 42
2.5.1 Acoustic-Phonetic Approach to Speech Recognition 45
2.5.2 Statistical Pattern-Recognition Approach to Speech Recognition 51
2.5.3 Artificial Intelligence (AI) Approaches to Speech Recognition 52
2.5.4 Neural Networks and Their Application to Speech Recognition 54
2.6 Summary 65

3 SIGNAL PROCESSING AND ANALYSIS METHODS FOR SPEECH RECOGNITION 69

3.1 Introduction 69
3.1.1 Spectral Analysis Models 70
3.2 The Bank-of-Filters Front-End Processor 73
3.2.1 Types of Filter Bank Used for Speech Recognition 77
3.2.2 Implementations of Filter Banks 80
3.2.3 Summary of Considerations for Speech-Recognition Filter Banks 92
3.2.4 Practical Examples of Speech-Recognition Filter Banks 93
3.2.5 Generalizations of Filter-Bank Analyzer 95
3.3 Linear Predictive Coding Model for Speech Recognition 97
3.3.1 The LPC Model 100
3.3.2 LPC Analysis Equations 101
3.3.3 The Autocorrelation Method 103
3.3.4 The Covariance Method 106
3.3.5 Review Exercise 107
3.3.6 Examples of LPC Analysis 108
3.3.7 LPC Processor for Speech Recognition 112
3.3.8 Review Exercises 117
3.3.9 Typical LPC Analysis Parameters 121
3.4 Vector Quantization 122
3.4.1 Elements of a Vector Quantization Implementation 123
3.4.2 The VQ Training Set 124
3.4.3 The Similarity or Distance Measure 125
3.4.4 Clustering the Training Vectors 125
3.4.5 Vector Classification Procedure 128
3.4.6 Comparison of Vector and Scalar Quantizers 129

3.4.7 Extensions of Vector Quantization 129
3.4.8 Summary of the VQ Method 131
3.5 Auditory-Based Spectral Analysis Models 132
3.5.1 The EIH Model 134
3.6 Summary 139

4 PATTERN-COMPARISON TECHNIQUES 141

4.1 Introduction 141
4.2 Speech (Endpoint) Detection 143
4.3 Distortion Measures—Mathematical Considerations 149
4.4 Distortion Measures—Perceptual Considerations 150
4.5 Spectral-Distortion Measures 154
4.5.1 Log Spectral Distance 158
4.5.2 Cepstral Distances 163
4.5.3 Weighted Cepstral Distances and Liftering 166
4.5.4 Likelihood Distortions 171
4.5.5 Variations of Likelihood Distortions 177
4.5.6 Spectral Distortion Using a Warped Frequency Scale 183
4.5.7 Alternative Spectral Representations and Distortion Measures 190
4.5.8 Summary of Distortion Measures—Computational Considerations 193
4.6 Incorporation of Spectral Dynamic Features into the Distortion Measure 194
4.7 Time Alignment and Normalization 200
4.7.1 Dynamic Programming—Basic Considerations 204
4.7.2 Time-Normalization Constraints 208
4.7.3 Dynamic Time-Warping Solution 221
4.7.4 Other Considerations in Dynamic Time Warping 229
4.7.5 Multiple Time-Alignment Paths 232
4.8 Summary 238

5 SPEECH RECOGNITION SYSTEM DESIGN AND IMPLEMENTATION ISSUES 242

5.1 Introduction 242
5.2 Application of Source-Coding Techniques to Recognition 244
5.2.1 Vector Quantization and Pattern Comparison Without Time Alignment 244
5.2.2 Centroid Computation for VQ Codebook Design 246
5.2.3 Vector Quantizers with Memory 254
5.2.4 Segmental Vector Quantization 256
5.2.5 Use of a Vector Quantizer as a Recognition Preprocessor 257
5.2.6 Vector Quantization for Efficient Pattern Matching 263
5.3 Template Training Methods 264
5.3.1 Casual Training 265

5.3.2 Robust Training 266
5.3.3 Clustering 267
5.4 Performance Analysis and Recognition Enhancements 274
5.4.1 Choice of Distortion Measures 274
5.4.2 Choice of Clustering Methods and kNN Decision Rule 277
5.4.3 Incorporation of Energy Information 280
5.4.4 Effects of Signal Analysis Parameters 282
5.4.5 Performance of Isolated Word-Recognition Systems 284
5.5 Template Adaptation to New Talkers 285
5.5.1 Spectral Transformation 286
5.5.2 Hierarchical Spectral Clustering 288
5.6 Discriminative Methods in Speech Recognition 291
5.6.1 Determination of Word Equivalence Classes 294
5.6.2 Discriminative Weighting Functions 297
5.6.3 Discriminative Training for Minimum Recognition Error 302
5.7 Speech Recognition in Adverse Environments 305
5.7.1 Adverse Conditions in Speech Recognition 306
5.7.2 Dealing with Adverse Conditions 309
5.8 Summary 317

6 THEORY AND IMPLEMENTATION OF HIDDEN MARKOV MODELS 321

6.1 Introduction 321
6.2 Discrete-Time Markov Processes 322
6.3 Extensions to Hidden Markov Models 325
6.3.1 Coin-Toss Models 326
6.3.2 The Urn-and-Ball Model 328
6.3.3 Elements of an HMM 329
6.3.4 HMM Generator of Observations 330
6.4 The Three Basic Problems for HMMs 333
6.4.1 Solution to Problem 1—Probability Evaluation 334
6.4.2 Solution to Problem 2—"Optimal" State Sequence 337
6.4.3 Solution to Problem 3—Parameter Estimation 342
6.4.4 Notes on the Reestimation Procedure 347
6.5 Types of HMMs 348
6.6 Continuous Observation Densities in HMMs 350
6.7 Autoregressive HMMs 352
6.8 Variants on HMM Structures—Null Transitions and Tied States 356
6.9 Inclusion of Explicit State Duration Density in HMMs 358
6.10 Optimization Criterion—ML, MMI, and MDI 362
6.11 Comparisons of HMMs 364
6.12 Implementation Issues for HMMs 365
6.12.1 Scaling 365
6.12.2 Multiple Observation Sequences 369
6.12.3 Initial Estimates of HMM Parameters 370

6.12.4 Effects of Insufficient Training Data 370
6.12.5 Choice of Model 371
6.13 Improving the Effectiveness of Model Estimates 372
6.13.1 Deleted Interpolation 372
6.13.2 Bayesian Adaptation 373
6.13.3 Corrective Training 376
6.14 Model Clustering and Splitting 377
6.15 HMM System for Isolated Word Recognition 378
6.15.1 Choice of Model Parameters 379
6.15.2 Segmental K-Means Segmentation into States 382
6.15.3 Incorporation of State Duration into the HMM 384
6.15.4 HMM Isolated-Digit Performance 385
6.16 Summary 386

7 SPEECH RECOGNITION BASED ON CONNECTED WORD MODELS 390

7.1 Introduction 390
7.2 General Notation for the Connected Word-Recognition Problem 393
7.3 The Two-Level Dynamic Programming (Two-Level DP) Algorithm 395
7.3.1 Computation of the Two-Level DP Algorithm 399
7.4 The Level Building (LB) Algorithm 400
7.4.1 Mathematics of the Level Building Algorithm 401
7.4.2 Multiple Level Considerations 405
7.4.3 Computation of the Level Building Algorithm 407
7.4.4 Implementation Aspects of Level Building 410
7.4.5 Integration of a Grammar Network 414
7.4.6 Examples of LB Computation of Digit Strings 416
7.5 The One-Pass (One-State) Algorithm 416
7.6 Multiple Candidate Strings 420
7.7 Summary of Connected Word Recognition Algorithms 423
7.8 Grammar Networks for Connected Digit Recognition 425
7.9 Segmental K-Means Training Procedure 427
7.10 Connected Digit Recognition Implementation 428
7.10.1 HMM-Based System for Connected Digit Recognition 429
7.10.2 Performance Evaluation on Connected Digit Strings 430
7.11 Summary 432

8 LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION 434

8.1 Introduction 434
8.2 Subword Speech Units 435
8.3 Subword Unit Models Based on HMMs 439
8.4 Training of Subword Units 441

8.5 Language Models for Large Vocabulary Speech Recognition 447
8.6 Statistical Language Modeling 448
8.7 Perplexity of the Language Model 449
8.8 Overall Recognition System Based on Subword Units 450
8.8.1 Control of Word Insertion/Word Deletion Rate 454
8.8.2 Task Semantics 454
8.8.3 System Performance on the Resource Management Task 454
8.9 Context-Dependent Subword Units 458
8.9.1 Creation of Context-Dependent Diphones and Triphones 460
8.9.2 Using Interword Training to Create CD Units 461
8.9.3 Smoothing and Interpolation of CD PLU Models 462
8.9.4 Smoothing and Interpolation of Continuous Densities 464
8.9.5 Implementation Issues Using CD Units 464
8.9.6 Recognition Results Using CD Units 467
8.9.7 Position Dependent Units 469
8.9.8 Unit Splitting and Clustering 470
8.9.9 Other Factors for Creating Additional Subword Units 475
8.9.10 Acoustic Segment Units 476
8.10 Creation of Vocabulary-Independent Units 477
8.11 Semantic Postprocessor for Recognition 478
8.12 Summary 478

9 TASK ORIENTED APPLICATIONS OF AUTOMATIC SPEECH RECOGNITION 482

9.1 Introduction 482
9.2 Speech-Recognizer Performance Scores 484
9.3 Characteristics of Speech-Recognition Applications 485
9.3.1 Methods of Handling Recognition Errors 486
9.4 Broad Classes of Speech-Recognition Applications 487
9.5 Command-and-Control Applications 488
9.5.1 Voice Repertory Dialer 489
9.5.2 Automated Call-Type Recognition 490
9.5.3 Call Distribution by Voice Commands 491
9.5.4 Directory Listing Retrieval 491
9.5.5 Credit Card Sales Validation 492
9.6 Projections for Speech Recognition 493

INDEX 497

LIST OF FIGURES

1.1 General block diagram of a task-oriented speech-recognition system. 3

2.1 Schematic diagram of speech-production/speech-perception process (after Flanagan [unpublished]). 12

2.2 Alternative view of speech-production/speech-perception process (after Rabiner and Levinson [1]). 13

2.3 Mid-sagittal plane X-ray of the human vocal apparatus (after Flanagan et al. [2]). 15

2.4 Schematic view of the human vocal mechanism (after Flanagan [3]). 16

2.5 Glottal volume velocity and resulting sound pressure at the start of a voiced sound (after Ishizaka and Flanagan [4]). 16

2.6 Schematic representation of the complete physiological mechanism of speech production (after Flanagan [3]). 17

2.7 Waveform plot of the beginning of the utterance "It's time." 18

2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea." 19

Ex.1022 / Page 10 of 333
AMAZON.COM, INC.

2.9 Wideband spectrogram and formant frequency representation of the utterance "Why do I owe you a letter" (after Atal and Hanauer [5]). 21
2.10 Wideband spectrogram and intensity contour of the phrase "Should we chase." 22
2.11 The speech waveform and a segmentation and labeling of the constituent sounds of the phrase "Should we chase." 23
2.12 Chart of the classification of the standard phonemes of American English into broad sound classes. 25
2.13 Articulatory configurations for typical vowel sounds (after Flanagan [3]). 25
2.14 Acoustic waveform plots of typical vowel sounds. 26
2.15 Spectrograms of the vowel sounds. 27
2.16 Measured frequencies of first and second formants for a wide range of talkers for several vowels (after Peterson & Barney [7]). 27
2.17 The vowel triangle with centroid positions of the common vowels. 28
2.18 Spectrogram plots of four diphthongs. 30
2.19 Time variation of the first two formants for the diphthongs (after Holbrook and Fairbanks [9]). 31
2.20 Waveforms for the sequences /ə-m-a/ and /ə-n-a/. 32
2.21 Spectrograms of the sequences /ə-m-a/ and ə-n-a/. 33
2.22 Waveforms for the sounds /f/, /s/ and /sh/ in the context /ə-x-a/ where /x/ is the unvoiced fricative. 34
2.23 Spectrogram comparisons of the sounds /ə-f-a/, /ə-s-a/ and /ə-sh-a/. 34
2.24 Waveforms for the sequences /ə-v-a/ and /ə-zh-a/. 35
2.25 Spectrograms for the sequences /ə-v-a/ and /ə-zh-a/. 36
2.26 Waveform for the sequence /ə-b-a/. 36
2.27 Waveforms for the sequences /ə-p-a/ and /ə-t-a/. 37
2.28 Spectrogram comparisons of the sequences of voiced (/ə-b-a/) and voiceless (/ə-p-a/ and /ə-t-a/) stop consonants. 38
2.29 Spectrograms of the 11 isolated digits, 0 through 9 plus oh, in random sequence. 40
2.30 Spectrograms of two connected digit sequences. 41
2.31 Phoneme lattice for word string. 43
2.32 Block diagram of acoustic-phonetic speech-recognition system. 45
2.33 Acoustic-phonetic vowel classifier. 47
2.34 Binary tree speech sound classifier. 48
2.35 Segmentation and labeling for word sequence "seven–six." 49
2.36 Segmentation and labeling for word sequence "did you." 50
2.37 Block diagram of pattern-recognition speech recognizer. 51

Ex.1022 / Page 11 of 333
AMAZON.COM, INC.

2.38 Illustration of the word correction capability of syntax in speech recognition (after Rabiner and Levinson [1]). 54
2.39 A bottom-up approach to knowledge integration for speech recognition. 55
2.40 A top-down approach to knowledge integration for speech recognition. 55
2.41 A blackboard approach to knowledge integration for speech recognition (after Lesser et al. [11]). 56
2.42 Conceptual block diagram of a human speech understanding system. 56
2.43 Simple computation element of a neural network. 57
2.44 McCullough-Pitts model of neurons (after McCullough and Pitts [12]). 58
2.45 Single-layer and three-layer perceptrons. 59
2.46 A multilayer perceptron for classifying steady vowels based on F_1, F_2 measurements (after Lippmann [13]). 59
2.47 Model of a recurrent neural network. 60
2.48 A fixed point interpretation of the Hopfield network. 60
2.49 The time delay neural network computational element (after Waibel et al. [14]). 63
2.50 A TDNN architecture for recognizing /b/, /d/ and /g/ (after Waibel et al. [14]). 64
2.51 A combination neural network and matched filter for speech recognition (after Tank & Hopfield [15]). 65
2.52 Example illustrating the combination of a neural network and a set of matched filters (after Tank & Hopfield [15]). 66
2.53 The hidden control neural network (after Levin [16]). 67

3.1 (a) Pattern recognition and (b) acoustic phonetic approaches to speech recognition. 71
3.2 Bank-of-filters analysis model. 72
3.3 LPC analysis model. 72
3.4 Complete bank-of-filters analysis model. 74
3.5 Typical waveforms and spectra for analysis of a pure sinusoid in the filter-bank model. 75
3.6 Typical waveforms and spectra of a voice speech signal in the bank-of-filters analysis model. 76
3.7 Ideal (a) and realistic (b) set of filter responses of a Q-channel filter bank covering the frequency range F_s/N to $(Q + {}^1/_2)F_s/N$. 77
3.8 Ideal specifications of a 4-channel octave band-filter bank (a), a 12-channel third-octave band filter bank (b), and a 7-channel critical band scale filter bank (c) covering the telephone bandwidth range (200–3200 Hz). 79

Ex.1022 / Page 12 of 333
AMAZON.COM, INC.

3.9 The variation of bandwidth with frequency for the perceptually based critical band scale. 79
3.10 The signals $s(m)$ and $w(n-m)$ used in evaluation of the short-time Fourier transform. 81
3.11 Short-time Fourier transform using a long (500 points or 50 msec) Hamming window on a section of voiced speech. 82
3.12 Short-time Fourier transform using a short (50 points or 5 msec) Hamming window on a section of voiced speech. 82
3.13 Short-time Fourier transform using a long (500 points or 50 msec) Hamming window on a section of unvoiced speech. 83
3.14 Short-time Fourier transform using a short (50 points or 5 msec) Hamming window on a section of unvoiced speech. 83
3.15 Linear filter interpretation of the short-time Fourier transform. 84
3.16 FFT implementation of a uniform filter bank. 89
3.17 Direct form implementation of an arbitrary nonuniform filter bank. 89
3.18 Two arbitrary nonuniform filter-bank ideal filter specifications consisting of either 3 bands (part a) or 7 bands (part b). 90
3.19 Tree structure implementation of a 4-band, octave-spaced, filter bank. 92
3.20 Window sequence, $w(n)$, (part a), the individual filter response (part b), and the composite response (part c) of a $Q = 15$ channel, uniform filter bank, designed using a 101-point Kaiser window smoothed lowpass window (after Dautrich et al. [4]). 94
3.21 Window sequence, $w(n)$, (part a), the individual filter responses (part b), and the composite response (part c) of a $Q = 15$ channel, uniform filter bank, designed using a 101-point Kaiser window directly as the lowpass window (after Dautrich et al. [4]). 95
3.22 Individual channel responses (parts a to d) and composite filter response (part c) of a $Q = 4$ channel, octave band design, using 101-point FIR filters in each band (after Dautrich et al. [4]). 96
3.23 Individual channel responses and composite filter response of a $Q = 12$ channel, $1/3$ octave band design, using 201-point FIR filters in each band (after Dautrich et al. [4]). 97
3.24 Individual channel responses (parts a to g) and composite filter response (part h) of a $Q = 7$ channel critical band filter bank design (after Dautrich et al. [4]). 98
3.25 Individual channel responses and composite filter response of a $Q = 13$ channel, critical band spacing filter bank, using highly overlapping filters in frequency (after Dautrich et al. [4]). 99
3.26 Generalization of filter-bank analysis model. 99
3.27 Linear prediction model of speech. 100
3.28 Speech synthesis model based on LPC model. 101

Ex.1022 / Page 13 of 333
AMAZON.COM, INC.

3.29 Illustration of speech sample, weighted speech section, and prediction error for voiced speech where the prediction error is large at the beginning of the section. 104
3.30 Illustration of speech sample, weighted speech section, and prediction error for voiced speech where the prediction error is large at the end of the section. 104
3.31 Illustration of speech sample, weighted speech section, and prediction error for unvoiced speech where there are almost no artifacts at the boundaries of the section. 105
3.32 Typical signals and spectra for LPC autocorrelation method for a segment of speech spoken by a male speaker (after Rabiner et al. [8]). 108
3.33 Typical signals and spectra for LPC autocorrelation method for a segment of speech spoken by a female speaker (after Rabiner et al. [8]). 109
3.34 Examples of signal (differentiated) and prediction error for several vowels (after Strube [9]). 110
3.35 Variation of the RMS prediction error with the number of predictor coefficients, p (after Atal and Hanauer [10]). 110
3.36 Spectra for a vowel sound for several values of predictor order, p. 111
3.37 Block diagram of LPC processor for speech recognition. 113
3.38 Magnitude spectrum of LPC preemphasis network for $\tilde{a} = 0.95$. 113
3.39 Blocking of speech into overlapping frames. 114
3.40 Block diagram of the basic VQ training and classification structure. 124
3.41 Partitioning of a vector space into VQ cells with each cell represented by a centroid vector. 126
3.42 Flow diagram of binary split codebook generation algorithm. 127
3.43 Codebook distortion versus codebook size (measured in bits per frame) for both voiced and unvoiced speech (after Juang et al. [12]). 128
3.44 Codebook vector locations in the $F_1 - F_2$ plane (for a 32-vector codebook) superimposed on the vowel ellipses (after Juang et al. [12]). 128
3.45 Model and distortion error spectra for scalar and vector quantizers (after Juang et al. [12]). 130
3.46 Plots and histograms of temporal distortion for scalar and vector quantizers (after Juang et al. [12]). 131
3.47 Physiological model of the human ear. 132
3.48 Expanded view of the middle and inner ear mechanics. 133
3.49 Block diagram of the EIH model (after Ghitza [13]). 135

3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]). 136
3.51 Magnitude of EIH for vowel /o/ showing the time-frequency resolution (after Ghitza [13]). 136
3.52 Operation of the EIH model for a pure sinusoid (after Ghitza [13]). 137
3.53 Comparison of Fourier and EIH log spectra for clean and noisy speech signals (after Ghitza [13]). 138

4.1 Contour of digit recognition accuracy (percent correct) as a function of endpoint perturbation (in ms) in a multispeaker digit-recognition experiment. Both the initial (beginning point) and the final (ending point) boundary of the detected speech signal were varied (after Wilpon et al. [2]). 144
4.2 Example of mouth click preceding a spoken word (after Wilpon et al. [2]). 145
4.3 Example of breathy speech due to heavy breathing while speaking (after Wilpon et al. [2]). 146
4.4 Example of click produced at the end of a spoken word (after Wilpon et al. [2]). 147
4.5 Block diagram of the explicit approach to speech endpoint detection. 147
4.6 Block diagram of the implicit approach to speech-endpoint detection. 148
4.7 Examples of word boundaries as determined by the implicit endpoint detection algorithm. 148
4.8 Block diagram of the hybrid approach to speech endpoint detection. 149
4.9 Block diagram of typical speech activity detection algorithm. 149
4.10 LPC pole frequency JNDs as a function of the pole bandwidth; the blank circles denote positive frequency perturbations, and the solid circles represent negative frequency perturbations; the fitting curves are parabolic (after Erell et al. [7]). 153
4.11 LPC pole bandwidth JNDs, in a logarithmic scale, as a function of the pole bandwidth itself (after Erell et al. [7]). 154
4.12 Two typical FFT power spectra, $S(\omega)$, of the sound /æ/ in a log scale and their difference magnitude $|V(\omega)|$ as a function of frequency. 159
4.13 LPC model spectra corresponding to the FFT spectra in Figure 4.12, plotted also in a log scale, and their difference magnitude $|V(\omega)|$ as a function of frequency. 159
4.14 Two typical FFT power spectra, $S(\omega)$, of the sound /sh/ in a log scale and their difference magnitude $|V(\omega)|$ as a function of frequency. 160

Ex.1022 / Page 15 of 333
AMAZON.COM, INC.

4.15 LPC model spectra corresponding to the FFT spectra in Figure 4.14, plotted also in a log scale, and their difference magnitude $|V(\omega)|$ as a function of frequency. 160

4.16 Typical FFT power spectra of the sounds /æ/ and /i/ respectively and their difference magnitude as a function of frequency. 161

4.17 LPC model spectra corresponding to the FFT spectra in Figure 4.16 and their difference magnitude $|V(\omega)|$ as a function of frequency. 161

4.18 Scatter plot of d_2^2, the cepstral distance, versus $2d_c^2(L)$, the truncated cepstral distance (multiplied by 2), for 800 pairs of all-pole model spectra; the truncation is at $L = 20$ (after Gray and Markel [9]). 165

4.19 Scatter plot of d_2^2, the cepstral distance, versus $2d_c^2(L)$, the truncated cepstral distance (multiplied by 2), for 800 pairs of all-pole model spectra; the truncation is at $L = 30$ (after Gray and Markel [9]). 165

4.20 Effects of cepstral liftering on a log LPC spectrum, as a function of the lifter length ($L = 8$ to 16) (after Juang et al. [11]). 169

4.21 Comparison of (a) original sequence of LPC log magnitude spectra; (b) liftered LPC log magnitude spectra, and (c) liftered log magnitude spectra (after Juang et al. [11]). 170

4.22 Comparison of the distortion integrands $V^2(\omega)/2$ and $e^{V(\omega)} - V(\omega) - 1$ (after Gray and Markel [9]). 173

4.23 A scatter plot of $d_{IS}(1/|A_p|^2, 1/|A|^2) + 1$ versus $d_{IS}(1/|A|^2, 1/|A_p|^2) + 1$ as measured from 6800 pairs of speech model spectra. 176

4.24 A linear system with transfer function $H(z) = A(z)/B(z)$. 177

4.25 A scatter diagram of the log spectral distance versus the COSH distortion as measured from a database of 6800 pairs of speech spectra. 178

4.26 LPC spectral pair and various spectral weighting functions; $W_1(\omega)$, $W_2(\omega)$, $W_3(\omega)$ and $W_4(\omega)$ are defined in (4.71), (4.72), (4.73), and (4.74), respectively. 180

4.27a An example of the cosh spectral deviation $F_4(\omega)$ and its weighted version using $W_3(\omega) = 1/|A(e^{j\omega})|^2$ as the weighting function; in this case the two spectra are of comparable power levels. 182

4.27b An example of the cosh spectral deviation $F_4(\omega)$ and its weighted version using $W_3(\omega) = 1/|A(e^{j\omega})|^2$ as the weighting function; in this case, the two spectra have significantly different power levels. 182

4.28 Subjectively perceived pitch, in mels, of a tone as a function of the frequency, in Hz; the upper curve relates the subjective pitch to frequency in a linear scale and the lower curve shows the subjective pitch as a function of the frequency in a logarithmic scale (after Stevens and Volkmann [13]). 184
4.29 The critical bandwidth phenomenon; the critical bandwidth as a function of the frequency at the center of the band (after Zwicker, Flottorp and Stevens [14]). 185
4.30 Real part of exp $[j\theta(b)k]$ as a function of b, the Bark scale, for different values of k (after Nocerino et al. [16]). 188
4.31 A filter-bank design in which each filter has a triangle bandpass frequency response with bandwidth and spacing determined by a constant mel frequency interval (spacing = 150 mels, bandwidth = 300 mels) (after Davis and Mermelstein [17]). 190
4.32 (a) Series of cylindrical sections concatenated as an acoustic tube model of the vocal tract; (b) the area function of the cylindrical sections in (a) (after Markel and Gray [10]). 191
4.33 A critical band spectrum (a) of a typical vowel sound and the corresponding log spectral differences V_{LM} (b) and V_{GM} (c) as functions of the critical band number (after Nocerino et al. [16]). 193
4.34 A trajectory of the (2nd) cepstral coefficient with 2nd-order polynomial $(h_1 + h_2 t + h_3 t^2)$ fitting on short portions of the trajectory; the width for polynomial fitting is 7 points. 197
4.35 Scatter diagram showing the correlation between the "instantaneous" cepstral distance, d_2, and the "differential" or "dynamic" cepstral distance, $d_{2\delta(1)}$; the correlation index is 0.6. 199
4.36 Linear time alignment for two sequences of different durations. 202
4.37 An example of time normalization of two sequential patterns to a common time index; the time warping functions ϕ_x and ϕ_y map the individual time index i_x and i_y, respectively, to the common time index k. 203
4.38 The optimal path problem—finding the minimum cost path from point 1 to point i in as many moves as needed. 205
4.39 A trellis structure that illustrates the problem of finding the optimal path from point i to point j in M steps. 207
4.40 An example of local continuity constraints expressed in terms of coordinate increments (after Myers et al. [23]). 210
4.41 The effects of global path constraints and range limiting on the allowable regions for time warping functions. 215
4.42 Illustration of the extreme cases, $T_x - 1 = Q_{\max}(T_y - 1)$ or $T_y - 1 = Q_{\max}(T_x - 1)$, where only linear time warping (single straight path) is allowed. 216

Ex.1022 / Page 17 of 333
AMAZON.COM, INC.

4.43 Type III local continuity constraints with four types of slope weighting (after Myers et al. [23]). 217
4.44 Type II local continuity constraints with 4 types of slope weighting and their smoothed version in which the slope weights are uniformly redistributed along paths where abrupt weight changes exist (after Myers et al. [23]). 218
4.45 Set of allowable grid points for dynamic programming implementation of local path expansion and contraction by 2 to 1. 225
4.46 The allowable path region for dynamic time alignment with relaxed endpoint constraints. 230
4.47 Set of allowable grid points when opening up the initial point range to 5 frames and the final point range to 9 frames. 231
4.48 The allowable path region for dynamic time alignment with localized range constraints. 232
4.49 Dynamic programming for finding K-best paths implemented in a parallel manner. 234
4.50 The serial dynamic programming algorithm for finding the K-best paths. 236
4.51 Example illustrating the need for nonlinear time alignment of two versions of a spoken word. 239
4.52 Illustration of the effectiveness of dynamic time warping alignment of two versions of a spoken word. 240

5.1 A vector-quantizer–based speech-recognition system. 246
5.2 A trellis quantizer as a finite state machine. 255
5.3 Codebook training for segmental vector quantization. 256
5.4 Block diagram of isolated word recognizer incorporating a word-based VQ preprocessor and a DTW-based postprocessor (after Pan et al. [4]). 258
5.5 Plots of the variation of preprocessor performance parameters P_1, E_1, and $(1 - \gamma)$ as a function of the distortion threshold D' for several codebook sizes for the digits vocabulary (after Pan et al. [4]). 260
5.6 Plots of the variation of preprocessor performance parameters E_2 and β as a function of the distortion threshold D' for several codebook sizes for the digits vocabulary (after Pan et al. [4]). 260
5.7 Plots of average fraction of decisions made by the preprocessor, γ, versus preprocessor decision threshold D'' for several codebook sizes for the digits vocabulary (after Pan et al. [4]). 262
5.8 Plots of average fraction of candidate words, β, passed on to the postprocessor, versus preprocessor decision threshold D'' for several codebook sizes for the digits vocabulary (after Pan et al. [4]). 262

5.9 Accumulated DTW distortion scores versus test frame based on casual training with two reference patterns per word (after Rabiner et al. [5]). 265
5.10 A flow diagram of the UWA clustering procedure (after Wilpon and Rabiner [7]). 269
5.11 A flow diagram of the MKM clustering procedure (after Wilpon and Rabiner [7]). 272
5.12 Recognition accuracy (percent correct) as a function of the number of templates per word based on template clustering with kNN decisions ($k = 1, 2, 3, 4$); (a) the top candidate is the correct word, (b) the correct word is among the top 5 candidates (after Rabiner et al. [5]). 278
5.13 Average digit error rate as a function of the number of templates per word for a database of 1000 digits with clusters generated by the UWA and the MKM clustering procedures respectively (after Wilpon and Rabiner [7]). 280
5.14 The nonlinearity function, g, applied to the log energy difference between two frames for the energy distance calculation. 282
5.15 Plots of word-recognition error rate versus the analysis frame size (in samples) with various frame shift intervals (33, 67, 100 and 200 samples) (after Rabiner and Wilpon [13]). 283
5.16 Plots of word recognition error rate versus LPC analysis order for three LPC-related distortion measures, the likelihood ratio measure (LR), the bandpass liftered cepstral measure (BPCEP), and the cepstral measure with additional delta cepstrum (DCEP) (after Rabiner and Wilpon [13]). 284
5.17 Four different speaker adaptation scenarios. 286
5.18 Hierarchical codebook adaptation algorithm (after Furui [14]). 290
5.19 Cepstral distortion between input speech pattern and reference templates resulted from hierarchical code-word adaptation (NA=no adaptation); — progressive adaptation, --- direct adaptation (after Furui [14]). 290
5.20 Frame-distortion sequences between pairs of speech utterances: (a) utterances of the same word, (b) utterances of acoustically different words, and (c) utterances of acoustically confusing words (after Rabiner and Wilpon [15]). 293
5.21 Examples illustrating "word" alignment based on dynamic "phone" warping for word equivalence class clustering (after Rabiner and Wilpon [15]). 295
5.22 Plots of average distortion sequences (—) and sequences of standard deviations of the frame distortion (- - - -) for various word pairs (after Rabiner and Wilpon [15]). 299
5.23 Weighting curves for discriminatively comparing words "I" and "Y" (after Rabiner and Wilpon [15]). 300

5.24 Speech-recognition performance in noisy conditions; •: training and testing have matched S/N as indicated by the abscissa; △: only clean training reference is used and the abscissa indicates that the test S/N; □: training and testing S/N's are mismatched with test S/N fixed at 18 dB and the abscissa indicates the training S/N (after Dautrich et al. [17]). 306
5.25 Noise spectrum in a typical personal office with a SUN 3/110. 307
5.26 Mean customer-premises-to-customer-premises attenuation distortion relative to 1004 Hz in a telephone channel (after Carey et al. [21]). 308
5.27 Schematic of the two-input noise cancelling approach. 310
5.28 Noisy speech-recognition performance of several distortion measures and methods (after Mansour and Juang [41]). 316

6.1 A Markov chain with five states (labeled 1 to 5) with selected state transitions. 323
6.2 Markov model of the weather. 323
6.3 Three possible Markov models that can account for the results of hidden coin-tossing experiments. (a) one-coin model, (b) two-coins model, (c) three-coins model. 328
6.4 An N-state urn-and-ball model illustrating the general case of a discrete symbol HMM. 329
6.5 (a) Illustration of the sequence of operations required for the computation of the forward variable $\alpha_{t+1}(j)$. (b) Implementation of the computation of $\alpha_t(i)$ in terms of a lattice of observations t, and states i. 336
6.6 Sequence of operations required for the computation of the backward variable $\beta_t(i)$. 338
6.7 Illustration of the sequence of operations required for the computation of the joint event that the system is in state i at time t and state j at time $t + 1$. 342
6.8 Illustration of three distinct types of HMMs. (a) A 4-state ergodic model. (b) A 4-state left-right model. (c) A 6-state parallel path left-right model. 349
6.9 Equivalence of a state with a mixture density to a multistate single-density distribution (after Juang et al. [21]). 351
6.10 Examples of networks incorporating null transitions. (a) Left-right model. (b) Finite state network. (c) Grammar network. 357
6.11 Illustration of general interstate connections of (a) a normal HMM with exponential state duration density, and (b) a variable duration HMM with specified state densities and no self-transitions from a state back to itself. 358
6.12 Example of how the process of deleted interpolation can be represented using a state diagram. 373

6.13 Block diagram of an isolated word HMM recognizer (after Rabiner [38]). 379
6.14 Average word error rate (for a digits vocabulary) versus the number of states N in the HMM (after Rabiner et al. [18]). 380
6.15 Comparison of estimated density (jagged contour) and model density (smooth contour) for each of the nine components of the observation vector (eight cepstral components, one log energy component) for state 1 of the digit zero (after Rabiner et al. [38]). 381
6.16 Average word error rate as a function of the minimum discrete density value ϵ (after Rabiner et al. [18]). 382
6.17 The segmental k-means training procedure used to estimate parameter values for the optimal continuous mixture density fit to a finite number of observation sequences (after Rabiner et al. [38]). 383
6.18 Plots of (a) log energy; (b) accumulated log likelihood; and (c) state assignment for one occurrence of the word "six" (after Rabiner et al. [38]). 384
6.19 Histograms of the normalized duration density for the five states of the digit "six" (after Rabiner et al. [38]). 385

7.1 Illustration of an isolated string of digits (upper panel) and a fluently spoken version of the same digit string (lower panel). 391
7.2 Illustration of the connected word-recognition problem. 392
7.3 Determination of the optimum alignment of super-reference pattern R^s to T, along with corresponding word boundary frames. 394
7.4 Computation ranges for matching R_v against portions of T. 395
7.5 Use of range limiting to reduce the size of individual time warps. 396
7.6 Series of paths ending at frame e. 397
7.7 Illustration of standard DTW alignment of super-reference and test patterns (a), and level building alignment (b) (after Myers and Rabiner [4]). 400
7.8 Implementation of level 1 of the level building algorithm (after Myers and Rabiner [4]). 402
7.9 Implementation of level 2 of the level building algorithm (after Myers and Rabiner [4]). 404
7.10 Simple example illustrating level building on two reference patterns of equal length (after Myers and Rabiner [4]). 405
7.11 Computation region of level building algorithm for a fixed-length reference pattern (after Myers and Rabiner [4]). 406
7.12 Reduced computation region using upper- and lower-level constraints (after Myers and Rabiner [4]). 407
7.13 Overall computation pattern of level building algorithm for variable length reference patterns (after Myers and Rabiner [4]). 408

Ex.1022 / Page 21 of 333
AMAZON.COM, INC.

7.14 Illustration of beginning range reduction (after Myers and Rabiner [4]). 412
7.15 Illustration of global range reduction (after Myers and Rabiner [4]). 413
7.16 Illustration of the use of reference pattern uncertainty regions (after Myers and Rabiner [4]). 413
7.17 Summary of level building implementation of computation reduction methods (after Myers and Rabiner [4]). 414
7.18 Level building of the string "51560" (after Myers and Rabiner [4]). 417
7.19 Level building of the string "99211" (after Myers and Rabiner [4]). 418
7.20 The one-pass connected word recognition algorithm (after Bridle et al. [6]). 418
7.21 Combinatorics for the one-pass algorithm. 419
7.22 Description of procedure for determining multiple candidate scores (after Myers and Rabiner [4]). 421
7.23 Candidate strings for a four-level search (after Myers and Rabiner [4]). 422
7.24 Illustration of level building flaw for determining the second-best candidate string (after Myers and Rabiner [10]). 422
7.25 Use of HMMs in the level building procedure. 424
7.26 A typical grammar node of an FSN grammar network (after Lee and Rabiner [9]). 425
7.27 Block diagram of connected word recognition computation. 425
7.28 Three possible grammar networks for connected digit recognition (after Lee and Rabiner [9]). 426
7.29 The segmental *k*-means training algorithm for connected word strings (after Rabiner et al. [13]). 428
7.30 Block diagram of connected digit recognition method (after Rabiner et al. [13]). 429
7.31 Connected digit HMM (after Rabiner et al. [13]). 430
7.32 Average speaking rate of talkers in the two connected digit databases as a function of the number digits per string (after Rabiner et al. [13]). 431

8.1 HMM representations of a word (a) and a subword unit (b). 439
8.2 Representations of the acoustic space of speech by (a) partitioned VQ cells, (b) sets of continuous mixture Gaussian densities, and (c) a continuous-density codebook (after Lee et al. [7]). 440
8.3 Representation of a sentence, word, and subword unit in terms of FSNs. 442
8.4 Creation of composite FSN for sentence "Show all alerts." 443

8.5 Segmentations of a training utterance resulting from the segmental k-means training for the first several iterations (after Lee et al. [7]). 444
8.6 Segmentation of an utterance into PLUs (after Lee et al. [7]). 445
8.7 Overall block diagram of subword unit based continuous speech recognizer. 451
8.8 FSN for the NG syntax. 452
8.9 FSN of the WP syntax. 453
8.10 Word and sentence accuracies versus number of mixtures per state for the training subset using the WP syntax. 455
8.11 Word and sentence accuracies versus number of mixtures per state for the training subset using the NG syntax. 456
8.12 Word and sentence accuracies versus number of mixtures per state for the four test sets using the WP syntax. 457
8.13 Word and sentence accuracies versus number of mixtures per state for the four test sets using the NG syntax. 458
8.14 Plots of the number of intraword units, interword units, and combined units as a function of the count threshold. 462
8.15 Deleted interpolation model for smoothing discrete density models. 463
8.16 FSN representation of words with three or more units (a), two units (b), and one unit (c). 466
8.17 Word accuracy (%) as a function of the number of generalized triphone models for several training set sizes (after Lee [2]). 469
8.18 Histograms of the unit separation distance for (a) combined intraword and interword units (1282 PLUs) and (b) separate intraword and interword units (1769 PLUs) (after Lee et al. [19]). 471
8.19 Splitting of subword unit p_i into three clusters (after Lee et al. [7]). 472
8.20 Histograms of likelihood score for four context-independent units (vowels) (after Lee et al. [7]). 472
8.21 Average likelihood scores for sets of PLU models obtained from model splitting (after Lee et al. [7]). 474
8.22 Word networks based on all combinations of units (a), or selected combinations of units (b) (after Lee et al. [7]). 474
8.23 Word and sentence accuracy improvements in RM after semantic processing (after Pieraccini and Lee [26]). 479

9.1 Block diagram of a task-specific voice control and dialog system. 483
9.2 Market sales for speech-recognition hardware over time, in each of five market segments. (Data estimated for 1990–1992 sales.) 488

Ex.1022 / Page 23 of 333
AMAZON.COM, INC.

9.3 Average LPC distance scores for first- and second-recognition candidates for each of the six test speakers (after Rabiner et al. [1]). 489
9.4 Block diagram of a directory listing retrieval system based on spelled spoken names (after Rabiner et al. [1]). 492

Ex.1022 / Page 24 of 333
AMAZON.COM, INC.

LIST OF TABLES

2.1	A condensed list of phonetic symbols for American English.	24
2.2	Formant frequencies for typical vowels.	29
2.3	Sound Lexicon of Digits	41
4.1	Measured DLs and JNDs for Synthetic Vowel Sounds	152
4.2	Examples of Critical Bandwidth	186
4.3	Values of the Elements in the Warping Matrix	189
4.4	Summary of Spectral Distortion Measures	195
4.5	Summary of sets of local constraints and the resulting path specifications	211
4.6	Values of Q_{max} and Q_{min} for different types of paths	214
4.7	Summary of sets of local constraints, slope weights, and DP recursion formulas	223
5.1	Comparison of single-frame vowel-recognition error rates with various distortion measures (after Rabiner and Soong [8]).	275
5.2	Comparison of recognition error rates for a 39 alphadigit-word vocabulary with various distortion measures (after Nocerino et al. [9]).	276

Ex.1022 / Page 25 of 333
AMAZON.COM, INC.

5.3 Comparison of recognition error rate pertaining to several distortion measures with and without energy information (after Nocerino et al. [9]). 282
5.4 Performance of Isolated Word-Recognition Systems 285
5.5 Attenuation distortion relative to 1004 Hz: short connections. 308

6.1 Average Digit Error Rates for Several Recognizers and Evaluation Sets 386

7.1 Average String Error Rates (%) for Connected Digit Recognition Tests 432

8.1 Set of basic PLUs for speech. 438
8.2 Typical word pronunciations (word lexicon) based on context-independent PLUs. 438
8.3 PLU statistics on count and average likelihood score. 446
8.4 Number of intra-word CD units as a function of count threshold, T. 461
8.5 Word error rates as a function of occurrence threshold for the feb 89 test set using intraword units with a 38 component/vector analysis. 468
8.6 Word error rates as a function of occurrence threshold for the feb 89 test set using both intraword and interword units (independently) with a 38 component/vector analysis. 468
8.7 Recognition performance on 122-word, office correspondence task with both VI and VD models (after Hon & Lee [25]). 478
8.8 Recognition performance on 991-word, RM task, with both VI and VD models (after Hon & Lee [25]). 478

9.1 Performance scores for several types of speech-recognition systems as measured under laboratory conditions. 484
9.2 Projections for speech recognition. 494

PREFACE

This book is an outgrowth of an association between the authors which started over 10 years ago when one of us (BHJ) was a graduate student at the University of California at Santa Barbara and the other (LRR) was a supervisor at AT&T Bell Laboratories. We began our relationship with a mutual interest in the problem of designing and implementing vector quantization for speech processing. This association turned into a full technical partnership and strong friendship when Fred Juang joined Bell Laboratories, initially in the development area and subsequently in research. The spark that ignited formal work on this book was a series of short courses taught by one of us (LRR) on speech recognition. After several iterations of teaching, it became clear that the area of speech recognition, although still changing and growing, had matured to the point where a book that covered its theoretical underpinnings was warranted.

Once we had decided to write this book, there were several key issues that had to be resolved, including how deep to go into areas like linguistics, natural language processing, and the practical side of the problem; whether to discuss individual systems proposed by various research labs around the world; and how extensively to cover applications. Although there were no simple answers to these questions, it rapidly became obvious to us that the fundamental goal of the book would be to provide a theoretically sound, technically accurate, and reasonably complete description of the basic knowledge and ideas that constitute a modern system for speech recognition by machine. With these basic guiding principles in mind, we were able to decide consistently (and hopefully reasonably) what material had to be included, and what material would be presented in only a cursory

Ex.1022 / Page 27 of 333
AMAZON.COM, INC.

manner. We leave it up to you, the reader, to decide if our choices have been wise ones.

The formal organization of the book is as follows. Chapter 1, called "Fundamentals of Speech Recognition," provides an overview of the entire field with a discussion of the breadth and depth of the various disciplines that are required for a deep understanding of all aspects of speech recognition. The concept of a task-oriented, speech-recognition system is introduced and it is shown that "base level" speech or sound recognition is only one step in a much larger process where higher-level task information, in the form of syntax, semantics, and pragmatics, can often play a major role. After a formal description of the material to be covered in each of the chapters, we give a brief history of speech recognition research in order to put the material presented in this book in its proper perspective.

Chapter 2, entitled the "The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization," provides a review of the theory of acoustic-phonetics in which we try to characterize basic speech sounds according to both their linguistic properties and the associated acoustic measurements. We show that although there is a solid basis for the linguistic description of sounds and a good understanding of the associated acoustics of sound production, there is, at best, a tenuous relationship between a given linguistic sound and a repeatable, reliable, measureable set of acoustic parameters. As such a wide variety of approaches to speech recognition have been proposed, including those based on the ideas of acoustic-phonetics, statistical pattern-recognition methods, and artificial intelligence (so-called expert system) ideas. We discuss the relative advantages and disadvantages of each of these approaches and show why, on balance, the pattern-recognition approach has become the method of choice for most modern systems.

In Chapter 3, entitled "Signal Processing and Analysis Methods for Speech Recognition," we discuss the fundamental techniques used to provide the speech features used in all recognition systems. In particular we discuss two well-known and widely used methods of spectrum analysis, namely the filter bank approach and the linear prediction method. We also show how the method of vector quantization can be used to code a spectral vector into one of a fixed number of discrete symbols in order to reduce the computation required in a practical system. Finally we discuss an advanced spectral analysis method that is based on processing within the human auditory system—an ear model. The ultimate goal of such a system is to increase the robustness of the signal representation and make the system relatively insensitive to noise and reverberation, in much the same way as the human ear.

Chapter 4, entitled "Pattern-Comparison Techniques," deals with the fundamental problems of defining speech feature vector patterns (from spoken input), and comparing pairs of feature vector patterns both locally (i.e., at some point in time), and globally (i.e., over the entire pattern) so as to derive a measure of similarity between patterns. To solve this pattern-comparison problem requires three types of algorithms, namely a speech-detection method (which essentially separates the speech signal from the background), a spectral vector comparison method (which compares two individual spectral vectors), and a global pattern comparison method which aligns the two patterns locally in time and compares the aligned patterns over the entire duration of the patterns. It is shown that a key issue is the way in which time alignment between patterns is achieved.

Chapter 5, entitled "Speech Recognition System Design and Implementation Issues," discusses the key issues of training a speech recognizer and adapting the recognizer pa-

rameters to different speakers, transmission conditions, and speaking environments. A key concept in most modern systems is that of learning, namely improving recognizer performance over time based on additional training provided by the user of the system. Adaptation methods provide a formalism for such learning.

In Chapter 6, "Theory and Implementation of Hidden Markov Models," we discuss a basic set of statistical modeling techniques for characterizing speech. The collection of methods, popularly called Hidden Markov Models, is a powerful set of tools for providing a statistical model of both the static properties of sounds and the dynamical changes that occur across sounds. Methods for time aligning patterns with models are discussed along with different ways of building the statistical models based on the type of representation, the sound being modeled, the class of talkers, and so forth.

Chapters 7 and 8, entitled "Speech Recognition Based on Connected Word Models" and "Large Vocabulary Continuous Speech Recognition," extend the speech-recognition problem from single word sequences to fluent speech. Modeling techniques based on whole word models are discussed in Chapter 7 where we assume that we are interested in recognizing sequences of digits, alphanumerics, and so forth. For this type of system whole-word models are most reasonable since the vocabulary is typically small and highly constrained. Hence the statistical properties of the word models, in all word contexts, can be learned from a reasonably sized training set. Modeling techniques based on subword units are discussed in Chapter 8 where we assume unlimited size vocabulary. Hence a key issue is what units are used, how context dependent the units should be, how unit models are trained reliably (and robustly to different vocabularies and tasks), and how large vocabulary recognition systems based on such units are efficiently implemented.

Finally, in Chapter 9, entitled "Task-Oriented Applications of Automatic Speech Recognition," we come full circle and return to the concept of a task-oriented system. We discuss the basic principles that make some tasks successful while others fail. By way of example we discuss, in fairly general terms, a couple of task-oriented recognizers and show how they perform in practice.

The material in this book is primarily intended for the practicing engineer, scientist, linguist, programmer, and so forth, who wants to learn more about this fascinating field. We assume a basic knowledge of signal processing and linear system theory as provided in an entry level course in digital signal processing. Although not intended as a formal university course, the material in this book is indeed suitable for a one-semester course at the graduate or high undergraduate level. Within almost every chapter we have provided "exercises" for the student to assess how well he or she understands the material. Solutions to the exercises are provided immediately following the exercise. Hence, for maximum effectiveness, each student must exercise self-discipline to work through an answer before comparing it with the published solution.

In order to truly understand the fundamentals of speech recognition, a person needs hands-on experience with the software, hardware, and platforms. Hence we strongly encourage all serious readers of this book to program the algorithms, implement the systems, and literally build applications. Without such practical experience the words in this book will not come alive for most people.

ACKNOWLEDGMENTS

Although the authors take full responsibility for the material presented in this book, we owe a great debt to our colleagues, both with AT&T Bell Laboratories and outside, for their many technical contributions which underly the material presented. In particular the authors owe a debt of gratitude to Dr. James L. Flanagan (currently director of the CAIP Institute at Rutgers University) for his roles in guiding and shaping both our careers and the field of speech processing. Without Jim's understanding and inspiration, this book would never have existed.

The number of people who have made substantial contributions to speech recognition are too numerous to mention. However there are three individuals who have had a profound influence on the field and they deserve special mention. The first is Professor Raj Reddy of Carnegie-Mellon University who was essentially the first person to realize the vast potential of speech recognition and has devoted over 25 years as a leader, innovator, and educator in this field. The second individual of note is Dr. Jack Ferguson (retired from Institute for Defense Analyses in Princeton) who is the person most responsible for development of the theory of the Hidden Markov Model as applied to speech recognition. Dr. Ferguson, as editor of the key textbook in this area and lecturer, par excellence, has spread the word on Hidden Markov Models so that this technology has rapidly risen from technical obscurity to become the preeminent method of speech recognition today. Finally, the third individual of note is Dr. Fred Jelinek of IBM, who has led the world's largest speech-recognition research group for almost two decades and has been responsible for a large number of major innovations in large vocabulary speech recognition. These three individuals have played major roles in nurturing the technology and enabling it to reach the state of maturity it has achieved today.

Within Bell Laboratories the authors have drawn freely from the research of our former and current colleagues. We would like to acknowledge the direct support and contributions of the following individuals: Prof. Ronald Schafer (currently at Georgia Tech.), Dr. Steve Levinson, Dr. Bishnu Atal, Dr. Esther Levin, Dr. Tali Tishby (currently at the Hebrew University in Jerusalem), Dr. Oded Ghitza, Jay Wilpon, Dr. Frank Soong, Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Cory Myers (currently at Atlantic Aerospace), Dr. Aaron Rosenberg, Dr. Chin Lee and Dr. Roberto Pieraccini. We thank these colleagues for their research contributions and for their friendship and guidance over the years.

Several individuals provided technical comments on a preliminary version of the manuscript. These comments were invariably insightful and provided valuable feedback on ways to improve the book. We would like to acknowledge and thank the following individuals for their help: Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Esther Levin, Dr. Oded Ghitza, Dr. Roberto Pieraccini, Dr. Chin Lee, Dr. Wu Chou and Dr. Sadaoki Furui (NTT Corporation, JAPAN).

The production of a book is an essentially infinite task which is seemingly an endless one. However all good things do come to an end and this one was no exception. The authors owe a great deal to Ms. Martina Sharp of the Bell Labs Word Processing group, who entered all the text material for the book using the LaTeX system. Tina worked on three

generations of the manuscript with a degree of skill that is essentially unmatched among her peers. It was a pleasure working with Tina and we look forward to future projects with her. Most of the art work for the book was produced by Ms. Danuta Sowinska-Khan of the Bell Labs art department. Danuta was a pleasure to work with and is a professional in her trade. The material she produced always met the highest professional standards and did honor to the manuscript in which it appears. A great deal of assistance in the preparation of drafts of the book, getting figures ready for production, and so forth was provided by Irene Morrongiello. We thank her and wish to express our appreciation for a job well done.

Finally, to the crew at Prentice Hall we owe a debt of gratitude for the professional way the book was handled from inception to final production. Karen Gettman provided the incentive to the authors to actually produce the manuscript and make the book a reality. We thank Lisa Garboski, the production editor, for help in all phases of the book production cycle.

Lawrence R. Rabiner
Biing-Hwang Juang

Chapter 1

FUNDAMENTALS OF SPEECH RECOGNITION

1.1 INTRODUCTION

Automatic recognition of speech by machine has been a goal of research for more than four decades and has inspired such science fiction wonders as the computer HAL in Stanley Kubrick's famous movie *2001—A Space Odyssey* and the robot R2D2 in the George Lucas classic *Star Wars* series of movies. However, in spite of the glamour of designing an intelligent machine that can recognize the spoken word and comprehend its meaning, and in spite of the enormous research efforts spent in trying to create such a machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments. Thus, an important question in this book is, What do we mean by "speech recognition by machine." Another important question is, How can we build a series of bridges that will enable us to advance both our knowledge as well as the capabilities of modern speech-recognition systems so that the "holy grail" of conversational speech recognition and understanding by machine is attained?

Because we do not know how to solve the ultimate challenge of speech recognition, our goal in this book is to give a series of presentations on the fundamental principles of most modern, successful speech-recognition systems so as to provide a framework from which researchers can expand the frontier. We will attempt to avoid making absolute judgments on the relative merits of various approaches to particular speech-recognition problems. Instead we will provide the theoretical background and justification for each topic discussed so that the reader is able to understand why the techniques have proved

Ex.1022 / Page 32 of 333
AMAZON.COM, INC.

valuable and how they can be used to advantage in practical situations.

One of the most difficult aspects of performing research in speech recognition by machine is its interdisciplinary nature, and the tendency of most researchers to apply a monolithic approach to individual problems. Consider the disciplines that have been applied to one or more speech-recognition problems:

1. **signal processing**—the process of extracting relevant information from the speech signal in an efficient, robust manner. Included in signal processing is the form of spectral analysis used to characterize the time-varying properties of the speech signal as well as various types of signal preprocessing (and postprocessing) to make the speech signal robust to the recording environment (signal enhancement).
2. **physics (acoustics)**—the science of understanding the relationship between the physical speech signal and the physiological mechanisms (the human vocal tract mechanism) that produced the speech and with which the speech is perceived (the human hearing mechanism).
3. **pattern recognition**—the set of algorithms used to cluster data to create one or more prototypical patterns of a data ensemble, and to match (compare) a pair of patterns on the basis of feature measurements of the patterns.
4. **communication and information theory**—the procedures for estimating parameters of statistical models; the methods for detecting the presence of particular speech patterns, the set of modern coding and decoding algorithms (including dynamic programming, stack algorithms, and Viterbi decoding) used to search a large but finite grid for a best path corresponding to a "best" recognized sequence of words.
5. **linguistics**—the relationships between sounds (phonology), words in a language (syntax), meaning of spoken words (semantics), and sense derived from meaning (pragmatics). Included within this discipline are the methodology of grammar and language parsing.
6. **physiology**—understanding of the higher-order mechanisms within the human central nervous system that account for speech production and perception in human beings. Many modern techniques try to embed this type of knowledge within the framework of artificial neural networks (which depend heavily on several of the above disciplines).
7. **computer science**—the study of efficient algorithms for implementing, in software or hardware, the various methods used in a practical speech-recognition system.
8. **psychology**—the science of understanding the factors that enable a technology to be used by human beings in practical tasks.

Successful speech-recognition systems require knowledge and expertise from a wide range of disciplines, a range far larger than any single person can possess. Therefore, it is especially important for a researcher to have a good understanding of the fundamentals of speech recognition (so that a range of techniques can be applied to a variety of problems), without necessarily having to be an expert in each aspect of the problem. It is the purpose of this book to provide this expertise by giving in-depth discussions of a number of

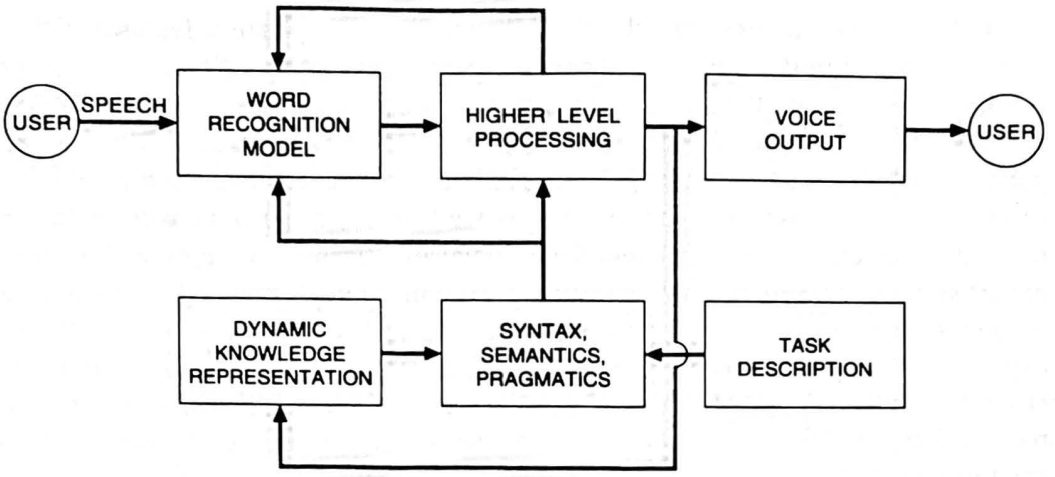


Figure 1.1 General block diagram of a task-oriented speech-recognition system.

fundamental topics in speech-recognition research.

1.2 THE PARADIGM FOR SPEECH RECOGNITION

A general model for speech recognition, shown in Figure 1.1, is used throughout this book. The model begins with a user creating a speech signal (speaking) to accomplish a given task. The spoken output is first recognized in that the speech signal is decoded into a series of words that are meaningful according to the syntax, semantics, and pragmatics of the recognition task. The meaning of the recognized words is obtained by a higher-level processor that uses a dynamic knowledge representation to modify the syntax, semantics, and pragmatics according to the context of what it has previously recognized. In this manner, things such as non sequitors are omitted from consideration at the risk of misunderstanding, but at the gain of minimizing errors for sequentially meaningful inputs. The feedback from the higher-level processing box reduces the complexity of the recognition model by limiting the search for valid (acceptable) input sentences (speech) from the user. The recognition system responds to the user in the form of a voice output, or equivalently, in the form of the requested action being performed, with the user being prompted for more input.

1.3 OUTLINE

The material in this book is organized into nine chapters. Chapters 2 through 9 each deals with a basic concept or a fundamental technique used in various speech-recognition systems. The material discussed in these chapters is as follows.

Chapter 2—The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization. In this chapter we review the speech production/perception process in human beings. We show how different speech sounds can be characterized by a set of

spectral and temporal properties that depend on the acoustic-phonetic features of the sound and are manifest in the waveform, the sound spectrogram, or both. Included in the chapter is an overview of the three most common approaches to speech recognition, namely the acoustic-phonetic approach (which tries to directly exploit individual sound properties), the pattern recognition approach (which relies only on gross spectral and temporal properties of speech sounds and uses conventional as well as neural network pattern recognition technology to classify sounds), and the artificial intelligence (AI) approach in which an expert system or a self-organizing (learning) system, as implemented by neural networks, is used to classify sounds. We discuss the strengths and weaknesses of each approach and explain why the pattern-recognition approach is the one most heavily relied on in practical systems. We conclude the chapter with a discussion of the fundamental issues in speech recognition (i.e., those factors that most influence overall system performance), and with a brief overview of current applications.

Chapter 3—Signal Processing and Analysis Methods for Speech Recognition. In this chapter we present the two fundamental signal-processing approaches to speech spectral analysis: filter bank and linear predictive methods. We specialize the presentation of these two fundamental techniques to aspects related to speech analysis and compare and contrast the two methods in terms of robustness to speech sounds and required computation. For completeness we also discuss the popular source-coding technique referred to as vector quantization (VQ). Here, a codebook is created to represent the anticipated range of spectral vectors. This enables us to code an arbitrary continuous speech spectral vector into one of a fixed number of discrete codebook symbols at the cost of increased error in signal representation but with the benefit of significantly reduced computation in the recognition process. We conclude this chapter with a discussion of a spectral analysis model that attempts to mimic the processing in the human auditory system—the so-called ear model. Although our knowledge of the higher-order processing in the central nervous system is rudimentary, the importance of ear models is related to their robustness to noise, reverberation, and other environmental factors that often seriously degrade performance of current speech recognizers.

Chapter 4—Pattern-Comparison Techniques. In this chapter we discuss three fundamental aspects of comparing a pair of speech patterns. These are the basic concept of detecting speech (i.e., finding the speech signal in a background of noise or other acoustic interference), the idea of computing a measure of the local distance (or similarity) of a pair of spectral representations of a short-time piece of speech signal (a distance or distortion measure), and the concept of temporally aligning and globally comparing a pair of speech patterns corresponding to different speech utterances that may or may not be the same sequence of sounds or words (dynamic time warping algorithms). We show in this chapter how the basic pattern-comparison techniques can be combined in a uniform framework for speech-recognition applications.

Chapter 5—Speech-Recognition System Design and Implementation Issues. In this chapter we discuss the remaining pieces (after signal processing and pattern comparison) that enable us to build and study performance of a practical speech-recognition system.

Ex.1022 / Page 35 of 333
AMAZON.COM, INC.

In particular we discuss how speech recognizers are trained and how we can enhance the basic recognition procedure by adding features, by exploiting a preprocessor, by the use of methods of adaptation or by postprocessing the recognizer outputs using a set of pattern discriminators (as opposed to the pattern classifiers used in a conventional implementation). We conclude the chapter with a discussion of various ways of recognizing speech in adverse environments (e.g., noise, stress conditions, or mismatch between training and testing).

Chapter 6—Theory and Implementation of Hidden Markov Models. In this chapter we discuss aspects of the theory and implementation of the set of statistical modeling techniques collectively referred to as hidden Markov modeling. Included within these techniques are the algorithms for scoring a statistical (Markovian) model against a speech pattern, the techniques for aligning the model with the speech pattern so as to recover an estimate of the alignment path between different speech sounds and different model states, and the techniques for estimating parameters of the statistical models from a training set of utterances of the sounds being modeled. Also included is a discussion of the practical aspects of building hidden Markov models, including the issues of scaling of data, handling of multiple observation sequences, providing initial estimates of model parameters, and combating the problems of insufficient training data. We conclude the chapter with a practical example illustrating how a simple, isolated word recognizer would be implemented using hidden Markov models.

Chapter 7—Speech Recognition Based on Connected Word Models. In this chapter we show how the basic set of techniques developed for recognizing an isolated word or phrase can be readily extended to recognizing a sequence of words (e.g., a string of digits of a credit card number) spoken in a fluent or connected manner. We make the simplifying assumption that the connected word string is recognized by finding the optimal sequence of word models that best matches the spoken string. Hence we assume that the word is the basic recognition unit for these systems, and therefore the training problem is one of estimating the optimal parameters of word models on the basis of training data, which need not contain isolated versions of the words. We describe three "optimal" approaches to solving the recognition part of connected word-recognition problems: (1) the two-level dynamic programming method, (2) the level building method, and (3) the time synchronous level building (or the one-pass) method and discuss the properties, and the relative strengths and weaknesses of each method. We then show how we can optimally train connected word systems, even if isolated versions of the vocabulary words are not available. We conclude the chapter with a discussion of a connected digit recognizer implemented using the methods described in the chapter.

Chapter 8—Large Vocabulary Continuous Speech Recognition. In this chapter we discuss the issues in applying speech-recognition technology to the problem of recognizing fluently spoken speech with vocabulary sizes of 1000 or more words (with unlimited vocabularies as the ultimate goal). It is shown that a number of fundamental problems must be solved to implement such a system, including the choice of a basic subword speech unit (from which words, phrases, and sentences can be built up), an effective way

of modeling the basic speech unit, a way of deriving models of the unit, a way of designing and implementing a word lexicon (which provides a mapping between words and subword units), a way of implementing task syntax (the system grammar), a way of implementing the overall recognition part of the system (via some type of network search), and a way of imposing task semantics onto the solution. We concentrate primarily on the issues involved in building large vocabulary recognition systems. For illustrative purposes we describe one reasonable way of building such a system and discuss the resulting performance on a standard database management task.

Chapter 9—Task Oriented Applications of Automatic Speech Recognition. The final chapter of the book provides a brief overview of how one might apply the ideas discussed in the book to building a real, task-oriented, speech recognition system. It includes discussions of how one would evaluate recognizer performance and how one might decide whether a proposed task is viable for speech recognition. We also discuss a set of broad classes of applications, which appear to be the most promising ones at this time, along with typical examples of how recognizers have been successfully employed within these broad classes. The chapter concludes with some broad performance projections through the year 2000.

1.4 A BRIEF HISTORY OF SPEECH-RECOGNITION RESEARCH

Research in automatic speech recognition by machine has been done for almost four decades. To gain an appreciation for the amount of progress achieved over this period, it is worthwhile to briefly review some research highlights. The reader is cautioned that such a review is cursory, at best, and must therefore suffer from errors of judgment as well as omission.

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [1]. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words [2]. The system again relied on spectral measurements (as provided by an analog filter bank) primarily during vowel regions. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [3]. They used a spectrum analyzer and a pattern matcher to make the recognition decision. A novel aspect of this research was the use of statistical information about allowable sequences of phonemes in English (a rudimentary form of language syntax) to improve overall phoneme accuracy for words consisting of two or more phonemes. Another effort of note in this period was the vowel recognizer of Forgie and Forgie, constructed at MIT Lincoln Laboratories in 1959, in which 10 vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker-independent manner [4]. Again a filter bank analyzer was used to provide spectral information, and a

time varying estimate of the vocal tract resonances was made to decide which vowel was spoken.

In the 1960s several fundamental ideas in speech recognition surfaced and were published. However, the decade started with several Japanese laboratories entering the recognition arena and building special-purpose hardware as part of their systems. One early Japanese system, described by Suzuki and Nakata of the Radio Research Lab in Tokyo [5], was a hardware vowel recognizer. An elaborate filter bank spectrum analyzer was used along with logic that connected the outputs of each channel of the spectrum analyzer (in a weighted manner) to a vowel-decision circuit, and a majority decision logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita of Kyoto University in 1962, who built a hardware phoneme recognizer [6]. A hardware speech segmenter was used along with a zero-crossing analysis of different regions of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963 [7]. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program.

In the 1960s three key research projects were initiated that have had major implications on the research and development of speech recognition for the past 20 years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems associated with nonuniformity of time scales in speech events. Martin developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduced the variability of the recognition scores [8]. Martin ultimately developed the method and founded one of the first companies, Threshold Technology, which built, marketed, and sold speech-recognition products. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances [9]. Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected word recognition, were embodied in Vintsyuk's work, it was largely unknown in the West and did not come to light until the early 1980s; this was long after the more formal methods were proposed and implemented by others.

A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes [10]. Reddy's research eventually spawned a long and highly successful speech-recognition research program at Carnegie Mellon University (to which Reddy moved in the late 1960s) which, to this day, remains a world leader in continuous-speech-recognition systems.

In the 1970s speech-recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in Russia [11], Sakoe and Chiba in Japan [12], and Itakura in the United States [13]. The Russian studies helped advance the use of pattern-recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura's research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low-bit-rate speech coding, could be extended to speech-

recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.

Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which researchers studied three distinct tasks over a period of almost two decades, namely the New Raleigh language [14] for simple database queries, the laser patent text language [15] for transcribing laser patents, and the office correspondence task, called Tangora [16], for dictation of simple memos.

Finally, at AT&T Bell Labs, researchers began a series of experiments aimed at making speech-recognition systems that were truly speaker independent [17]. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker-independent patterns are now well understood and widely used.

Just as isolated word recognition was a key focus of research in the 1970s, the problem of connected word recognition was a focus of research in the 1980s. Here the goal was to create a robust system capable of recognizing a fluently spoken string of words (e.g., digits) based on matching a concatenated pattern of individual words. A wide variety of connected word-recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) [18], the one-pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in England [19], the level building approach of Myers and Rabiner at Bell Labs [20], and the frame synchronous level building approach of Lee and Rabiner at Bell Labs [21]. Each of these "optimal" matching procedures had its own implementational advantages, which were exploited for a wide range of tasks.

Speech research in the 1980s was characterized by a shift in technology from template-based approaches to statistical modeling methods—especially the hidden Markov model approach [22, 23]. Although the methodology of hidden Markov modeling (HMM) was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analyses (IDA), and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs, in the mid-1980s, that the technique became widely applied in virtually every speech-recognition research laboratory in the world.

Another "new" technology that was reintroduced in the late 1980s was the idea of applying neural networks to problems in speech recognition. Neural networks were first introduced in the 1950s, but they did not prove useful initially because they had many practical problems. In the 1980s, however, a deeper understanding of the strengths and limitations of the technology was obtained, as well as the relationships of the technology to classical signal classification methods. Several new ways of implementing systems were also proposed [24, 25].

Finally, the 1980s was a decade in which a major impetus was given to large vocabulary, continuous-speech-recognition systems by the Defense Advanced Research Projects Agency (DARPA) community, which sponsored a large research program aimed at achieving high word accuracy for a 1000-word, continuous-speech-recognition, database management task. Major research contributions resulted from efforts at CMU (notably the well-

Ex.1022 / Page 39 of 333
AMAZON.COM, INC.

known SPHINX system) [26], BBN with the BYBLOS system [27], Lincoln Labs [28], SRI [29], MIT [30], and AT&T Bell Labs [31]. The DARPA program has continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer, and the task shifting to retrieval of air travel information. At the same time, speech-recognition technology has been increasingly used within telephone networks to automate as well as enhance operator services.

REFERENCES

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits," *J. Acoust. Soc. Am.*, 24 (6): 637–642, 1952.

[2] H. F. Olson and H. Belar, "Phonetic Typewriter," *J. Acoust. Soc. Am.*, 28 (6): 1072–1081, 1956.

[3] D. B. Fry, "Theoretical Aspects of Mechanical Speech Recognition"; and P. Denes, "The Design and Operation of the Mechanical Speech Recognizer at University College London," *J. British Inst. Radio Engr.*, 19: 4, 211–229, 1959.

[4] J. W. Forgie and C. D. Forgie, "Results Obtained From a Vowel Recognition Computer Program," *J. Acoust. Soc. Am.*, 31 (11): 1480–1489, 1959.

[5] J. Suzuki and K. Nakata, "Recognition of Japanese Vowels—Preliminary to the Recognition of Speech," *J. Radio Res. Lab*, 37 (8): 193–212, 1961.

[6] T. Sakai and S. Doshita, "The Phonetic Typewriter, Information Processing 1962," *Proc. IFIP Congress*, Munich, 1962.

[7] K. Nagata, Y. Kato, and S. Chiba, "Spoken Digit Recognizer for Japanese Language," *NEC Res. Develop.*, No. 6, 1963.

[8] T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[9] T. K. Vintsyuk, "Speech Discrimination by Dynamic Programming," *Kibernetika*, 4 (2): 81–88, Jan.–Feb. 1968.

[10] D. R. Reddy, "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," Tech. Report No. C549, Computer Science Dept., Stanford Univ., September 1966.

[11] V. M. Velichko and N. G. Zagoruyko, "Automatic Recognition of 200 Words," *Int. J. Man-Machine Studies*, 2: 223, June 1970.

[12] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-26 (1): 43-49, February 1978.

[13] F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23 (1): 67-72, February 1975.

[14] C. C. Tappert, N. R. Dixon, A. S. Rabinowitz, and W. D. Chapman, "Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery," Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146, 1971.

[15] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a Linguistic Statistical Decoder for the

Recognition of Continuous Speech," *IEEE Trans. Information Theory*, IT-21: 250–256, 1975.

[16] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proc. IEEE*, 73 (11): 1616–1624, 1985.

[17] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27: 336–349, August 1979.

[18] H. Sakoe, "Two Level DP Matching—A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27: 588–595, December 1979.

[19] J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," *Proc. Inst. Acoust. Autumn Conf.*, 25–28, November 1979.

[20] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-29: 284–297, April 1981.

[21] C. H. Lee and L. R. Rabiner, "A Frame Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37 (11): 1649–1658, November 1989.

[22] J. Ferguson, Ed., *Hidden Markov Models for Speech*, IDA, Princeton, NJ, 1980.

[23] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, 77 (2): 257–286, February 1989.

[24] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Mag.*, 4 (2): 4–22, April 1987.

[25] A. Weibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37: 393–404, 1989.

[26] K. F. Lee, H. W. Hon, and D. R. Reddy, "An Overview of the SPHINX Speech Recognition System," *IEEE Trans. Acoustics, Speech, Signal Proc.*, 38: 600–610, 1990.

[27] Y. L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, S. Roucos, and R. M. Schwartz, "BBYLOS: The BBN Continuous Speech Recognition System," *Proc. ICASSP 87*, 89–92, April 1987.

[28] D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proc. ICASSP 89*, Glasgow, Scotland, 449–452, May 1989.

[29] M. Weintraub et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. ICASSP 89*, Glasgow, Scotland, 699–702, May 1989.

[30] V. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT Summit Speech Recognition System: A Progress Report," *Proc. DARPA Speech and Natural Language Workshop*, 179–189, February 1989.

[31] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, 4: 127–165, 1990.

Chapter 2

THE SPEECH SIGNAL: PRODUCTION, PERCEPTION, AND ACOUSTIC-PHONETIC CHARACTERIZATION

2.1 INTRODUCTION

In this chapter we discuss the mechanics of producing and perceiving speech in human beings, and we show how an understanding of these processes leads naturally to several different approaches to speech recognition by machine. We begin by showing how the different classes of speech sounds, or phonetics, can each be characterized in terms of broad acoustic features whose properties are relatively invariant across words and speakers. The ideas of acoustic-phonetic characterization of sounds lead naturally to straightforward implementation of a speech-recognition algorithm based on sequential detection of sounds and sound classes. The strengths and weaknesses of such an approach are discussed. An alternative approach to speech recognition is to use standard pattern-recognition techniques in a framework in which all speech knowledge is "learned" via a training phase. We show that such a "blind" approach has some natural advantages for a wide range of speech-recognition systems. Finally we show how aspects of both the acoustic-phonetic approach and the pattern-recognition approach can be integrated into a hybrid method that includes techniques from artificial intelligence as well as neural network methods.

2.1.1 The Process of Speech Production and Perception in Human Beings

Figure 2.1 shows a schematic diagram of the speech-production/speech-perception process in human beings. The production (speech-generation) process begins when the talker

Ex.1022 / Page 42 of 333
AMAZON.COM, INC.

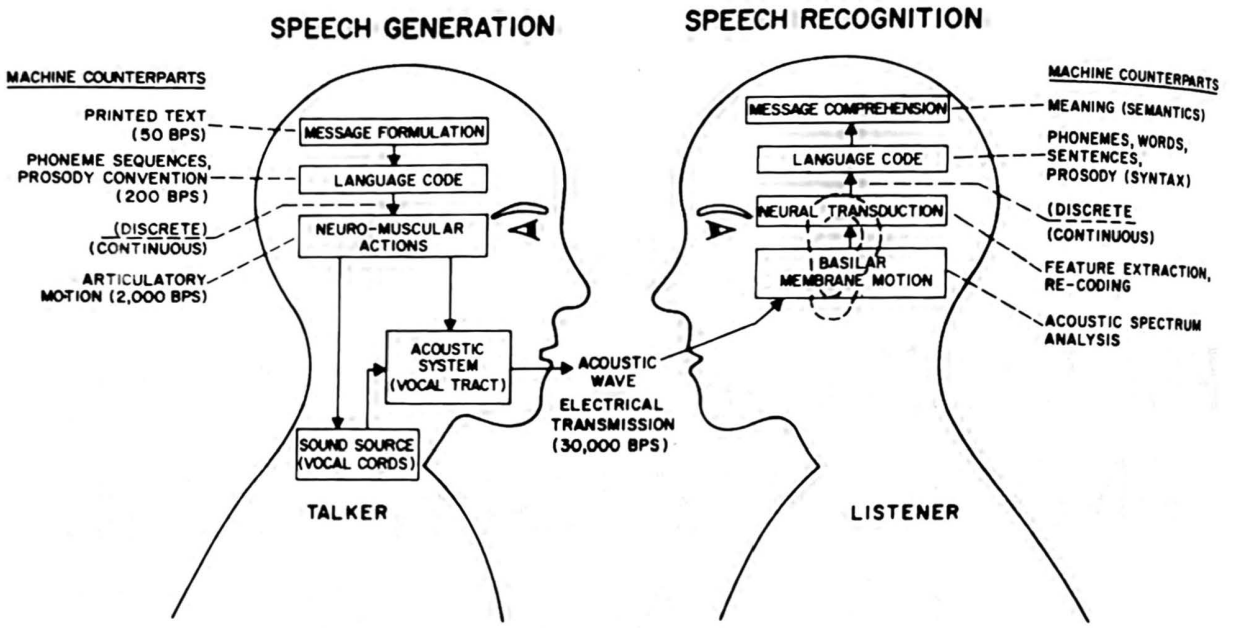


Figure 2.1 Schematic diagram of speech-production/speech-perception process (after Flanagan [unpublished]).

formulates a message (in his mind) that he wants to transmit to the listener via speech. The machine counterpart to the process of message formulation is the creation of printed text expressing the words of the message. The next step in the process is the conversion of the message into a language code. This roughly corresponds to converting the printed text of the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody markers denoting duration of sounds, loudness of sounds, and pitch accent associated with the sounds. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output. The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the lips, jaw, tongue, and velum (a "trapdoor" controlling the acoustic flow to the nasal mechanism).

Once the speech signal is generated and propagated to the listener, the speech-perception (or speech-recognition) process begins. First the listener processes the acoustic signal along the basilar membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process. In a manner that is not well understood, the neural activity along the auditory nerve is converted into a language code at the higher centers of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

A slightly different view of the speech-production/speech-perception process is shown in Figure 2.2. Here we see the steps in the process laid out along a line corresponding to the basic information rate of the signal (or control) at various stages of the

Ex.1022 / Page 43 of 333
AMAZON.COM, INC.

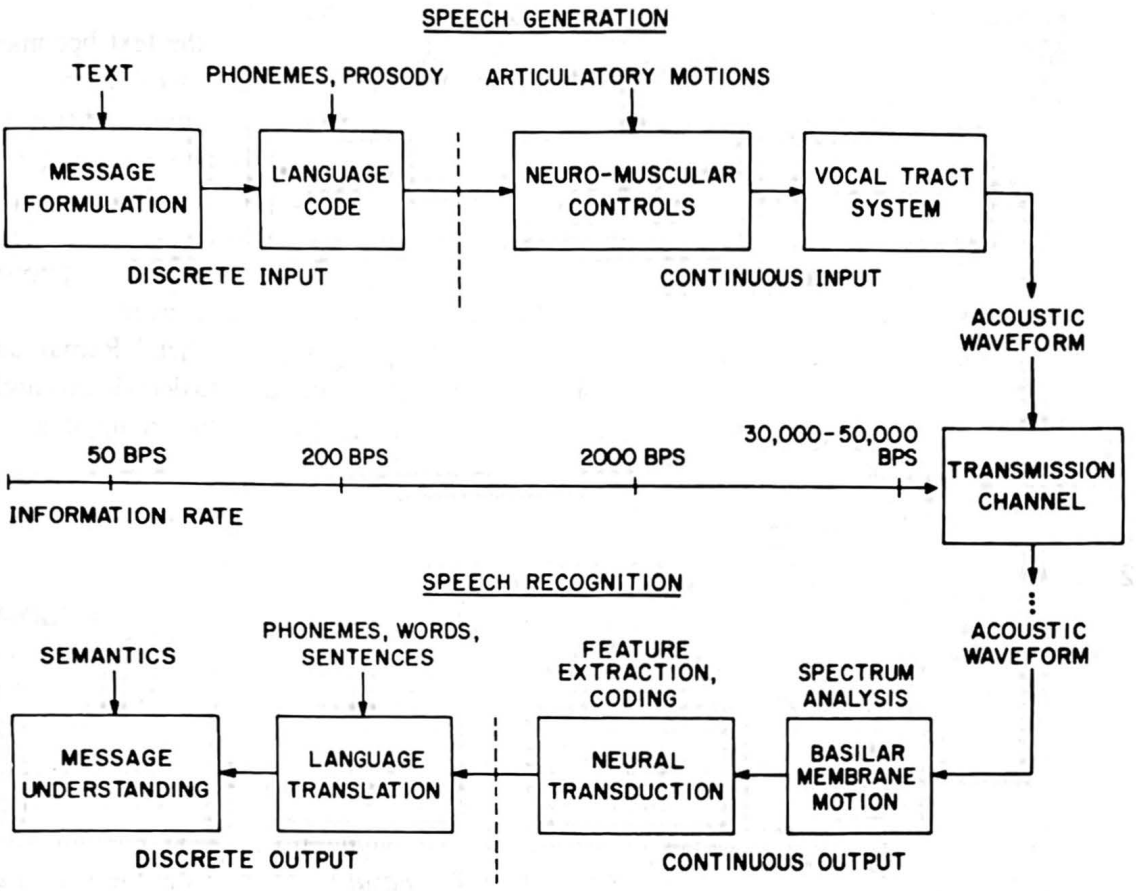


Figure 2.2 Alternative view of speech-production/speech-perception process (after Rabiner and Levinson [1]).

process. The discrete symbol information rate in the raw message text is rather low (about 50 bps [bits per second] corresponding to about 8 sounds per second, where each sound is one of about 50 distinct symbols). After the language code conversion, with the inclusion of prosody information, the information rate rises to about 200 bps. Somewhere in the next stage the representation of the information in the signal (or the control) becomes continuous with an equivalent rate of about 2000 bps at the neuromuscular control level, and about 30,000–50,000 bps at the acoustic signal level.

A transmission channel is shown in Figure 2.2 [1], indicating that any of several well-known coding techniques could be used to transmit the acoustic waveform from the talker to the listener. The steps in the speech-perception mechanism can also be interpreted in terms of information rate in the signal or its control and follows the inverse pattern of the production process. Thus the continuous information rate at the basilar membrane is in the range of 30,000–50,000 bps, while at the neural transduction stage it is about 2000 bps. The higher-level processing within the brain converts the neural signals to a discrete representation, which ultimately is decoded into a low-bit-rate message.

To illustrate, in a trivial way, how the speech-production/speech-perception process works, consider that the speaker has a goal of finding out whether his office mate has eaten his lunch yet. To express this thought, the speaker formulates the message "Did you eat

Ex.1022 / Page 44 of 333
AMAZON.COM, INC.

yet?" In the process of converting the message to a language code, the text becomes a phonetic sequence of sounds of the form /dId yu it yɛt?/, in which each word is expressed as a sequence of phonemes constituting the ideal pronunciation of the sounds of the word (as spoken in isolation) within the spoken language. However, because the words are not spoken in isolation, and a physical mechanism is used to produce the sounds (the human vocal tract system), and because physical systems obey continuity and smoothness constraints, by the time the message is spoken the sounds become more like the phonetic string /dI jə it jɛt?/. The final d in dId is dropped, the word *you* becomes converted to a word that sounds a lot like "juh," and finally the word *yet* is pronounced as "jet." Remarkably, through the speech-perception process, human beings are usually able to decode this highly stylized version of the text into the correct string; sadly, however, this remains a most difficult task for almost all speech-recognition machines.

2.2 THE SPEECH-PRODUCTION PROCESS

Figure 2.3 shows a mid-sagittal plane (longitudinal cross-section) X-ray of the human vocal apparatus [2]. The *vocal tract*, outlined by the dotted lines in Figure 2.3, begins at the opening of the vocal cords, or *glottis*, and ends at the lips. The vocal tract consists of the *pharynx* (the connection from the esophagus to the mouth) and the mouth, or *oral cavity*. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract, determined by the positions of the tongue, lips, jaw, and velum, varies from zero (complete closure) to about 20 cm^2. The *nasal tract* begins at the velum and ends at the nostrils. When the *velum*, (a trapdoor-like mechanism at the back of the mouth cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

A schematic diagram of the human vocal mechanism is shown in Figure 2.4 [3]. Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the *trachea* (or windpipe), the tensed vocal cords within the *larynx* are caused to vibrate (in the mode of a relaxation oscillator) by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the *pharynx* (the throat cavity), the mouth cavity, and possibly the nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, mouth), different sounds are produced.

Figure 2.5 shows plots of the glottal air flow (volume velocity waveform) and the resulting sound pressure at the mouth for a typical vowel sound [4]. The glottal waveform shows a gradual build-up to a quasi-periodic pulse train of air, taking about 15 msec to reach steady state. This build-up is also reflected in the acoustic waveform shown at the bottom of the figure.

A simplified representation of the complete physiological mechanism for creating speech is shown in Figure 2.6 [3]. The lungs and the associated muscles act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the bronchi and trachea. When the vocal cords are tensed, the air flow causes them to vibrate, producing

Ex.1022 / Page 45 of 333
AMAZON.COM, INC.

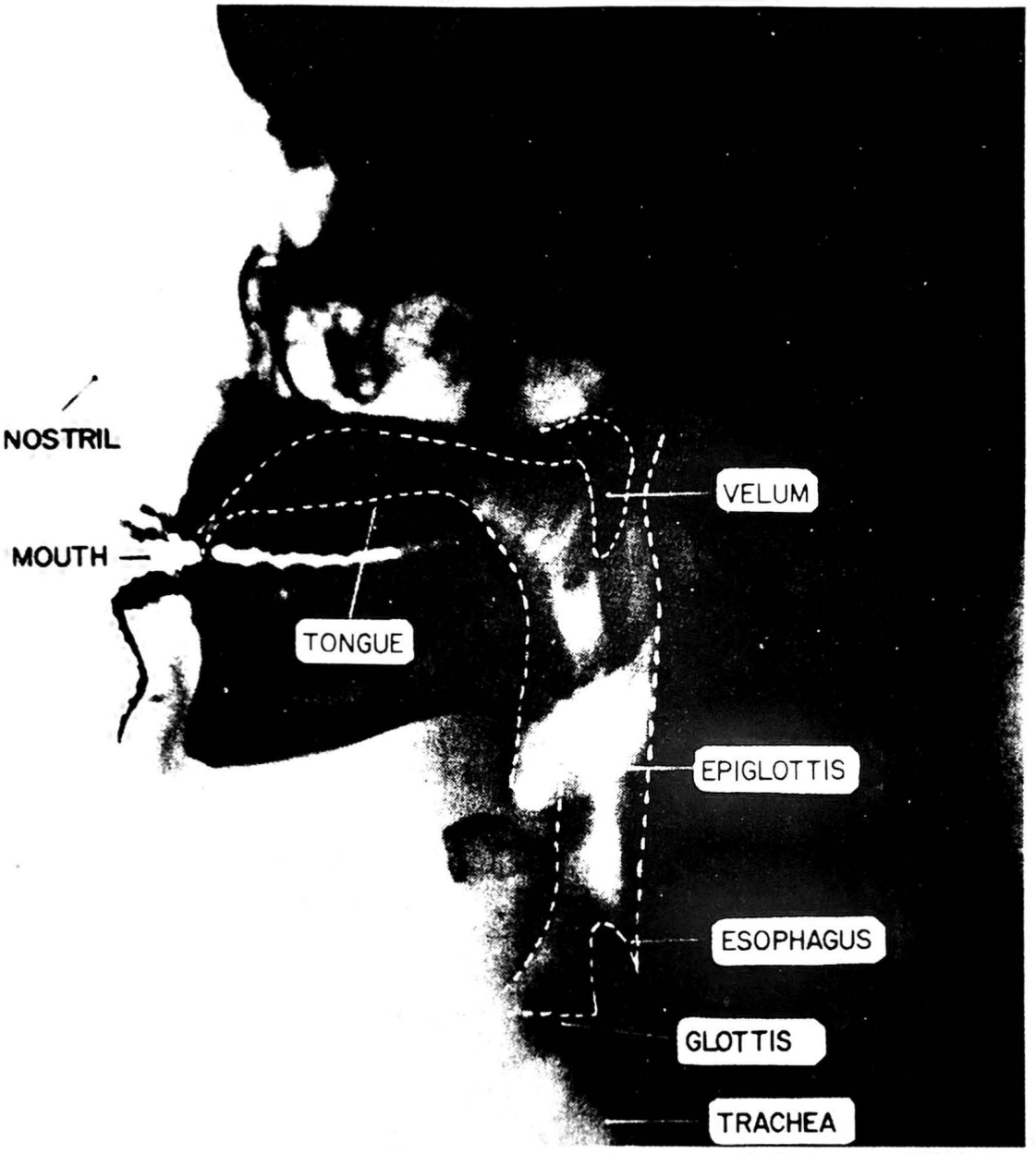


Figure 2.3 Mid-sagittal plane X-ray of the human vocal apparatus (after Flanagan et al. [2]).

so-called voiced speech sounds. When the vocal cords are relaxed, in order to produce a sound, the air flow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sounds, or it can build up pressure behind a point of total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly released, causing a brief transient sound.

Speech is produced as a sequence of sounds. Hence the state of the vocal cords, as well as the positions, shapes, and sizes of the various articulators, changes over time to reflect the sound being produced. The manner in which different sounds are created will be described later in this chapter. First we divert to a brief discussion of the speech waveform

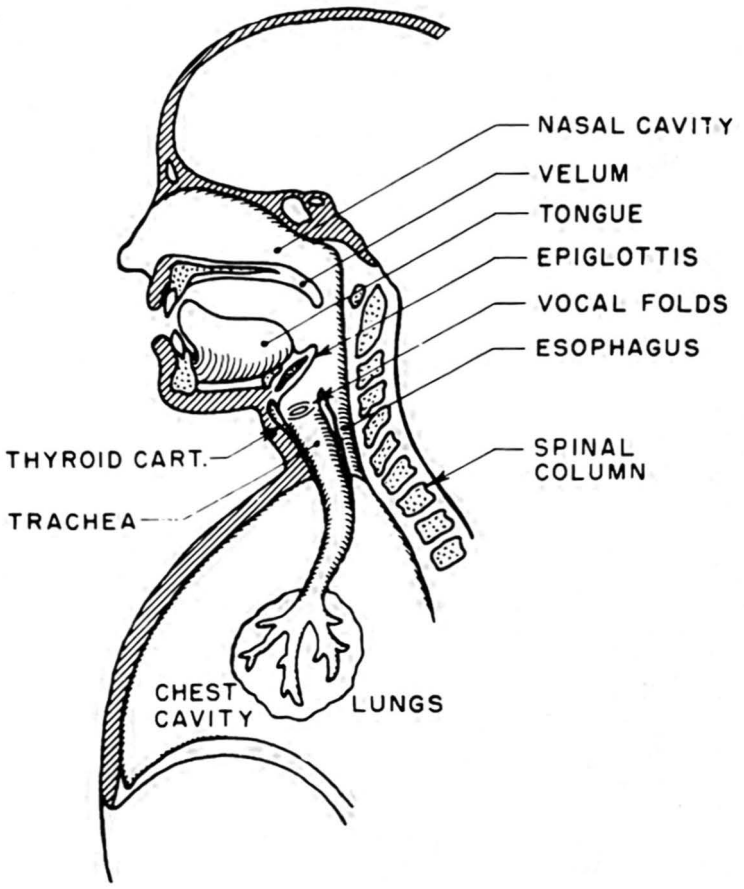


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).

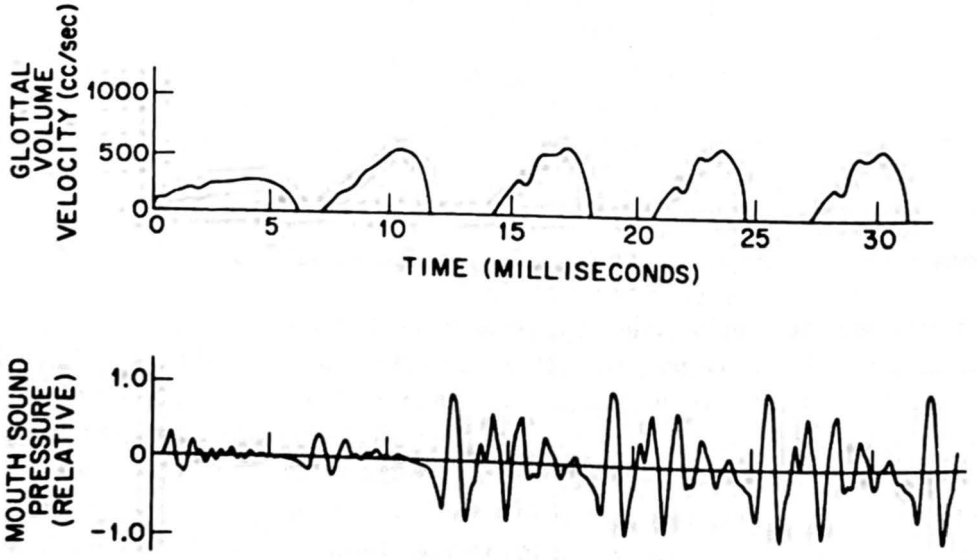


Figure 2.5 Glottal volume velocity and resulting sound pressure at the start of a voiced sound (after Ishizaka and Flanagan [4]).

Ex.1022 / Page 47 of 333
AMAZON.COM, INC.

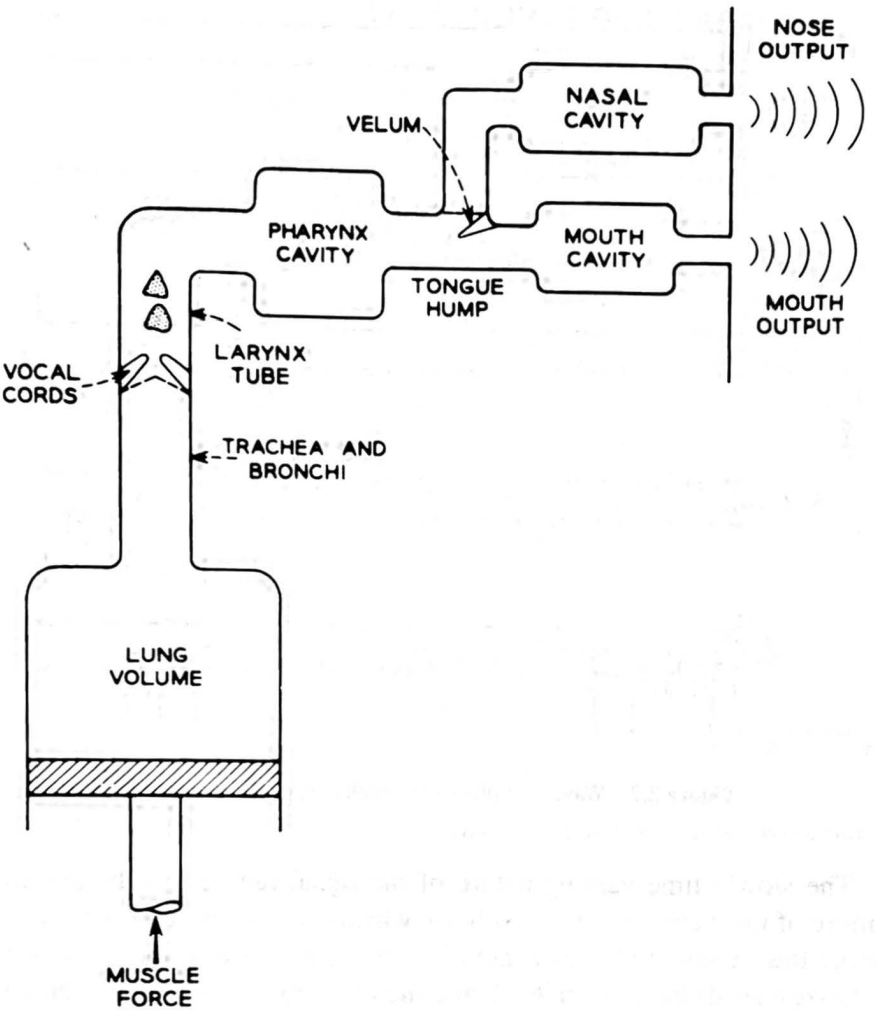


Figure 2.6 Schematic representation of the complete physiological mechanism of speech production (after Flanagan [3]).

and its spectral representation.

2.3 REPRESENTING SPEECH IN THE TIME AND FREQUENCY DOMAINS

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. An illustration of this effect is given in Figure 2.7, which shows the time waveform corresponding to the initial sounds in the phrase, "It's time . . ." as spoken by a male speaker. Each line of the waveform corresponds to 100 msec (1/10 second) of signal; hence the entire plot encompasses about 0.5 sec.

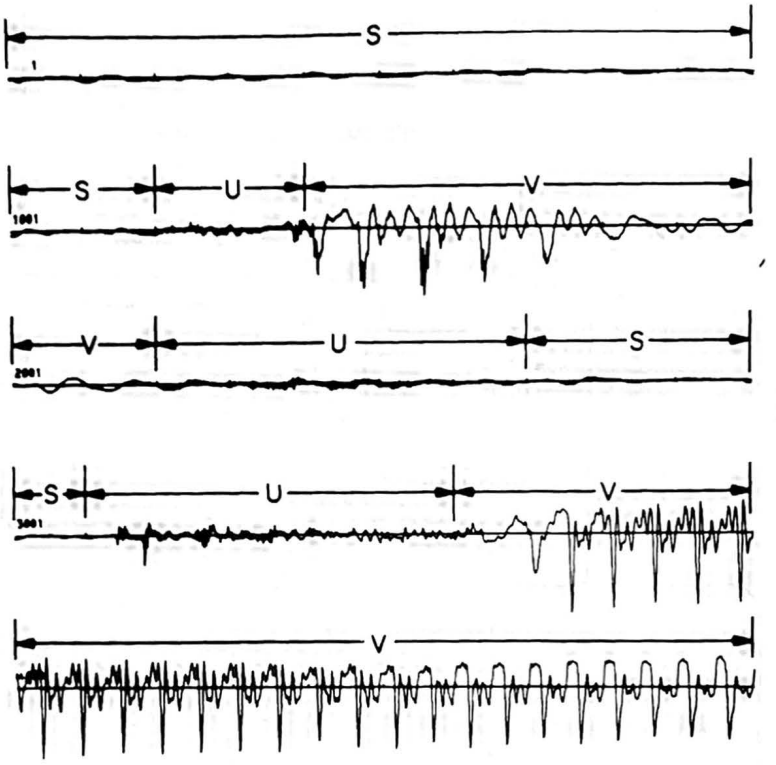


Figure 2.7 Waveform plot of the beginning of the utterance "It's time."

The slowly time varying nature of the signal can be seen by contrasting the first 100 msec of the waveform (the first line), which corresponds to background silence and is therefore low in amplitude, to the next 100 msec of the waveform (the second line), which first shows a small increase in level, and then a sharp increase in level and a gross change in waveform shape and regularity (it becomes almost periodic).

There are several ways of classifying (labeling) events in speech. Perhaps the simplest and most straightforward is via the state of the speech-production source—the vocal cords. It is accepted convention to use a three-state representation in which the states are (1) silence (S), where no speech is produced; (2) unvoiced (U), in which the vocal cords are not vibrating, so the resulting speech waveform is aperiodic or random in nature; and (3) voiced (V), in which the vocal cords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting speech waveform is quasi-periodic. The result of applying this type of classification to the waveform of Figure 2.7 is shown in the figure. Initially, before speaking begins, the waveform is classified as silence (S). A brief period of unvoiced (U) sound (whisper or aspiration) is seen prior to the voicing (V) corresponding to the initial vowel in the word *It's*. Following the voicing region, there is a brief, unvoiced aspiration (devoicing of the vowel), followed by a silence region (prior to the /t/ in *It's*), and then a relatively long, unvoiced (U) region corresponding to the /t/ release, followed by the /s/, followed by the /t/ in *time*. Finally there is a long voicing (V) region corresponding to the diphthong /a^y/ in *time*.

It should be clear that the segmentation of the waveform into well-defined regions of

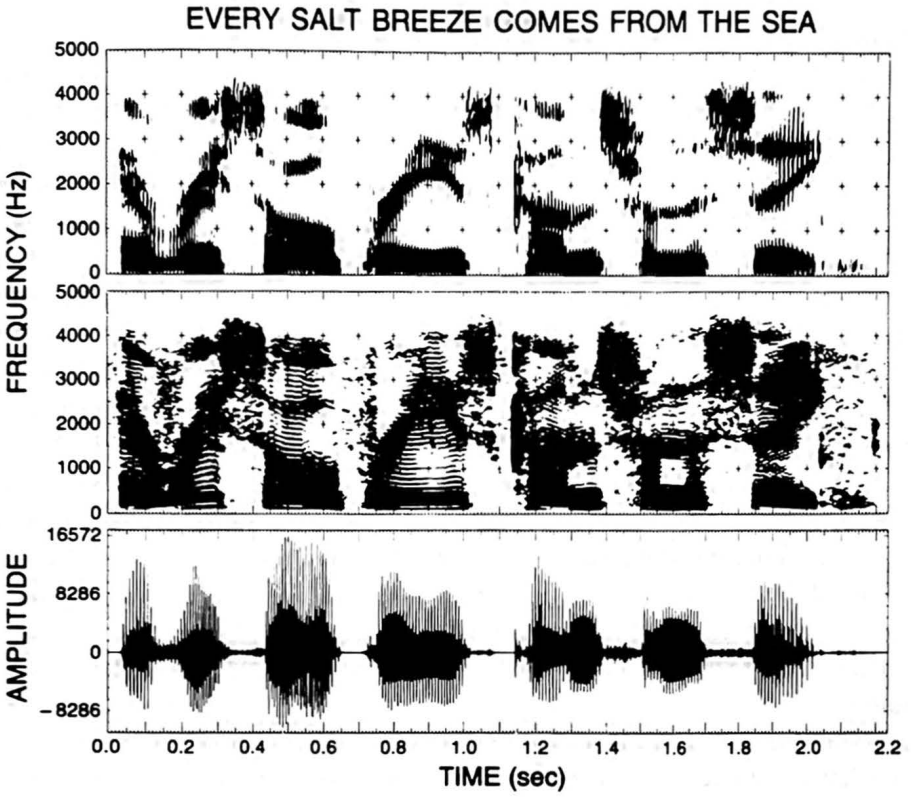


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea."

silence, unvoiced, and voiced signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or a weak, voiced sound (like /v/ or /m/) from unvoiced sounds or even silence. However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications.

An alternative way of characterizing the speech signal and representing the information associated with the sounds is via a spectral representation. Perhaps the most popular representation of this type is the sound spectrogram in which a three-dimensional representation of the speech intensity, in different frequency bands, over time is portrayed. An example of this type of speech representation is given in Figure 2.8, which shows a *wideband spectrogram* in the first panel, a *narrowband spectrogram* in the second panel, and a waveform amplitude plot in the third panel, of a spoken version of the utterance "Every salt breeze comes from the sea" by a male speaker. The wideband spectrogram corresponds to performing a spectral analysis on 15-msec sections of waveform using a broad analysis filter (125 Hz bandwidth) with the analysis advancing in intervals of 1 msec. The spectral intensity at each point in time is indicated by the intensity (darkness) of the plot at a particular analysis frequency. Because of the relatively broad bandwidth of the analysis filters, hence the relatively short duration of the analysis window, the spectral envelope of individual periods of the speech waveform during voiced sections are resolved

and are seen as vertical striations in the spectrogram.

The narrowband spectrogram (shown in the second panel of Figure 2.8) corresponds to performing a spectral analysis on 50-msec sections of waveform using a narrow analysis filter (40 Hz bandwidth), with the analysis again advancing in intervals of 1 msec. Because of the relatively narrow bandwidth of the analysis filters, individual spectral harmonics corresponding to the pitch of the speech waveform, during voiced regions, are resolved and are seen as almost-horizontal lines in the spectrogram. During periods of unvoiced speech, we see primarily high-frequency energy in the spectrograms; during periods of silence we essentially see no spectral activity (because of the reduced signal level).

A third way of representing the time-varying signal characteristics of speech is via a parameterization of the spectral activity based on the model of speech production. Because the human vocal tract is essentially a tube, or concatenation of tubes, of varying cross-sectional area that is excited either at one end (by the vocal cord puffs of air) or at a point along the tube (corresponding to turbulent air at a constriction), acoustic theory tells us that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube. Such resonances are called *formants* for speech, and they represent the frequencies that pass the most acoustic energy from the source to the output. Typically there are about three resonances of significance, for a human vocal tract, below about 3500 Hz. Figure 2.9 [5] shows a wideband spectrogram, along with the computed formant frequency estimates, for the utterance "Why do I owe you a letter," spoken by a male speaker. There is a good correspondence between the estimated formant frequencies and the points of high spectral energy in the spectrogram. The formant frequency representation is a highly efficient, compact representation of the time-varying characteristics of speech. The major problem, however, is the difficulty of reliably estimating the formant frequencies for low-level voiced sounds, and the difficulty of defining the formants for unvoiced or silence regions. As such, this representation is more of theoretical than of practical interest.

Figures 2.10 and 2.11 show spectral and temporal representations of the phrase "Should we chase," spoken by a male speaker, along with a detailed segmentation of the waveform into individual sounds. The ultimate goal of speech recognition is to uniquely and automatically provide such a segmentation and labeling of speech into constituent sounds or sound groups such as words, then sentences. To understand the limitations on this approach, we will next discuss, in detail, the general sounds of English and the relevant acoustic and phonetic features of the sounds.

2.4 SPEECH SOUNDS AND FEATURES

The number of linguistically distinct speech sounds (phonemes) in a language is often a matter of judgment and is not invariant to different linguists. Table 2.1 shows a condensed list of phonetic symbols of American English, their ARPABET representation [6], and an example word in which the sound occurs. Shown in this table are 48 sounds, including 18 vowels or vowel combinations (called diphthongs), 4 vowel-like consonants, 21 standard consonants, 4 syllabic sounds, and a phoneme referred to as a glottal stop (literally a symbol

Ex.1022 / Page 51 of 333
AMAZON.COM, INC.

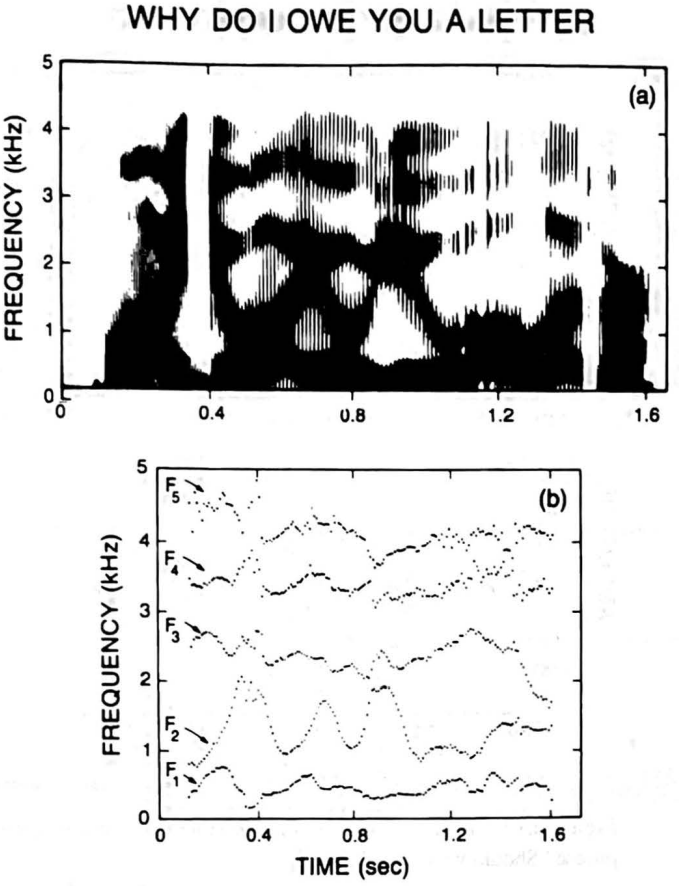


Figure 2.9 Wideband spectrogram and formant frequency representation of the utterance "Why do I owe you a letter" (after Atal and Hanauer [5]).

for a sound corresponding to a break in voicing within a sound).

Many of the sounds or phonemes shown in Table 2.1 are not considered standard; they represent specialized cases such as the so-called barred I (/ɨ/) in the word *roses*. As such, a more standard representation of the basic sounds and sound classes of American English is shown in Figure 2.12. Here we see the conventional set of 11 vowels, classified as front, mid, or back, corresponding to the position of the tongue hump in producing the vowel; 4 vowel combinations or diphthongs; the 4 semivowels broken down into 2 liquids and 2 glides; the nasal consonants, the voiced and unvoiced stop consonants; the voiced and unvoiced fricatives; whisper; and the affricates. There are a total of 39 of the 48 sounds of Table 2.1 represented in Figure 2.12.

2.4.1 The Vowels

The vowel sounds are perhaps the most interesting class of sounds in English. Their importance to the classification and representation of written text is very low; however, most practical speech-recognition systems rely heavily on vowel recognition to achieve

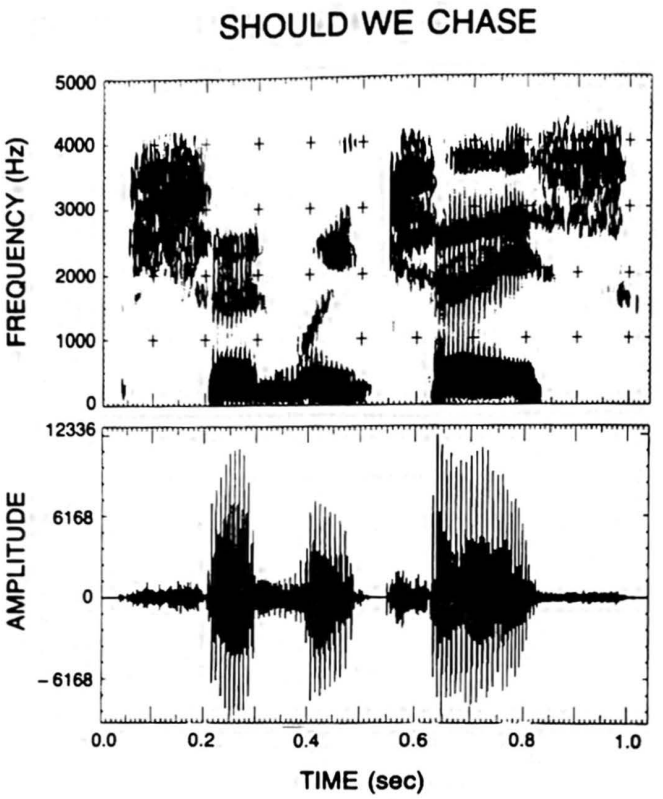


Figure 2.10 Wideband spectrogram and intensity contour of the phrase "Should we chase."

high performance. To partially illustrate this point, consider the following sections of text:

Section I

Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nts i_ th_ c_mp_ny's _m_g_, s_p_rv_s__n, th__r w_rk_ng c_nd_t__ns, b_n_f_ts _nd _pp_rt_n_t__s f_r gr_wth.

Section II

A__i_u_e_ _o_a___a___a_e_ e__e__ia___ __e _a_e, _i__ __e __o_e_ o_ o__u_a_io_a_ e___o_ee_ __i_____ _e__ea_i__.

In Section I we have omitted the conventional vowel letters (a,e,i,o,u); however, with a little effort the average reader can "fill in" the missing vowels and decode the section so that it reads

They noted significant improvements in the company's image, supervision, their working conditions, benefits and opportunities for growth.

In Section II we have omitted the conventional consonant letters; the resulting text is essentially not decodable. The actual text is

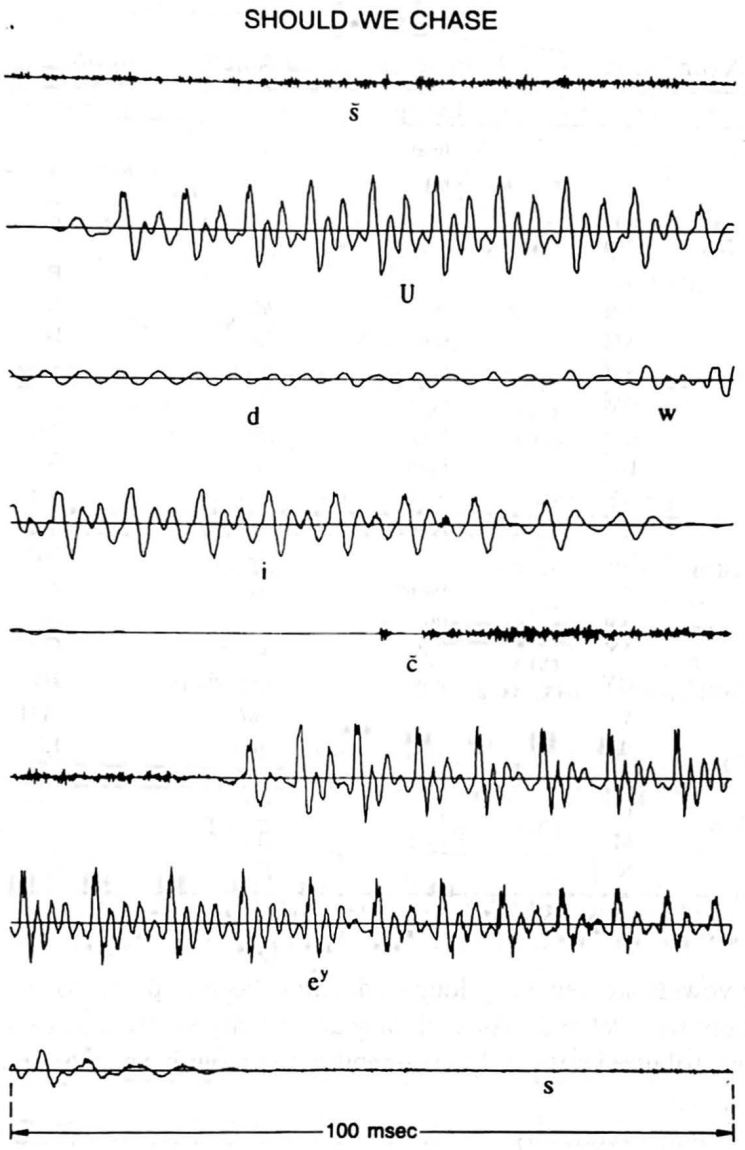


Figure 2.11 The speech waveform and a segmentation and labeling of the constituent sounds of the phrase "Should we chase."

Attitudes toward pay stayed essentially the same, with the scores of occupational employees slightly decreasing.

In speaking, vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. The way in which the cross-sectional area varies along the vocal tract determines the resonance frequencies of the tract (the formants) and thereby the sound that is produced. The vowel sound produced is determined primarily by the position of the tongue, but the positions of the jaw, lips, and to a small extent, the velum, also influence the resulting sound.

TABLE 2.1. A condensed list of phonetic symbols for American English.

Phoneme	ARPABET	Example	Phoneme	ARPABET	Example
/i/	IY	beat	/ŋ/	NX	sing
/ɪ/	IH	bit	/p/	P	pet
/e/ (e^y)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/a/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	H	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (o^w)	OW	boat	/f/	F	fat
/ʊ/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/š/ (sh)	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/a^w/	AW	down	/ž/ (zh)	ZH	azure
/a^y/	AY	buy	/č/ (tsh)	CH	church
/$ɔ^y$/	OY	boy	/ǰ/ (dzh, j)	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/l̩/	EL	battle
/ɾ/	R	rent	/m̩/	EM	bottom
/l/	L	let	/n̩/	EN	button
/m/	M	met	/ɾ̃/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

The vowels are generally long in duration (as compared to consonant sounds) and are spectrally well defined. As such they are usually easily and reliably recognized and therefore contribute significantly to our ability to recognize speech, both by human beings and by machine.

There are several ways to characterize and classify vowels, including the typical articulatory configuration required to produce the sounds, typical waveform plots, and typical spectrogram plots. Figures 2.13–2.15 show typical articulatory configurations of the vowels (2.13), examples of vowel waveforms (2.14), and examples of vowel spectrograms (2.15). A convenient and simplified way of classifying vowel articulatory configurations is in terms of the tongue hump position (i.e., front, mid, back), and tongue hump height (high, mid, low), where the tongue hump is the mass of the tongue at its narrowest constriction within the vocal tract. According to this classification the vowels /i/, /ɪ/, /æ/, and /ɛ/ are front vowels, (with different tongue heights) /a/, /ʌ/, and /ɔ/ are mid vowels, and /ʊ/, /u/, and /o/ are back vowels (see also Figure 2.12).

As shown in the acoustic waveform plots of the vowels, in Figure 2.14, the front vowels show a pronounced, high-frequency resonance, the mid vowels show a balance of energy over a broad frequency range, and the back vowels show a predominance of low-frequency spectral information. This behavior is evidenced in the vowel spectrogram

Ex.1022 / Page 55 of 333
AMAZON.COM, INC.

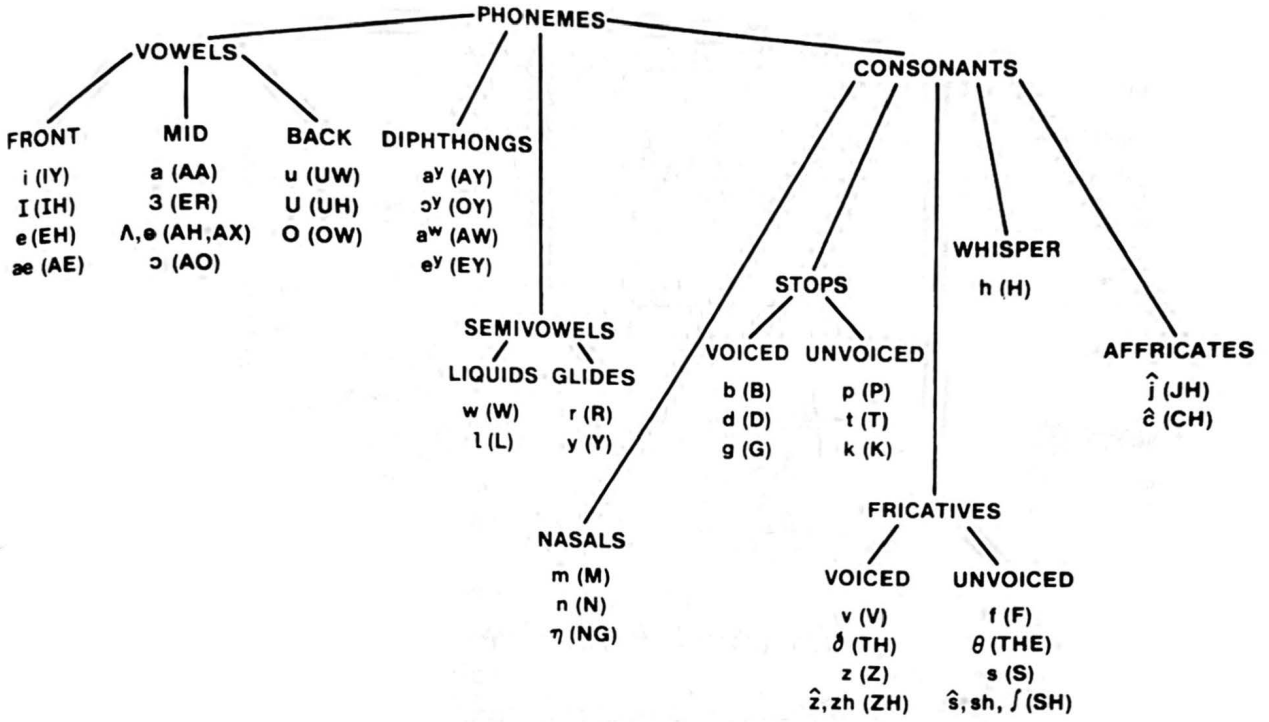


Figure 2.12 Chart of the classification of the standard phonemes of American English into broad sound classes.

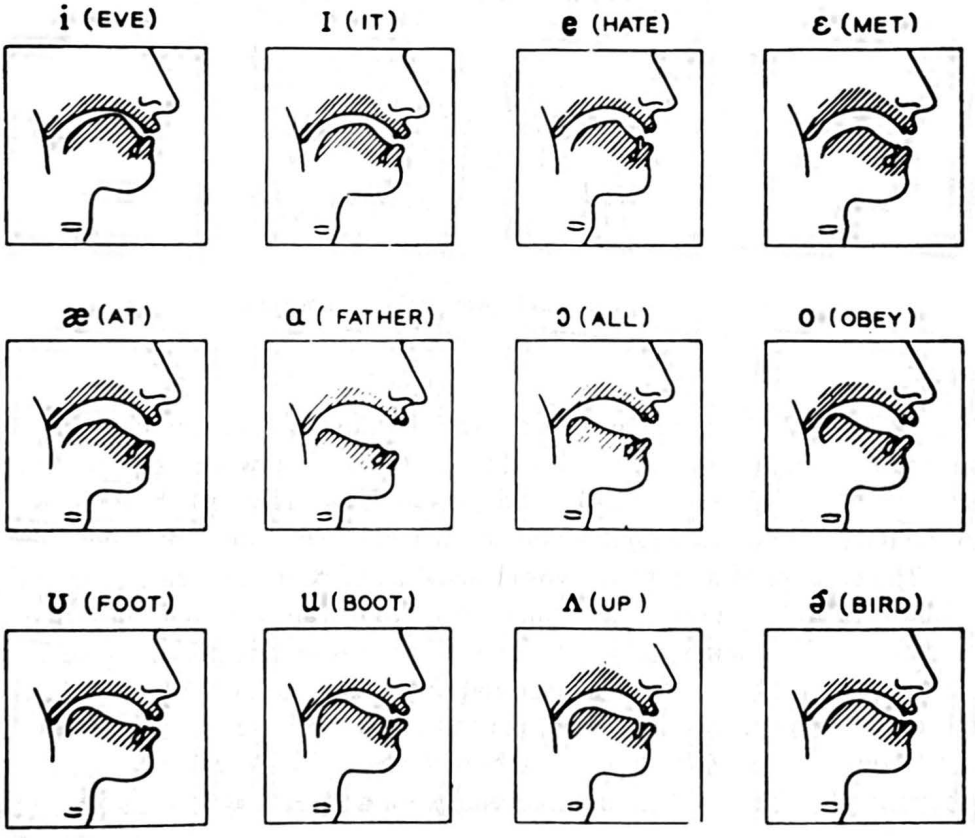


Figure 2.13 Articulatory configurations for typical vowel sounds (after Flanagan [3]).

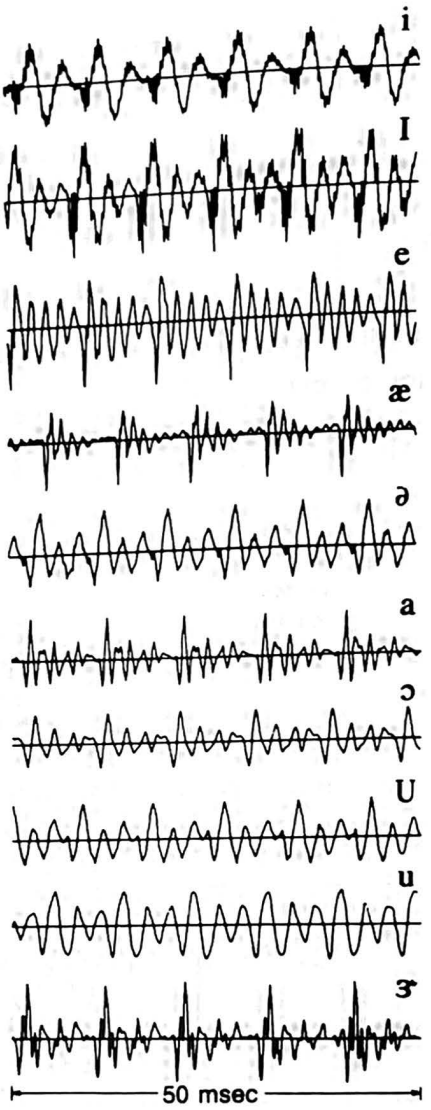


Figure 2.14 Acoustic waveform plots of typical vowel sounds.

plots of Figure 2.15, in which the front vowels show a relatively high second and third formant frequency (resonance), whereas the mid vowels show well-separated and balanced locations of the formants, and the back vowels (especially /u/) show almost no energy beyond the low-frequency region with low first and second formant frequencies.

The concept of a "typical" vowel sound is, of course, unreasonable in light of the variability of vowel pronunciation among men, women and children with different regional accents and other variable characteristics. To illustrate this point, Figure 2.16 shows a classic plot, made by Gordon Peterson and Harold Barney, of measured values of the first and second formant for 10 vowels spoken by a wide range of male and female talkers who attended the 1939 World's Fair in New York City [7].A wide range of variability can be seen in the measured formant frequencies for a given vowel sound, and also there is

Ex.1022 / Page 57 of 333
AMAZON.COM, INC.

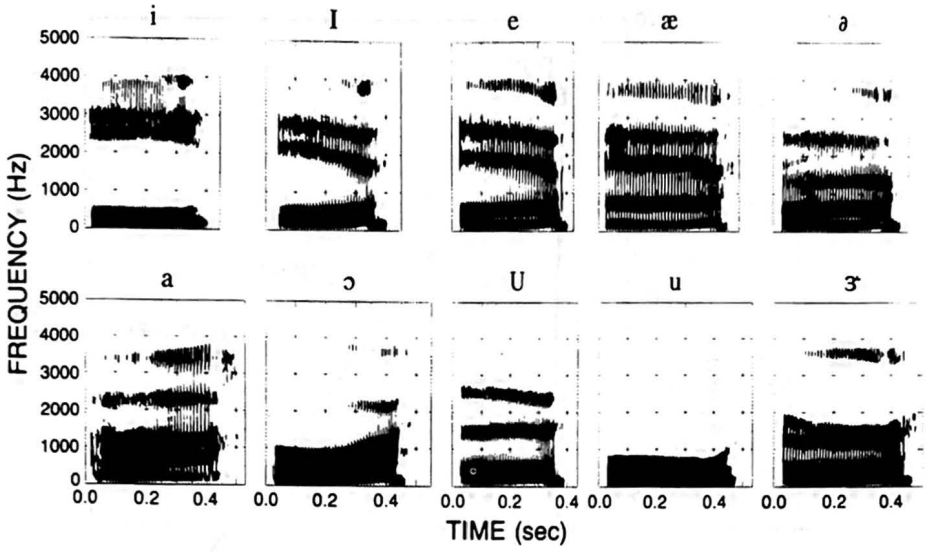


Figure 2.15 Spectrograms of the vowel sounds.

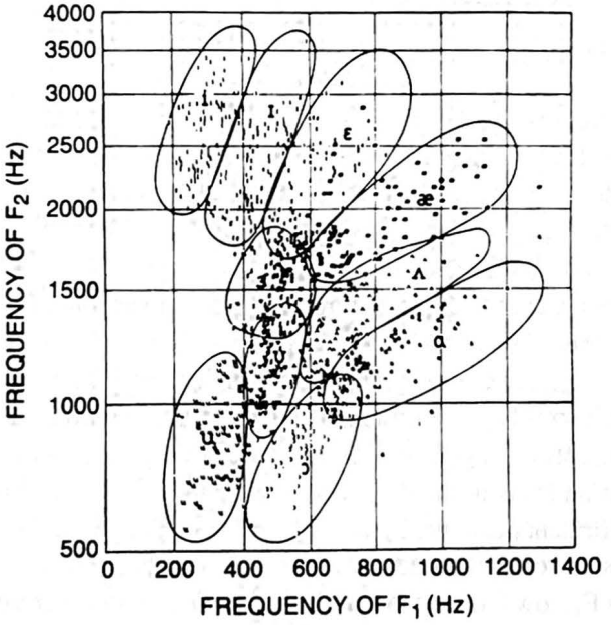


Figure 2.16 Measured frequencies of first and second formants for a wide range of talkers for several vowels (after Peterson & Barney [7]).

overlap between the formant frequencies for *different* vowel sounds by different talkers. The ellipses drawn in this figure represent gross characterizations of the regions in which most of the tokens of the different vowels lie. The message of Figure 2.16, for speech recognition by machine, is fairly clear; that is, it is not just a simple matter of measuring formant frequencies or spectral peaks accurately to accurately classify vowel sounds; one

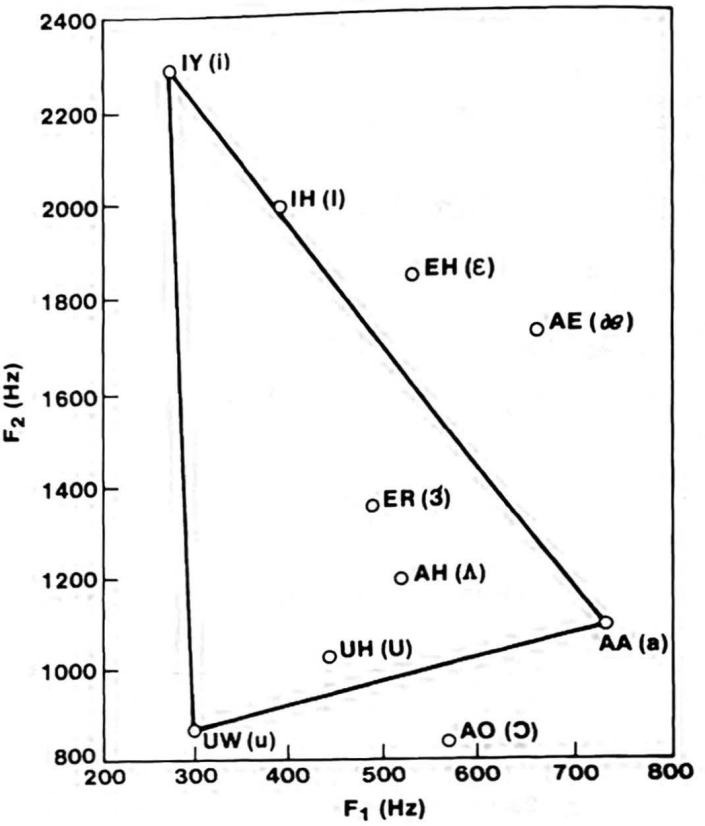


Figure 2.17 The vowel triangle with centroid positions of the common vowels.

must do some type of talker (accent) normalization to account for the variability in formants and overlap between vowels.

A common way of exploiting the information embodied in Figures 2.15 and 2.16 is to represent each vowel by a centroid in the formant space with the realization that the centroid, at best, represents average behavior and does not represent variability across talkers. Such a representation leads to the classic vowel triangle shown in Figure 2.17 and represented in terms of formant positions by the data given in Table 2.2. The vowel triangle represents the extremes of formant locations in the F_1-F_2 plane, as represented by /i/ (low F_1, high F_2), /u/ (low F_1, low F_2), and /a/ (high F_1, low F_2), with other vowels appropriately placed with respect to the triangle vertices. The utility of the formant frequencies of Table 2.2 has been demonstrated in text-to-speech synthesis in which high-quality vowel sounds have been synthesized using these positions for the resonances [8].

2.4.2 Diphthongs

Although there is some ambiguity and disagreement as to what is and what is not a diphthong, a reasonable definition is that a diphthong is a gliding monosyllabic speech

TABLE 2.2. Formant frequencies for typical vowels.

ARPABET Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	/i/	beet	270	2290	3010
IH	/ɪ/	bit	390	1990	2550
EH	/ɛ/	bet	530	1840	2480
AE	/æ/	bat	660	1720	2410
AH	/ʌ/	but	520	1190	2390
AA	/a/	hot	730	1090	2440
AO	/ɔ/	bought	570	840	2410
UH	/U/	foot	440	1020	2240
UW	/u/	boot	300	870	2240
ER	/ɝ/	bird	490	1350	1690

sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another. According to this definition, there are six diphthongs in American English, namely /a^y/ (as in buy), /a^w/ (as in down), /e^y/ (as in bait), and /ɔ^y/ (as in boy), /o/ (as in boat), and /ju/ (as in you).

The diphthongs are produced by varying the vocal tract smoothly between vowel configurations appropriate to the diphthong. Figure 2.18 shows spectrogram plots of four of the diphthongs spoken by a male talker. The gliding motions of the formants are especially prominent for the sounds /a^y/, /a^w/ and /ɔ^y/ and are somewhat weaker for /e^y/ because of the closeness (in vowel space) of the two vowel sounds comprising this diphthong.

An alternative way of displaying the time-varying spectral characteristics of diphthongs is via a plot of the values of the second formant versus the first formant (implicitly as a function of time) as shown in Figure 2.19 [9]. The arrows in this figure indicate the direction of motion of the formants (in the ($F_1 - F_2$) plane) as time increases. The dashed circles in this figure indicate average positions of the vowels. Based on these data, and other measurements, the diphthongs can be characterized by a time-varying vocal tract area function that varies between two vowel configurations.

2.4.3 Semivowels

The group of sounds consisting of /w/, /l/, /r/, and /y/ is quite difficult to characterize. These sounds are called semivowels because of their vowel-like nature. They are generally characterized by a gliding transition in vocal tract area function between adjacent phonemes. Thus the acoustic characteristics of these sounds are strongly influenced by the context in which they occur. For our purposes, they are best described as transitional, vowel-like sounds, and hence are similar in nature to the vowels and diphthongs.

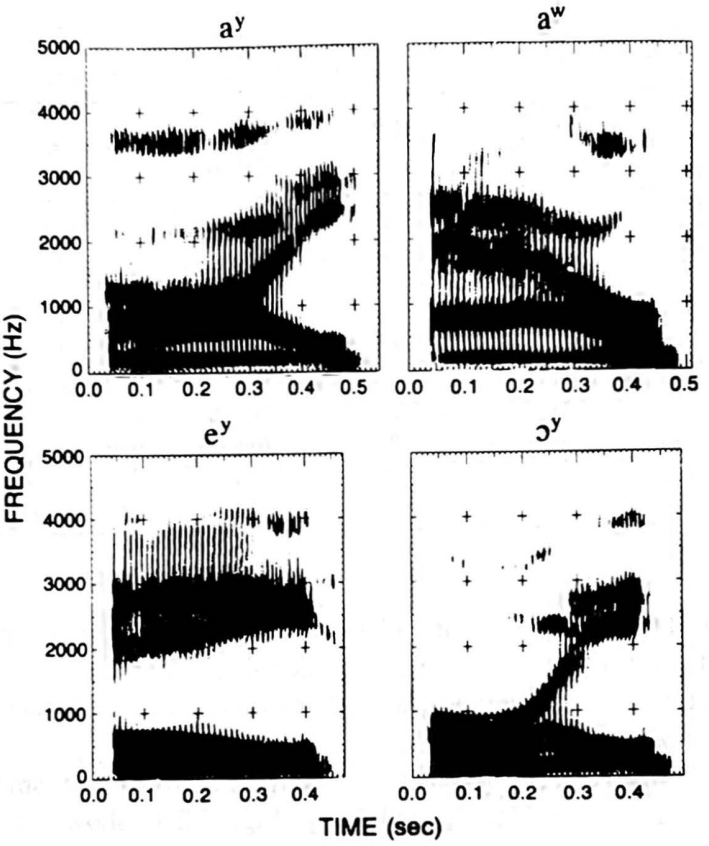


Figure 2.18 Spectrogram plots of four diphthongs.

2.4.4 Nasal Consonants

The nasal consonants /m/, /n/, and /η/ are produced with glottal excitation and the vocal tract totally constricted at some point along the oral passageway. The velum is lowered so that air flows through the nasal tract, with sound being radiated at the nostrils. The oral cavity, although constricted toward the front, is still acoustically coupled to the pharynx. Thus, the mouth serves as a resonant cavity that traps acoustic energy at certain natural frequencies. As far as the radiated sound is concerned, these resonant frequencies of the oral cavity appear as antiresonances, or zeros of the transfer function of sound transmission. Furthermore, nasal consonants and nasalized vowels (i.e., some vowels preceding or following nasal consonants) are characterized by resonances that are spectrally broader, or more highly damped, than those for vowels.

The three nasal consonants are distinguished by the place along the oral tract at which a total constriction is made. For /m/ the constriction is at the lips; for /n/ the constriction is just behind the teeth; and for /η/ the constriction is just forward of the velum itself. Figure 2.20 shows typical speech waveforms and Figure 2.21 spectrograms for two nasal consonants in the context vowel-nasal-vowel. The waveforms of /m/ and /n/ look very similar. The spectrograms show a concentration of low-frequency energy with a midrange of frequencies that contain no prominent peaks. This is because of the particular

Ex.1022 / Page 61 of 333
AMAZON.COM, INC.

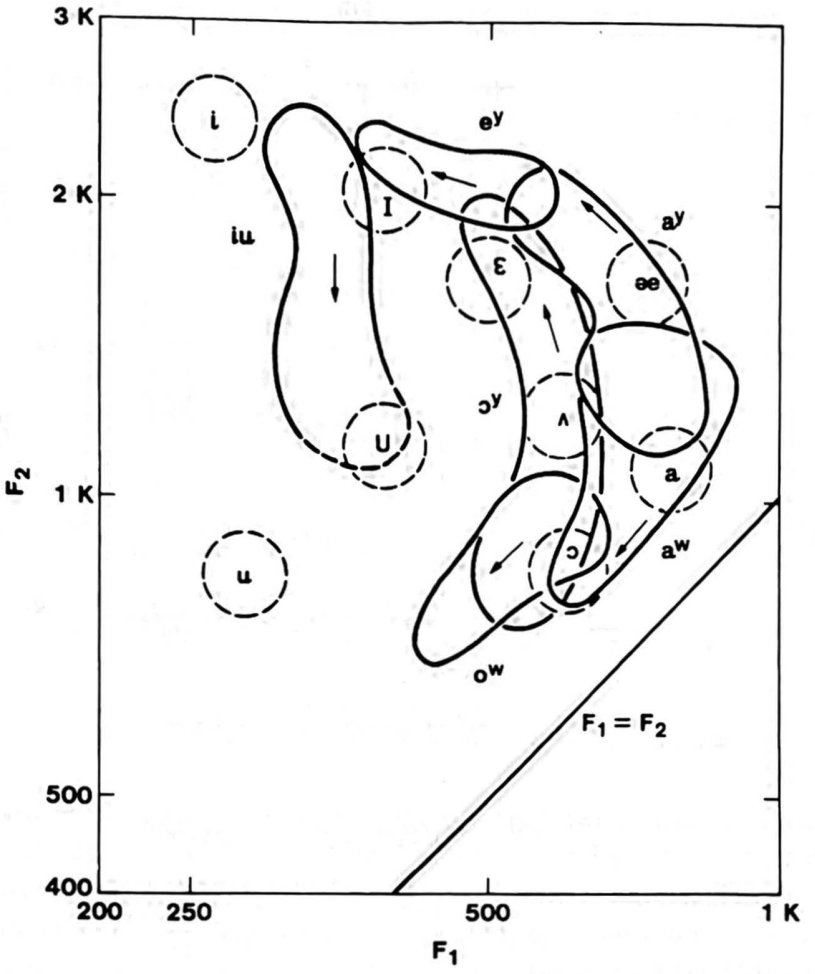


Figure 2.19 Time variation of the first two formants for the diphthongs (after Holbrook and Fairbanks [9]).

combination of resonances and antiresonances that result from the coupling of the nasal and oral tracts.

2.4.5 Unvoiced Fricatives

The unvoiced fricatives /f/, /θ/, /s/, and /sh/ are produced by exciting the vocal tract by a steady air flow, which becomes turbulent in the region of a constriction in the vocal tract. The location of the constriction serves to determine which fricative sound is produced. For the fricative /f/ the constriction is near the lips; for /θ/ it is near the teeth; for /s/ it is near the middle of the oral tract; and for /sh/ it is near the back of the oral tract. Thus the system for producing unvoiced fricatives consists of a source of noise at a constriction, which separates the vocal tract into two cavities. Sound is radiated from the lips—that is, from the front cavity. The back cavity serves, as in the case of nasals, to trap energy and thereby introduce antiresonances into the vocal output. Figure 2.22 shows the waveforms and Figure 2.23 the spectrograms of the fricatives /f/, /s/ and /sh/. The nonperiodic nature

Ex.1022 / Page 62 of 333
AMAZON.COM, INC.

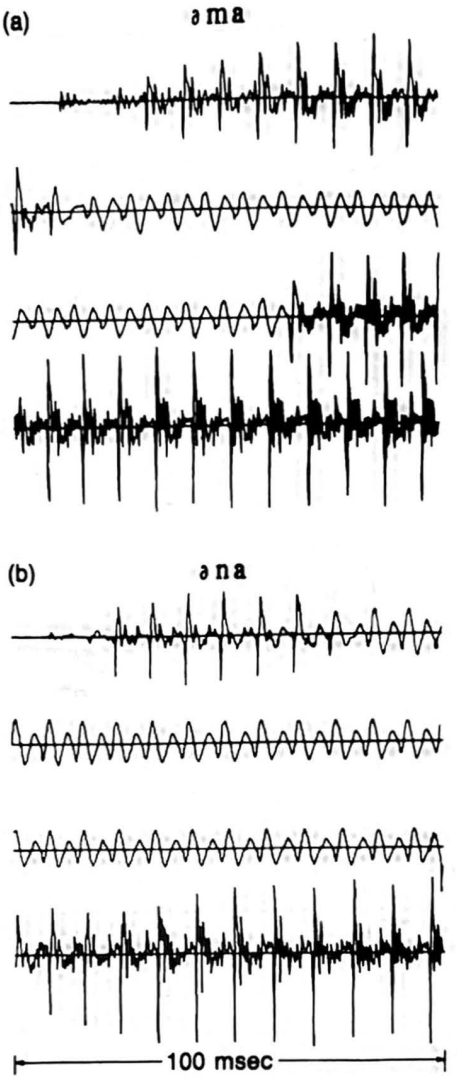


Figure 2.20 Waveforms for the sequences /ə-m-a/ and /ə-n-a/.

of fricative excitation is obvious in the waveform plots. The spectral differences among the fricatives are readily seen by comparing the three spectrograms.

2.4.6 Voiced Fricatives

The voiced fricatives /v/, /th/, /z/ and /zh/ are the counterparts of the unvoiced fricatives /f/, /θ/, /s/, and /sh/, respectively, in that the place of constriction for each of the corresponding phonemes is essentially identical. However, the voiced fricatives differ markedly from their unvoiced counterparts in that two excitation sources are involved in their production. For voiced fricatives the vocal cords are vibrating, and thus one excitation source is at the glottis. However, since the vocal tract is constricted at some point forward of the glottis, the air flow becomes turbulent in the neighborhood of the constriction. Thus the

Ex.1022 / Page 63 of 333
AMAZON.COM, INC.

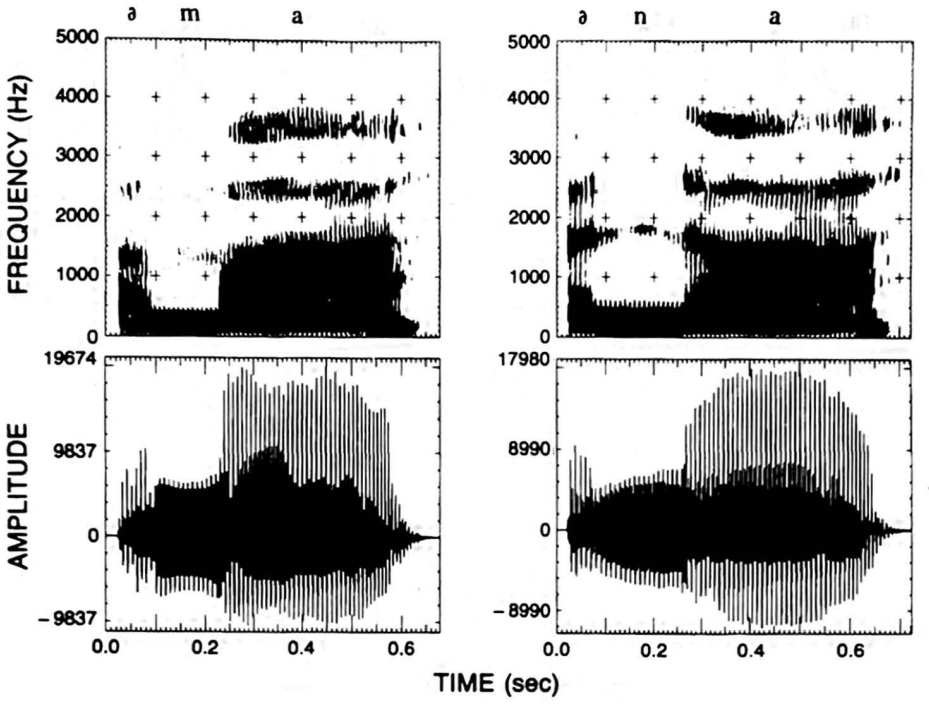


Figure 2.21 Spectrograms of the sequences /ə-m-a/ and ə-n-a/.

spectra of voiced fricatives can be expected to display two distinct components. These excitation features are readily observable in Figure 2.24, which shows typical waveforms, and in Figure 2.25, which shows spectra for two voiced fricatives. The similarity of the unvoiced fricative /f/ to the voiced fricative /v/ is easily shown in a comparison between corresponding spectrograms in Figures 2.23 and 2.25. Likewise, it is instructive to compare the spectrograms of /sh/ and /zh/.

2.4.7 Voiced and Unvoiced Stops

The voiced stop consonants /b/, /d/, and /g/, are transient, noncontinuant sounds produced by building up pressure behind a total constriction somewhere in the oral tract and then suddenly releasing the pressure. For /b/ the constriction is at the lips; for /d/ the constriction is at the back of the teeth; and for /g/ it is near the velum. During the period when there is total constriction in the tract, no sound is radiated from the lips. However, there is often a small amount of low-frequency energy radiated through the walls of the throat (sometimes called a voice bar). This occurs when the vocal cords are able to vibrate even though the vocal tract is closed at some point.

Since the stop sounds are dynamical in nature, their properties are highly influenced by the vowel that follows the stop consonant. As such, the waveforms for stop consonants give little information about the particular stop consonant. Figure 2.26 shows the waveform of the syllable /ə-b-a/. The waveform of /b/ shows few distinguishing features except for the voiced excitation and lack of high-frequency energy.

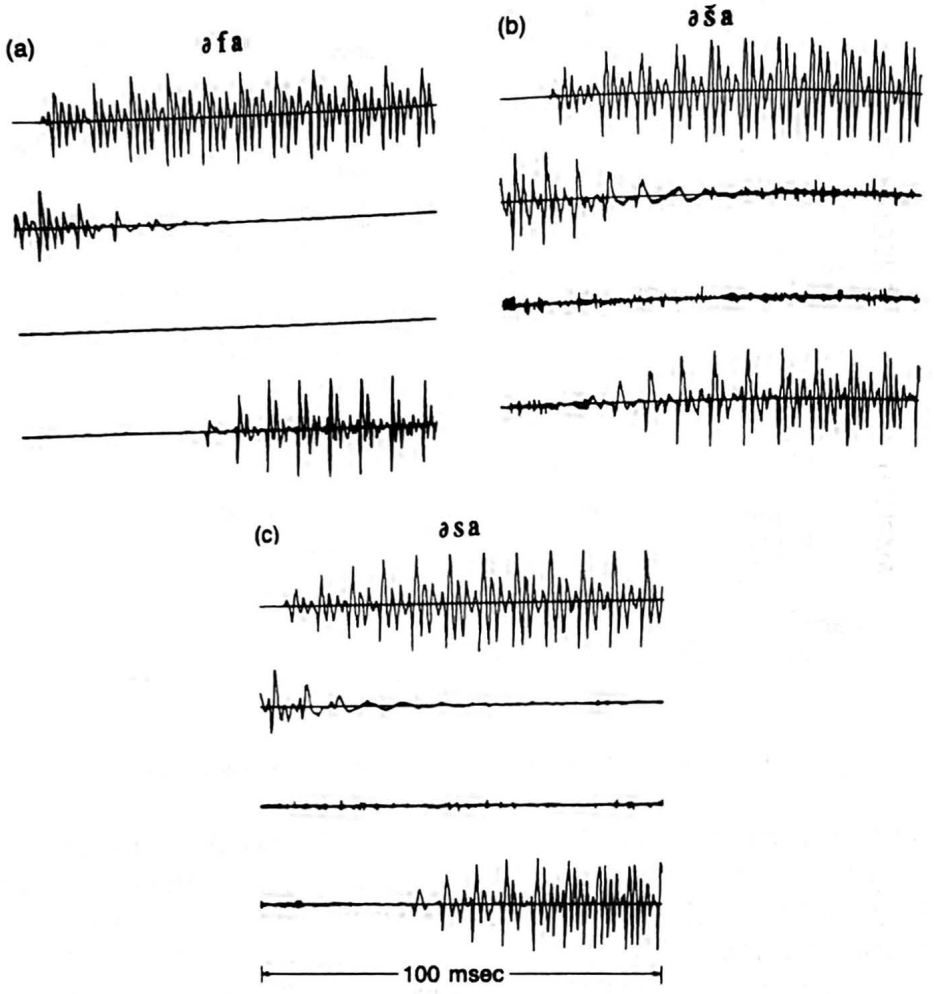


Figure 2.22 Waveforms for the sounds /f/, /s/ and /sh/ in the context /ə-x-a/ where /x/ is the unvoiced fricative.

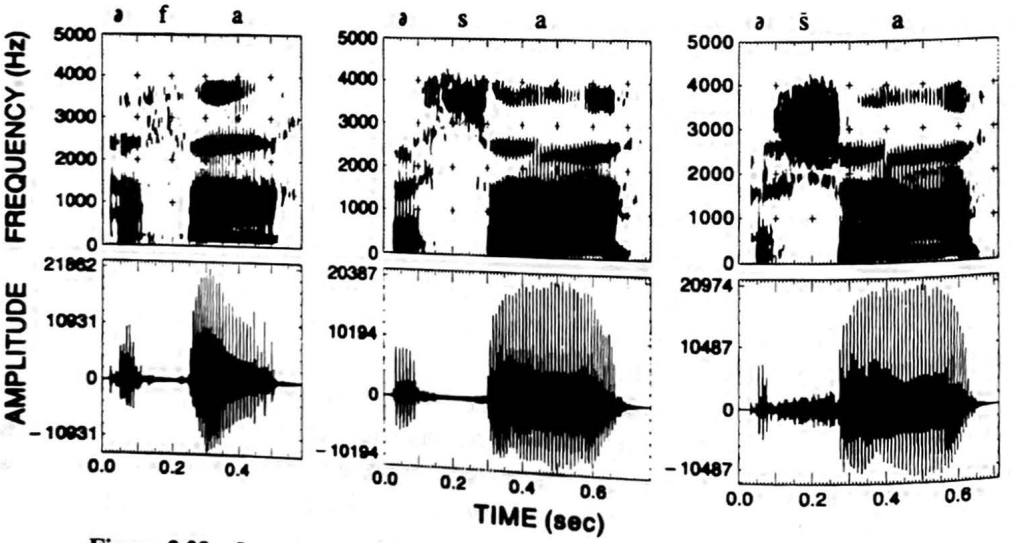


Figure 2.23 Spectrogram comparisons of the sounds /ə-f-a/, /ə-s-a/ and /ə-sh-a/.

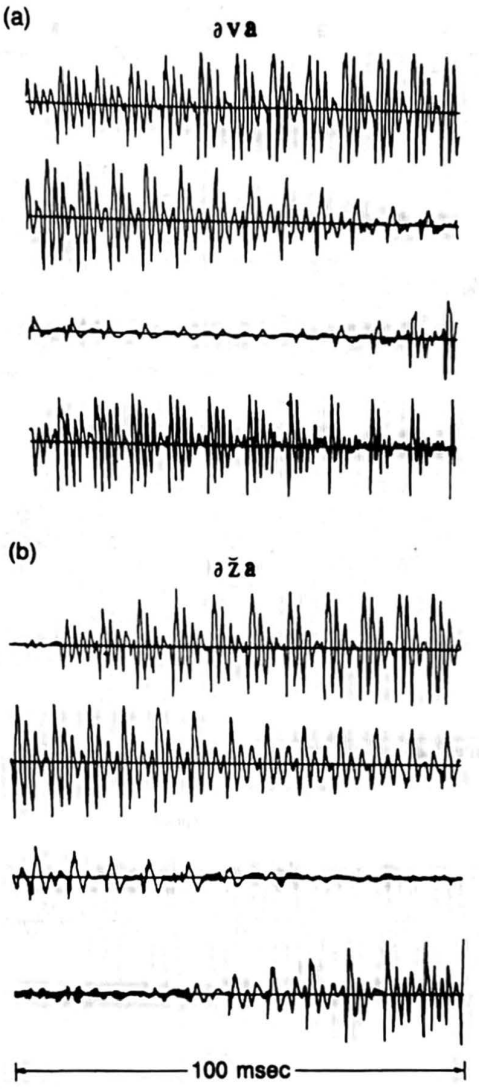


Figure 2.24 Waveforms for the sequences /ə-v-a/ and /ə-zh-a/.

The unvoiced stop consonants /p/, /t/, and /k/ are similar to their voiced counterparts /b/, /d/, and /g/, with one major exception. During the period of total closure of the tract, as the pressure builds up, the vocal cords do not vibrate. Then, following the period of closure, as the air pressure is released, there is a brief interval of friction (due to sudden turbulence of the escaping air) followed by a period of aspiration (steady air flow from the glottis exciting the resonances of the vocal tract) before voiced excitation begins.

Figure 2.27 shows waveforms and Figure 2.28 shows spectrograms of the voiced stop /b/ and the voiceless stop consonants /p/ and /t/. The "stop gap," or time interval, during which the pressure is built up is clearly in evidence. Also, it can be readily seen that the duration and frequency content of the frication noise and aspiration vary greatly with the stop consonant.

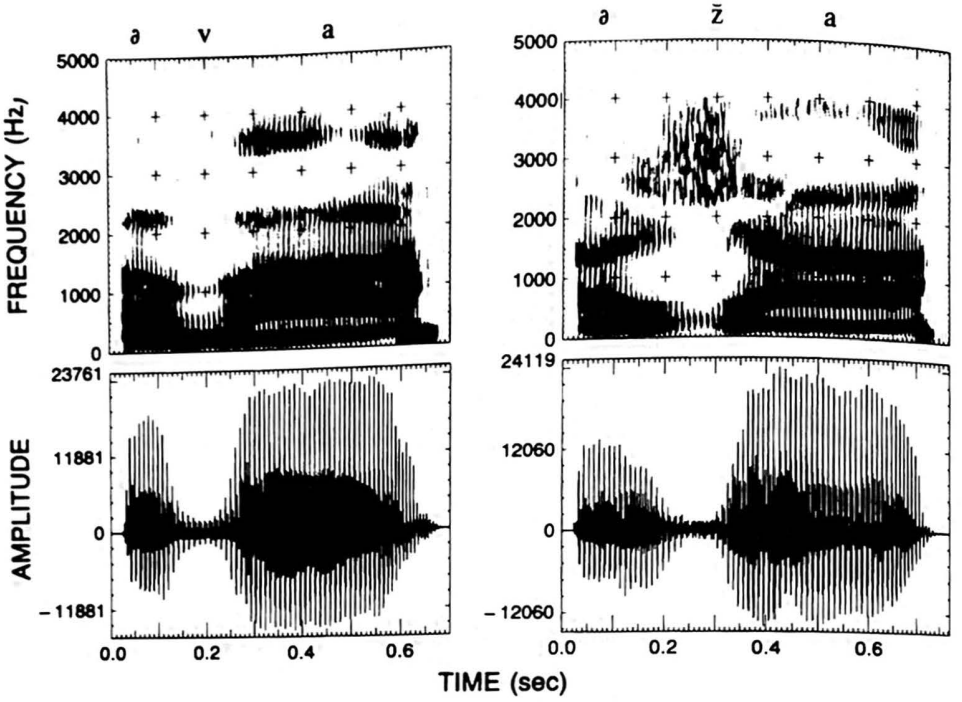


Figure 2.25 Spectrograms for the sequences /ə-v-a/ and /ə-zh-a/.

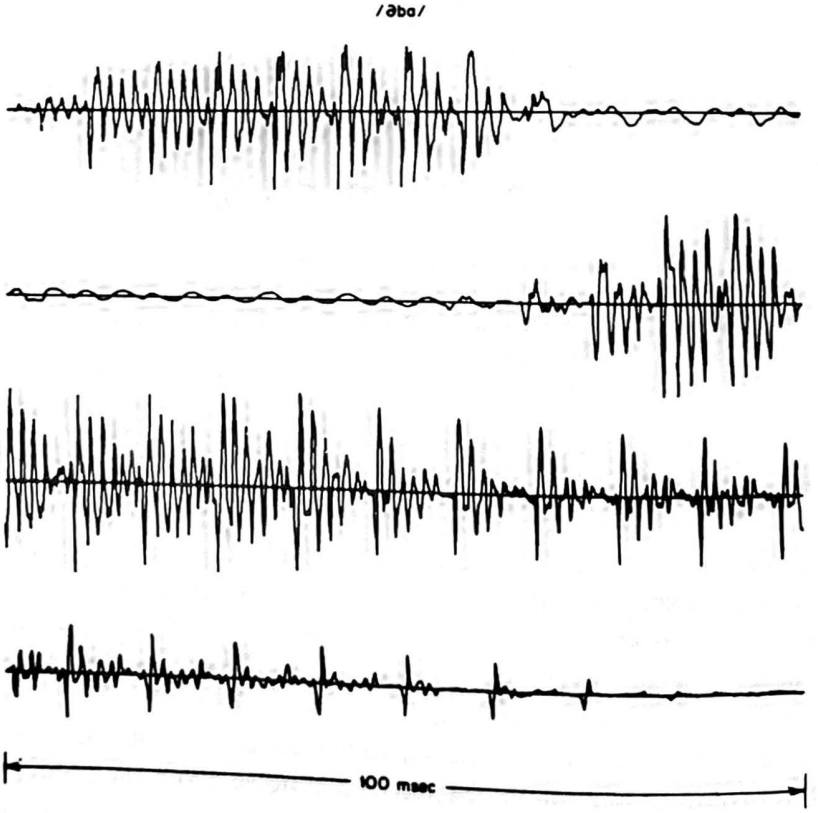


Figure 2.26 Waveform for the sequence /ə-b-a/.

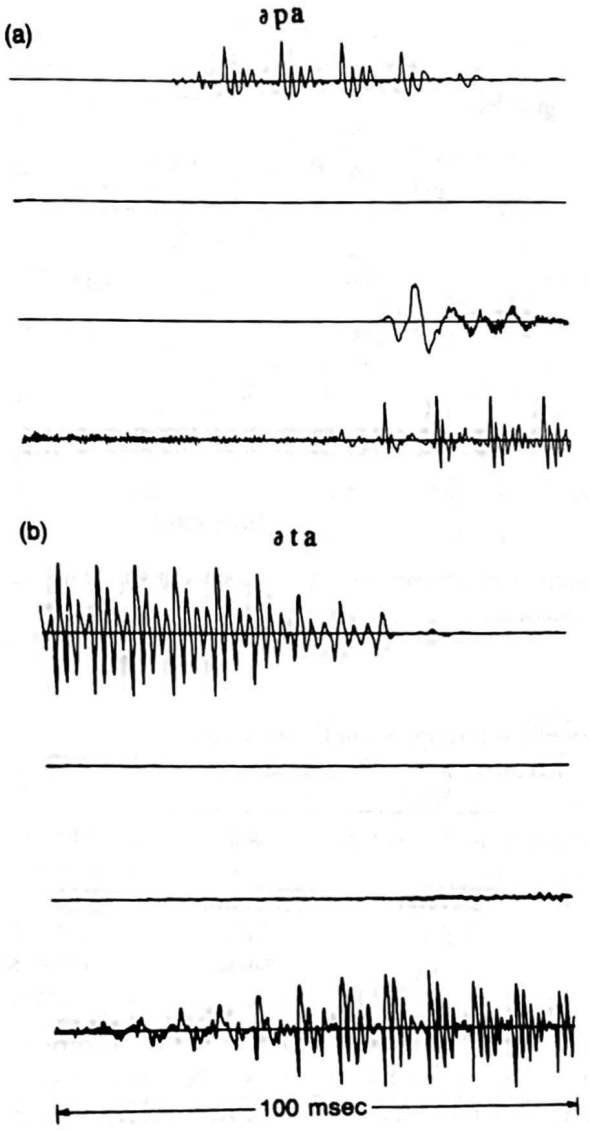


Figure 2.27 Waveforms for the sequences /ə-p-a/ and /ə-t-a/.

2.4.8 Review Exercises

As a self-check on the reader's understanding of the material on speech sounds and their acoustic manifestations, we now digress and present some simple exercises along with the solutions. For maximum effectiveness, the reader is encouraged to think through each exercise before looking at the solution.

Exercise 2.1

1. Write out the phonetic transcription for the following words:

   he, eats, several, light, tacos

2. What effect occurs when these five words are spoken in sequence as a sentence? What does this imply about automatic speech recognition?

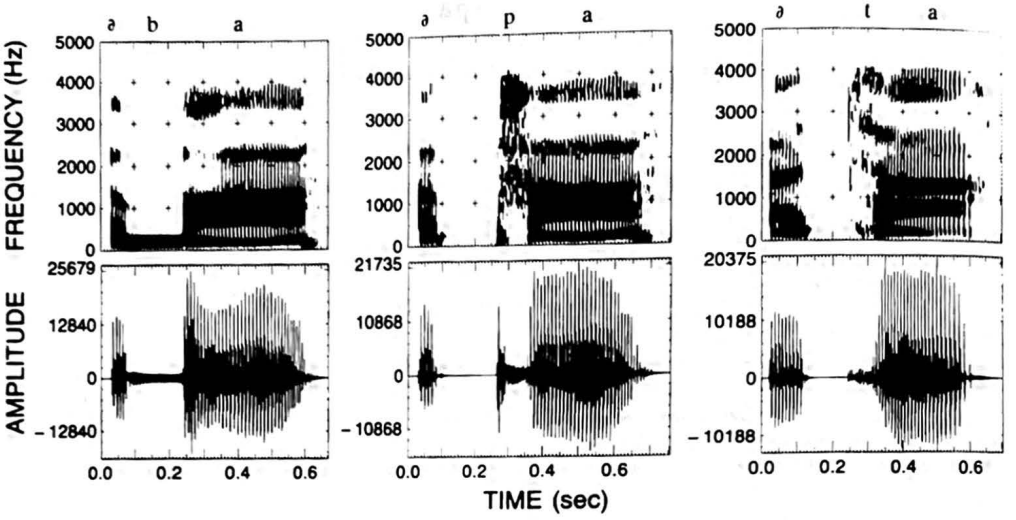


Figure 2.28 Spectrogram comparisons of the sequences of voiced (/ə-b-a/) and voiceless (/ə-p-a/ and /ə-t-a/) stop consonants.

Solution 2.1

1. The phonetic transcriptions of the words are

Word	Phoneme Sequence	ARPABET
he	/hi/	HH–IY
eats	/its/	IY–TS
several	/sɛvrʌl/	S–EH–V–R–AH–L
light	/laʸt/	L–AY–T
tacos	/takoz/	T–AA–K–OW–Z

2. When the words are spoken together, the last sound of each word merges with the first sound of the succeeding word (since they are the same sound), resulting in strong coarticulation of boundary sounds. The ARPABET transcription for the sentence is:

   HH–IY–T–S–EH–V–R–AH–L–AY–T–AA–K–OW–Z

   All information about word boundaries is totally lost; furthermore, the durations of the common sounds at the boundaries of words are much shorter than what would be predicted from the individual words.

Exercise 2.2

Some of the difficulties in large vocabulary speech recognition are related to the irregularities in the way basic speech sounds are combined to produce words. Exercise 2.2 highlights a couple of these difficulties.

1. In word initial position of American English, which phoneme or phonemes can never occur? Which hardly ever occur?
2. There are many word initial consonant clusters of length two, such as *sp*eak, *dr*ank, *pl*ead, and *pr*ess. How many word initial consonant clusters of length three are there in American English? What general rule can you give about the sounds in each of the three positions?

Ex.1022 / Page 69 of 333
AMAZON.COM, INC.

3. A nasal consonant can be combined with a stop consonant (e.g., ca*mp*, te*nd*) in a limited number of ways. What general rule do such combinations obey? There are several notable exceptions to this general rule. Can you give a couple of exceptions? What kind of speaking irregularity often results from these exceptions?

Solution 2.2

1. The only phoneme that never occurs in initial word position in English is the /ng/ sound (e.g., si*ng*). The only other sound that almost never occurs naturally in English, in initial word position, is /zh/ except some foreign words imported into English, such as gendarme, which does have an initial /zh/.
2. The word initial consonant clusters of length three in English include

/spl/	—	split
/spr/	—	spring
/skw/	—	squirt
/skr/	—	script
/str/	—	string

The general rule for such clusters is

/sound s/unvoiced stop/semivowel/

3. The general rule for a nasal-stop combination is that the nasal and stop have the same place of articulation, e.g., front/lips (/mp/), mid/dental (/nt/), back/velar (/ng k/). Exceptions occur in words like summed (/md/) or hanged (/ng d/) or dreamt (/mt/). There is often a tendency to insert an extra stop in such situations (e.g., dreamt → /drempt/).

Exercise 2.3

An important speech task is accurate digit recognition. This exercise seeks to exploit knowledge of acoustic phonetics to recognize first isolated digits, and next some simple connected digit strings. We first need a sound lexicon (a dictionary) for the digits. The sound lexicon describes the pronunciations of digits in terms of the basic sounds of English. Such a sound lexicon is given in Table 2.3. A single male adult talker (LRR) spoke each of the 11 digits in random sequence and in isolation, and spectrograms of these spoken utterances are shown in Figure 2.29. Figure 2.30 shows spectrograms of two connected digit sequences spoken by the same talker.

1. Identify each of the 11 digits based on the acoustic properties of the sounds within the digit (as expressed in the sound lexicon). Remember that each digit was spoken exactly once.
2. Try to identify the spoken digits in each of the connected digit strings.

Solution 2.3

1. The digits of the top row are 3 and 7:
   a. The digit 3 is cued by the distinctive brief initial fricative (/θ/), followed by the semivowel /r/ where the second and third formants both get very low in frequency, followed by the /i/ where F_2 and F_3 both become very high in frequency.
   b. The digit 7 is cued by the strong /s/ frication at the beginning, the distinctive /ɛ/, followed by the voiced fricative /v/, a short vowel /ə/ and ending in the strong

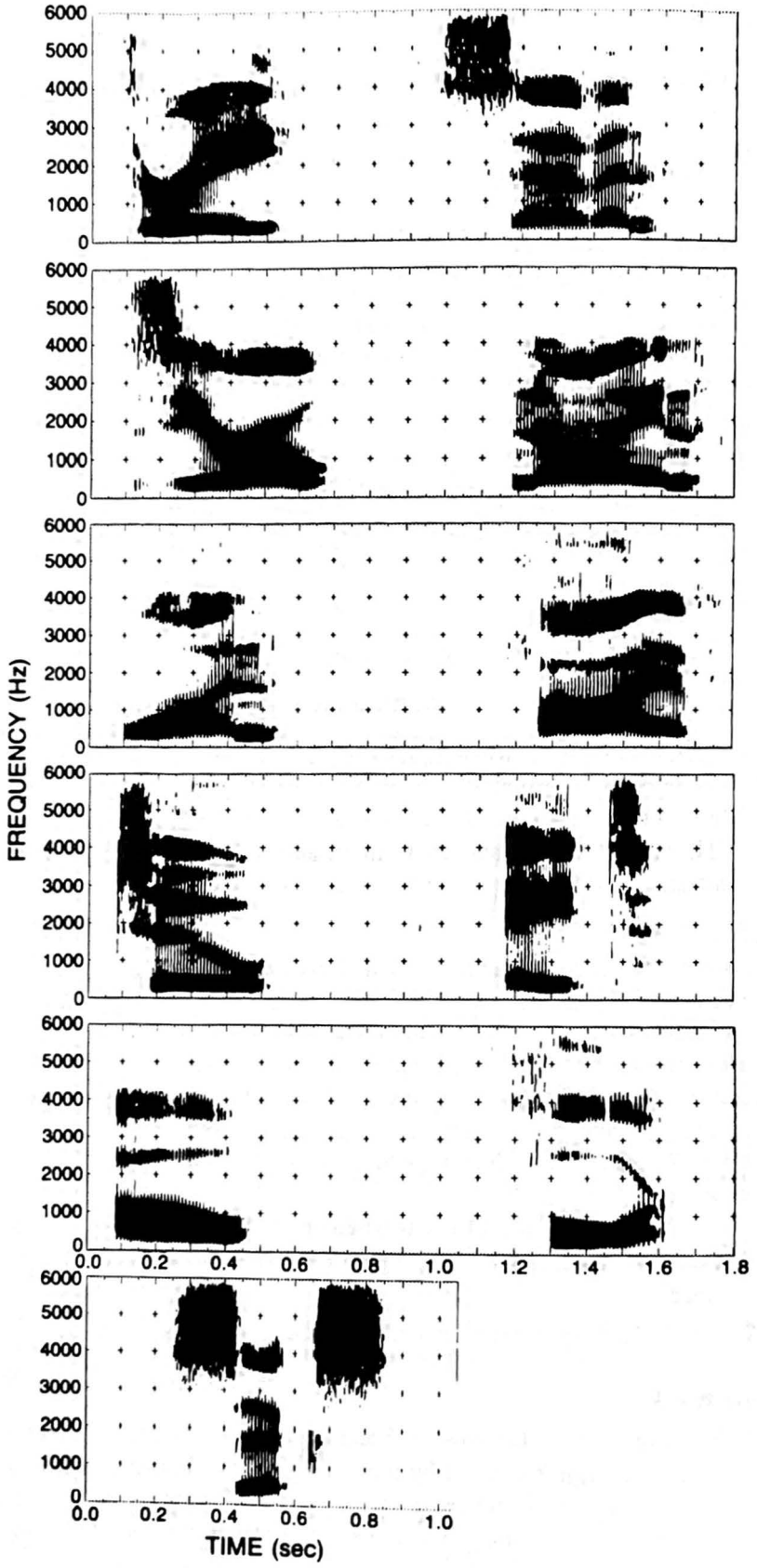


Figure 2.29 Spectrograms of the 11 isolated digits, 0 through 9 plus oh, in random sequence.

TABLE 2.3. Sound Lexicon of Digits

Word	Sounds	ARPABET
Zero	/z I r o/	Z–IH–R–OW
One	/w ʌ n/	W–AH–N
Two	/t u/	T–UW
Three	/θ r i/	TH–R–IY
Four	/f o r/	F–OW–R
Five	/f a^y v/	F–AY–V
Six	/s I k s/	S–IH–K–S
Seven	/s ɛ v ə n/	S–EH–V–AX–N
Eight	/e^y t/	EY–T
Nine	/n a^y n/	N–AY–N
Oh	/o/	OW

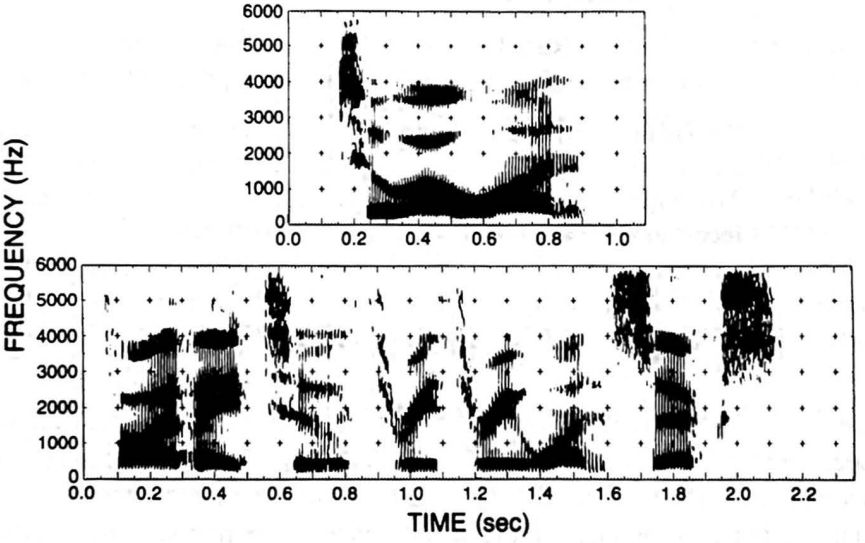


Figure 2.30 Spectrograms of two connected digit sequences.

nasal /n/.

The digits in the second row are 0 and 9:

a. The initial /z/ is cued by the strong frication with the presence of voicing at low frequencies; the following /I/ is seen by the high F_2 and F_3, the /r/ is signaled by the low F_2 and F_3, and the diphthong /o/ is signaled by the gliding motion of F_2 and F_3 toward an /u/-like sound.

b. The digit 9 is cued by the distinct initial and final nasals /n/ and by the /a^y/ glide between the nasals.

The digits in the third row are 1 and 5:

a. The digit 1 is cued by the strong initial semivowel /w/ with very low F_2 and by the strong final nasal /n/.

b. The digit 5 is cued by the weak initial frication of /f/, followed by the strong diphthong /a^y/ and ending in the very weak fricative /v/.

The digits in the fourth row are 2 and 8:

a. The digit 2 is cued by the strong /t/ burst and release followed by the glide to the /u/ sound.

b. The digit 8 is cued by the initial weak diphthong /e^{y}/ followed by a clear stop gap of the /t/ and then the /t/ release.

The digits in the fifth row are "oh" and 4:

a. The digit "oh" is virtually a steady sound with a slight gliding tendency toward /u/ at the end.

b. The digit 4 is cued by the weak initial fricative /f/, followed by the strong /o/ vowel and ending with a classic /r/ where F_2 and F_3 merge together.

The digit in the last row is 6:

a. The digit 6 is cued by the strong /s/ frication at the beginning and end, and by the steady vowel /I/ followed by the stop gap and release of the /k/.

2. By examining the isolated digit sequences, one can eventually (with a lot of work and some good luck) conclude that the two sequences are

Row 1:	2-oh-1	(telephone area code)
Row 2:	5-8-2-3-3-1-6	(7-digit telephone number)

We will defer any explanation of how any reasonable person, or machine, could perform this task until later in this book when we discuss connected word-recognition techniques. The purpose of this exercise is to convince the reader how difficult a relatively simple recognition task can be.

2.5 APPROACHES TO AUTOMATIC SPEECH RECOGNITION BY MACHINE

The material presented in the previous sections leads to a straightforward way of performing speech recognition by machine whereby the machine attempts to decode the speech signal in a sequential manner based on the observed acoustic features of the signal and the known relations between acoustic features and phonetic symbols. This method, appropriately called the acoustic-phonetic approach, is indeed viable and has been studied in great depth for more than 40 years. However, for a variety of reasons, the acoustic-phonetic approach has not achieved the same success in practical systems as have alternative methods. Hence, in this section, we provide an overview of several proposed approaches to automatic speech recognition by machine with the goal of providing some understanding as to the essentials of each proposed method, and the basic strengths and weaknesses of each approach.

Broadly speaking, there are three approaches to speech recognition, namely:

1. the acoustic-phonetic approach
2. the pattern recognition approach
3. the artificial intelligence approach

The acoustic-phonetic approach is based on the theory of acoustic phonetics that postulates that there exist finite, distinctive phonetic units in spoken language and that the phonetic

Ex.1022 / Page 73 of 333
AMAZON.COM, INC.