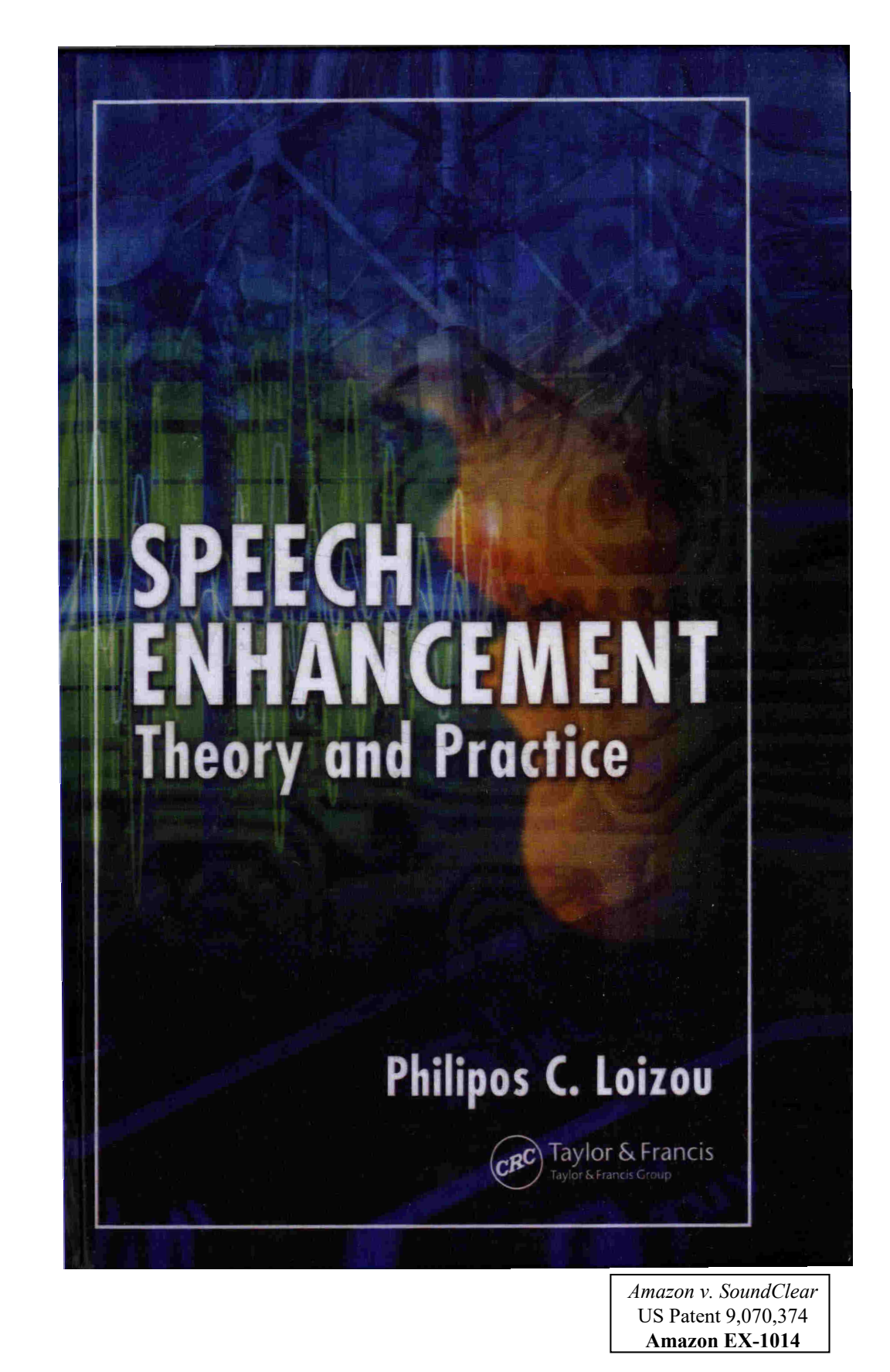


The background of the book cover is a dark, textured blue. Overlaid on this is a faint, stylized image of a human face, possibly a woman, with a warm, orange-gold glow emanating from the mouth area. Superimposed over the face and background are several thin, light blue lines that resemble sound waves or a spectrogram, adding a technical or scientific feel to the design.

SPEECH ENHANCEMENT

Theory and Practice

Philipos C. Loizou



Taylor & Francis
Taylor & Francis Group

Amazon v. SoundClear
US Patent 9,070,374
Amazon EX-1014

SPEECH ENHANCEMENT

Theory and Practice

Signal Processing and Communications

Editorial Board

Maurice G. Ballanger, *Conservatoire National
des Arts et Métiers (CNAM), Paris*

Ezio Biglieri, *Politecnico di Torino, Italy*

Sadaoki Furui, *Tokyo Institute of Technology*

Yih-Fang Huang, *University of Notre Dame*

Nikil Jayant, *Georgia Institute of Technology*

Aggelos K. Katsaggelos, *Northwestern University*

Mos Kaveh, *University of Minnesota*

P. K. Raja Rajasekaran, *Texas Instruments*

John Aasted Sorenson, *IT University of Copenhagen*

1. Digital Signal Processing for Multimedia Systems, *edited by Keshab K. Parhi and Takao Nishitani*
2. Multimedia Systems, Standards, and Networks, *edited by Atul Puri and Tsuhan Chen*
3. Embedded Multiprocessors: Scheduling and Synchronization, *Sundararajan Sriram and Shuvra S. Bhattacharyya*
4. Signal Processing for Intelligent Sensor Systems, *David C. Swanson*
5. Compressed Video over Networks, *edited by Ming-Ting Sun and Amy R. Reibman*
6. Modulated Coding for Intersymbol Interference Channels, *Xiang-Gen Xia*
7. Digital Speech Processing, Synthesis, and Recognition: Second Edition, Revised and Expanded, *Sadaoki Furui*
8. Modern Digital Halftoning, *Daniel L. Lau and Gonzalo R. Arce*
9. Blind Equalization and Identification, *Zhi Ding and Ye (Geoffrey) Li*
10. Video Coding for Wireless Communication Systems, *King N. Ngan, Chi W. Yap, and Keng T. Tan*
11. Adaptive Digital Filters: Second Edition, Revised and Expanded, *Maurice G. Bellanger*
12. Design of Digital Video Coding Systems, *Jie Chen, Ut-Va Koc, and K. J. Ray Liu*
13. Programmable Digital Signal Processors: Architecture, Programming, and Applications, *edited by Yu Hen Hu*

14. Pattern Recognition and Image Preprocessing: Second Edition, Revised and Expanded, *Sing-Tze Bow*
15. Signal Processing for Magnetic Resonance Imaging and Spectroscopy, *edited by Hong Yan*
16. Satellite Communication Engineering, *Michael O. Kolawole*
17. Speech Processing: A Dynamic and Optimization-Oriented Approach, *Li Deng*
18. Multidimensional Discrete Unitary Transforms: Representation: Partitioning and Algorithms, *Artyom M. Grigoryan, Sos S. Agaian, and S.S. Agaian*
19. High-Resolution and Robust Signal Processing, *Yingbo Hua, Alex B. Gershman and Qi Cheng*
20. Domain-Specific Processors: Systems, Architectures, Modeling, and Simulation, *Shuvra Bhattacharyya; Ed Deprettere; Jurgen Teich*
21. Watermarking Systems Engineering: Enabling Digital Assets Security and Other Applications, *Mauro Barni, Franco Bartolini*
22. Biosignal and Biomedical Image Processing: MATLAB-Based Applications, *John L. Semmlow*
23. Broadband Last Mile Technologies: Access Technologies for Multimedia Communications, *edited by Nikil Jayant*
24. Image Processing Technologies: Algorithms, Sensors, and Applications, *edited by Kiyoharu Aizawa, Katsuhiko Sakaue and Yasuhito Suenaga*
25. Medical Image Processing, Reconstruction and Restoration: Concepts and Methods, *Jiri Jan*
26. Multi-Sensor Image Fusion and Its Applications, *edited by Rick Blum and Zheng Liu*
27. Advanced Image Processing in Magnetic Resonance Imaging, *edited by Luigi Landini, Vincenzo Positano and Maria Santarelli*
28. Digital Video Image Quality and Perceptual Coding, *edited by H.R. Wu and K.R. Rao*
29. Document and Image Compression, *edited by Mauro Barni*
30. Speech Enhancement: Theory and Practice, *Philipos C. Loizou*

SPEECH ENHANCEMENT

Theory and Practice

Philipos C. Loizou



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2007 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

International Standard Book Number-13: 978-0-8493-5032-0 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Speech enhancement : theory and practice / author/editor, Philipos C. Loizou.

p. cm. -- (Signal processing and communications)

Includes bibliographical references and index.

ISBN 978-0-8493-5032-0 (alk. paper)

1. Speech processing systems. 2. Signal processing--Digital techniques. 3. Image processing--Digital techniques. I. Loizou, Philipos C. II. Title. III. Series.

TK7882.S65S662 2007

621.382'8--dc22

2006100005

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Dedication

To my parents, my wife, Demetria, and my son, Costakis

Table of Contents

Preface	xvii
The Author	xix

Chapter 1

Introduction	1
1.1 Understanding the Enemy: Noise	2
1.1.1 Noise Sources	2
1.1.2 Noise and Speech Levels in Various Environments	5
1.2 Classes of Speech Enhancement Algorithms	6
1.3 Book Organization	7
References	9

Part 1

Fundamentals	11
---------------------------	----

Chapter 2

Discrete-Time Signal Processing and Short-Time Fourier Analysis	13
2.1 Discrete-Time Signals	13
2.2 Linear Time-Invariant Discrete-Time Systems	16
2.2.1 Difference Equations	16
2.2.2 Linear Convolution	17
2.3 The z-Transform	18
2.3.1 Properties	18
2.3.2 The z-Domain Transfer Function	19
2.4 Discrete-Time Fourier Transform	21
2.4.1 DTFT Properties	22
2.4.2 Discrete Fourier Transform	24
2.4.3 Windowing	27
2.5 Short-Time Fourier Transform	32
2.5.1 Definition	32
2.5.2 Interpretations of the STFT	33
2.5.3 Sampling the STFT in Time and Frequency	35
2.5.4 Short-Time Synthesis of Speech	36
2.5.4.1 Filterbank Summation for Short-Time Synthesis of Speech	37
2.5.4.2 Overlap-and-Add Method for Short-Time Synthesis	39

2.6	Spectrographic Analysis of Speech Signals.....	42
2.7	Summary.....	43
	References	44

Chapter 3

	Speech Production and Perception	45
3.1	The Speech Signal.....	45
3.2	The Speech Production Process	45
	3.2.1 Lungs	46
	3.2.2 Larynx and Vocal Folds.....	47
	3.2.3 Vocal Tract.....	51
3.3	Engineering Model of Speech Production	54
3.4	Classes of Speech Sounds.....	55
3.5	Acoustic Cues in Speech Perception	57
	3.5.1 Vowels and Diphthongs.....	57
	3.5.2 Semivowels	60
	3.5.3 Nasals.....	61
	3.5.4 Stops	62
	3.5.4 Fricatives.....	64
3.6	Summary.....	66
	References	66

Chapter 4

	Noise Compensation by Human Listeners	69
4.1	Intelligibility of Speech in Multiple-Talker Conditions	70
	4.1.1 Effect of Masker's Spectral/Temporal Characteristics and Number of Talkers: Monaural Hearing.....	70
	4.1.2 Effect of Source Spatial Location: Binaural Hearing	73
4.2	Acoustic Properties of Speech Contributing to Robustness	78
	4.2.1 Shape of the Speech Spectrum	78
	4.2.2 Spectral Peaks.....	80
	4.2.3 Periodicity.....	83
	4.2.4 Rapid Spectral Changes Signaling Consonants.....	83
4.3	Perceptual Strategies for Listening in Noise	85
	4.3.1 Auditory Streaming	85
	4.3.2 Listening in the Gaps and Glimpsing	86
	4.3.3 Use of F0 Differences	87
	4.3.4 Use of Linguistic Knowledge	88
	4.3.5 Use of Spatial and Visual Cues.....	89
4.4	Summary.....	89
	References	90

Part 2	
Algorithms	95

Chapter 5

Spectral-Subtractive Algorithms	97
5.1 Basic Principles of Spectral Subtraction	97
5.2 A Geometric View of Spectral Subtraction	101
5.2.1 Upper Bounds on the Difference Between the Noisy and Clean Signals' Phases	102
5.2.2 Alternate Spectral-Subtractive Rules and Theoretical Limits	104
5.3 Shortcomings of the Spectral Subtraction Method	110
5.4 Spectral Subtraction Using Oversubtraction	112
5.5 Nonlinear Spectral Subtraction	119
5.6 Multiband Spectral Subtraction	120
5.7 MMSE Spectral Subtraction Algorithm	125
5.8 Extended Spectral Subtraction	128
5.9 Spectral Subtraction Using Adaptive Gain Averaging	130
5.10 Selective Spectral Subtraction	133
5.11 Spectral Subtraction Based on Perceptual Properties	135
5.12 Performance of Spectral Subtraction Algorithms	136
5.13 Summary	138
References	139

Chapter 6

Wiener Filtering	143
6.1 Introduction to Wiener Filter Theory	143
6.2 Wiener Filters in the Time Domain	144
6.3 Wiener Filters in the Frequency Domain	146
6.4 Wiener Filters and Linear Prediction	148
6.5 Wiener Filters for Noise Reduction	150
6.5.1 Square-Root Wiener Filter	158
6.5.2 Parametric Wiener Filters	158
6.6 Iterative Wiener Filtering	163
6.6.1 Mathematical Speech Production Model	164
6.6.2 Statistical Parameter Estimation of the All-Pole Model in Noise	165
6.7 Imposing Constraints on Iterative Wiener Filtering	172
6.7.1 Across-Time Spectral Constraints	172
6.7.2 Across-Iterations Constraints	176
6.8 Constrained Iterative Wiener Filtering	177
6.9 Constrained Wiener Filtering	180
6.9.1 Mathematical Definitions of Speech and Noise Distortions	180
6.9.2 Limiting the Noise Distortion Level	184

6.10	Estimating the Wiener Gain Function.....	187
6.11	Incorporating Psychoacoustic Constraints in Wiener Filtering.....	192
6.11.1	Shaping the Noise Distortion in the Frequency Domain.....	192
6.11.2	Using Masking Thresholds as Constraints	195
6.12	Codebook-Driven Wiener Filtering.....	198
6.13	Audible Noise Suppression Algorithm	202
6.14	Summary.....	208
	References	209

Chapter 7

	Statistical-Model-Based Methods	213
7.1	Maximum-Likelihood Estimators	213
7.2	Bayesian Estimators	219
7.3	MMSE Estimator.....	219
7.3.1	MMSE Magnitude Estimator.....	222
7.3.2	MMSE Complex Exponential Estimator	227
7.3.3	Estimating the <i>A Priori</i> SNR.....	228
7.3.3.1	Maximum-Likelihood Method.....	229
7.3.3.2	Decision-Directed Approach.....	230
7.4	Improvements to the Decision-Directed Approach	231
7.4.1	Reducing the Bias	232
7.4.2	Improving the Adaptation Speed	233
7.5	Implementation and Evaluation of the MMSE Estimator.....	237
7.6	Elimination of Musical Noise	238
7.7	Log-MMSE Estimator.....	240
7.8	MMSE Estimation of the p th-Power Spectrum.....	242
7.9	MMSE Estimators Based on Non-Gaussian Distributions.....	247
7.10	Maximum <i>A Posteriori</i> (MAP) Estimators.....	251
7.11	General Bayesian Estimators	254
7.12	Perceptually Motivated Bayesian Estimators	256
7.12.1	Psychoacoustically Motivated Distortion Measure	256
7.12.2	Weighted Euclidean Distortion Measure	257
7.12.3	Itakura-Saito Measure.....	262
7.12.4	Cosh Measure.....	263
7.12.5	Weighted Likelihood Ratio	266
7.12.6	Modified IS Distortion Measure	266
7.13	Incorporating Speech Absence Probability in Speech Enhancement	269
7.13.1	Incorporating Speech-Presence Uncertainty in Maximum-Likelihood Estimators.....	270
7.13.2	Incorporating Speech-Presence Uncertainty in MMSE Estimators	272

7.13.3	Incorporating Speech-Presence Uncertainty in Log-MMSE Estimators.....	277
7.13.4	Implementation Issues Regarding <i>A Priori</i> SNR Estimation.....	279
7.14	Methods for Estimating the <i>A Priori</i> Probability of Speech Absence.....	279
7.15	Summary.....	285
	References.....	286

Chapter 8

	Subspace Algorithms.....	291
8.1	Introduction	291
8.1.1	Definitions	292
8.1.2	Projections	293
8.1.3	Low-Rank Modeling	298
8.2	Using SVD for Noise Reduction: Theory	300
8.2.1	SVD Analysis of "Noisy" Matrices	300
8.2.2	Least-Squares and Minimum-Variance Estimates of the Signal Matrix.....	303
8.3	SVD-Based Algorithms: White Noise	306
8.3.1	SVD Synthesis of Speech	306
8.3.2	Determining the Effective Rank	311
8.3.4	Noise Reduction Algorithm	315
8.4	SVD-Based Algorithms: Colored Noise	316
8.5	SVD-Based Methods: A Unified View	320
8.6	EVD-Based Methods: White Noise	320
8.6.1	Eigenvalue Analysis of "Noisy" Matrices	320
8.6.2	Subspace Methods Based on Linear Estimators.....	325
8.6.2.1	Linear Minimum Mean-Square Estimator (LMMSE)	326
8.6.2.2	Time-Domain-Constrained Estimator	328
8.6.2.3	Spectral-Domain-Constrained Estimator	332
8.6.3	Implementation.....	338
8.6.3.1	Covariance Estimation	338
8.6.3.2	Estimating the Lagrange Multiplier.....	340
8.6.3.3	Estimating the Signal Subspace Dimension	342
8.7	EVD-Based Methods: Colored Noise.....	344
8.7.1	Prewhitening Approach	345
8.7.2	Signal/Noise KLT-Based Method	349
8.7.3	Adaptive KLT Approach	352
8.7.4	Subspace Approach with Embedded Prewhitening	354
8.7.4.1	Time-Domain-Constrained Estimator	354
8.7.4.2	Spectrum-Domain-Constrained Estimator	356

8.7.4.3	Implementation.....	359
8.7.4.4	Relationship Between Subspace Estimators and Prewhitening.....	361
8.8	EVD-Based Methods: A Unified View.....	366
8.9	Perceptually Motivated Subspace Algorithms.....	367
8.9.1	Fourier to Eigen-Domain Relationship.....	368
8.9.2	Incorporating Psychoacoustic Model Constraints.....	372
8.9.3	Incorporating Auditory Masking Constraints.....	374
8.10	Subspace-Tracking Algorithms.....	376
8.10.1	Block Algorithms.....	377
8.10.2	Recursive Algorithms.....	383
8.10.2.1	Modified Eigenvalue Problem Algorithms.....	384
8.10.2.2	Adaptive Algorithms.....	385
8.10.3	Using Subspace-Tracking Algorithms in Speech Enhancement.....	392
8.11	Summary.....	393
	References.....	394

Chapter 9

	Noise Estimation Algorithms.....	399
9.1	Voice Activity Detection Vs. Noise Estimation.....	399
9.2	Introduction to Noise Estimation Algorithms.....	401
9.3	Minimal-Tracking Algorithms.....	403
9.3.1	Minimum Statistics (MS) Noise Estimation.....	403
9.3.1.1	Principles.....	403
9.3.1.2	Derivation of the Bias Factor.....	405
9.3.1.3	Derivation of Optimal Time- and Frequency-Dependent Smoothing Factor.....	411
9.3.1.4	Searching for the Minimum.....	414
9.3.1.5	Minimum Statistics Algorithm.....	415
9.3.2	Continuous Spectral Minimum Tracking.....	417
9.4	Time-Recursive Averaging Algorithms for Noise Estimation.....	420
9.4.1	SNR-Dependent Recursive Averaging.....	421
9.4.2	Weighted Spectral Averaging.....	423
9.4.3	Recursive Averaging Algorithms Based on Signal-Presence Uncertainty.....	429
9.4.3.1	Likelihood Ratio Approach.....	430
9.4.3.2	Minima-Controlled Recursive Averaging (MCRA) Algorithms.....	434
9.5	Histogram-Based Techniques.....	446
9.6	Other Noise Estimation Algorithms.....	453
9.7	Objective Comparison of Noise Estimation Algorithms.....	455

9.8 Summary.....	459
References	460

Part 3

Evaluation	463
------------------	-----

Chapter 10

Evaluating Performance of Speech Enhancement Algorithms	465
10.1 Quality vs. Intelligibility.....	465
10.2 Evaluating Intelligibility of Processed Speech	466
10.2.1 Nonsense Syllable Tests.....	467
10.2.2 Word Tests.....	472
10.2.2.1 Phonetically Balanced Word Tests.....	472
10.2.2.2 Rhyming Word Tests.....	473
10.2.3 Sentence Tests	476
10.2.4 Measuring Speech Intelligibility	478
10.2.4.1 Speech Reception Threshold	478
10.2.4.2 Using Statistical Tests to Assess Significant Differences: Recommended Practice.....	480
10.3 Evaluating Quality of Processed Speech.....	486
10.3.1 Relative Preference Methods	486
10.3.2 Absolute Category Rating Methods.....	489
10.3.2.1 Mean Opinion Scores	490
10.3.2.2 Diagnostic Acceptability Measure.....	492
10.3.2.3 The ITU-T P.835 Standard	495
10.4 Evaluating Reliability of Quality Judgments: Recommended Practice.....	498
10.4.1 Intrarater Reliability Measures.....	498
10.4.2 Interrater Reliability Measures.....	500
10.5 Objective Quality Measures	502
10.5.1 Segmental SNR Measures: Time and Frequency	503
10.5.2 Spectral Distance Measures Based on LPC	506
10.5.3 Perceptually Motivated Measures	507
10.5.3.1 Weighted Spectral Slope (WSS) Distance Measure	508
10.5.3.2 Bark Distortion Measures	509
10.5.3.3 Perceptual Evaluation of Speech Quality (PESQ) Measure	514
10.5.4 Composite Measures	525
10.6 Nonintrusive Objective Quality Measures.....	527
10.7 Figures of Merit of Objective Quality Measures	528
10.8 Challenges and Future Directions in Objective Quality Evaluation	530

10.9 Summary	534
References	535

Chapter 11

Comparison of Speech Enhancement Algorithms	541
11.1 NOIZEUS: A Noisy Speech Corpus for Quality Evaluation of Speech Enhancement Algorithms	542
11.2 Comparison of Speech Enhancement Algorithms: Quality	543
11.2.1 Quality Evaluation: Procedure	544
11.2.2 Subjective Quality Evaluation: Results	545
11.2.3 Within-Class Algorithm Comparisons	545
11.2.4 Across-Class Algorithm Comparisons	550
11.2.5 Comparisons in Reference to Noisy Speech	554
11.2.6 Contribution of Speech and Noise Distortion to Judgment of Overall Quality	558
11.2.7 Summary of Findings	559
11.3 Comparison of Speech Enhancement Algorithms: Intelligibility	560
11.3.1 Listening Tests: Procedure	561
11.3.2 Intelligibility Evaluation: Results	562
11.3.3 Intelligibility Comparison Among Algorithms	564
11.3.4 Intelligibility Comparison Against Noisy Speech	564
11.3.5 Summary of Findings	567
11.4 Comparison of Objective Measures for Quality Evaluation	568
11.4.1 Objective Measures	568
11.4.2 Correlations of Objective Measures with Quality	573
11.4.3 Summary of Findings	576
11.5 Summary	578
References	579

Appendix A: Special Functions and Integrals	581
--	------------

Appendix B: Derivation of the MMSE Estimator	585
---	------------

Appendix C: Speech Databases and MATLAB Code	589
---	------------

Index	601
--------------------	------------

4 Noise Compensation by Human Listeners

In many everyday situations, we are surrounded by sounds originating from a variety of sources. These sounds might include a number of people speaking in the background, a car passing by, music playing on the radio, a telephone ringing, the wind blowing, and so on. The signals from each source are mixed together when they enter the two ears to form a composite waveform. The frequency components of each source are not constrained within an isolated region of the spectrum but overlap with frequency components of other sources. The problem of sorting out which components belong to which sound source is highly complex, but it is a task that the brain performs effortlessly, at least by listeners with normal hearing. In the context of signal processing, this task is equivalent to solving a blind source separation problem: given a mixture of signals recorded with multiple microphones, identify the target signal. How does the brain solve this blind source separation problem?

The purpose of this chapter is to review the literature investigating the cues used by human listeners when communicating in highly adverse conditions, particularly in situations where multiple talkers are speaking in the background. Understanding the strategies and acoustic cues used by human listeners to segregate speech from other signals in adverse conditions is important for several reasons. For one, such knowledge would tell us which spectral and temporal characteristics of the speech signal need to be preserved by speech enhancement algorithms. Second, it would help us design better speech enhancement algorithms that would mimic listening strategies used by humans. Such algorithms would hold promise in improving speech intelligibility in noise.

Research in this area was triggered by the seminal paper of Cherry [1] in the early 1950s on the intelligibility of speech in multiple-talker conditions. He referred to this listening task as the “cocktail party” problem, simply formulated as: How do we understand what one person is saying when others are speaking at the same time? The more general problem of segregating speech in the presence of multiple noise sources (not necessarily speech) is commonly referred to as the “auditory scene analysis” problem [2].

This chapter is organized into four sections. The first section reviews the research on speech intelligibility in multiple-talker conditions. The remaining sections focus on the various factors influencing speech segregation, including the properties of the speech signal itself that shield information from distortion. The perceptual strategies used by listeners for speech recognition in noisy environments are discussed in the last section.

4.1 INTELLIGIBILITY OF SPEECH IN MULTIPLE-TALKER CONDITIONS

The intelligibility of speech in adverse conditions is greatly influenced by the spectral and temporal characteristics of the noise, specifically, whether the noise is stationary or nonstationary, narrow-band, or wide-band, modulated or continuous (steady-state), and whether it is presented monaurally or binaurally [3]. In this section, we focus on intelligibility of speech in the presence of single or multiple talkers presented monaurally or binaurally. Because the interfering signal in multiple-talker listening conditions is speech, we will use the term *masker* to differentiate the interfering signal from the voice of interest, also called the *target* voice.

4.1.1 EFFECT OF MASKER'S SPECTRAL/TEMPORAL CHARACTERISTICS AND NUMBER OF TALKERS: MONAURAL HEARING

Several studies have demonstrated that the intelligibility of speech (presented monaurally) in multiple-talker conditions (e.g., in a restaurant) is greatly affected by the number of competing talkers present as well as by the temporal characteristics of the noise [4]. In an early study, Miller [4] obtained articulation scores — percent correct identification scores of monosyllabic words as a function of the intensity of the masker — for speech in the presence of one, two, four, six, and eight competing voices. The target voice was male, and the interfering voices consisted of equal number of male and female talkers. Results indicated that word identification in the presence of two voices was significantly more difficult than word identification in the presence of a single voice. Word identification became progressively worse as the number of voices increased, with only a small change in identification scores when the number of voices increased from four to eight. When a large number of voices are present, the masker waveform becomes nearly continuous, leaving no silent gaps in the waveform. As we will discuss in Sec. 4.3.2 these gaps or dips in the waveform produce high-SNR segments in the signal that enable the listeners to “hear out” (glimpse) segments of the target voice.

Evidence of the idea that normal-hearing listeners are exploiting the gaps or dips present in speech waveforms corrupted by a single competing voice comes from several studies. These studies compared speech intelligibility in the presence of steady-state continuous noise, amplitude-modulated noise, and single interfering voice. Results were reported in terms of the speech reception threshold (SRT), defined as the SNR (dB) at which 50% of the words are reliably identified (see Chapter 10 for more details on how to measure the SRT). Several studies confirmed that it is more difficult to identify words in the presence of steady-state noise than in the presence of a single competing speaker. The difference in SRT (50% intelligibility level) can range from 6 dB [5] to 10 dB [4]; that is to say, listeners can identify words in steady-state noise or in the presence of a single competing talker with the same accuracy (50% correct) even when the SNR of

the competing-talker sentences is 6–10 dB lower. The interfering voice is therefore a much less effective masker than steady-state noise when presented at the same level. The intelligibility advantage gained by using a single (or a small number) competing talker instead of steady-state noise as the interfering signal is described in the literature as *masking release* [6,7].

Studies that investigated the effect of the masker's temporal fluctuations showed that the difference in SRT between stationary noise and modulated noise was 4–6 dB [5,7,8]. Some studies used noise that was modulated by the envelope fluctuations of the speech signal, whereas others used noise constructed to have the same long-term spectrum as speech. The latter noise was used to eliminate or reduce spectral difference cues between the target and interfering signals. Other studies included time-reversed speech (i.e., speech played backwards) as maskers. These maskers were used to assess whether the difference in performance (with competing speech maskers) can be attributed to differences in temporal fluctuations or to the distractive linguistic contents of the interfering sentence. A large release of masking (5 dB) was obtained with speech-modulated maskers, suggesting that the temporal fluctuations in the interfering signal can have a significant effect on speech intelligibility in noise despite the similarities in the spectra of the signal and interfering noise. Masking release (of about 7 dB) was also obtained when time-reversed speech was used as the masker [5,8]. These findings provide evidence that the release of masking can be attributed solely to the temporal fluctuations of the interfering signal rather than to spectral differences between the masker and target signals.

Aside from the temporal fluctuations of the interfering signal, other factors can potentially influence intelligibility of speech in the presence of a single competing talker. These include gender differences and similarities in information content of the target and masker speech. As one would expect, recognition is easier when the talkers are of different sexes than when they are of the same sex [9]. Similarities in the information content of the target and masker sentences can also influence recognition, contributing to the so called “informational masking” [10]. Generally, it is believed that the difficulty normal-hearing listeners experience with single-talker maskers is attributed to a combination of (1) “energetic masking,” resulting from overlap of the target and masker in the auditory periphery and (2) “informational masking,” resulting from competition between the target and masker at more central stages of auditory processing [10]. Energetic masking is considered to be a peripheral masking phenomenon occurring when the energy from two (or more) sounds overlap both spectrally and temporally, making signal recovery difficult. Teasing out the contributions of energetic and informational masking has been a challenging task [10].

When the number of maskers or the number of voices increases beyond four, the temporal fluctuations average out, and the composite masker approximates a steady-state signal. Figure 4.1 (bottom panel) shows an example time waveform of eight-talker babble resembling that of stationary white noise. Unlike the waveform of a single competing talker, the waveform of the eight-talker babble does not contain any dips or gaps, at least in the time domain. Consequently, listeners

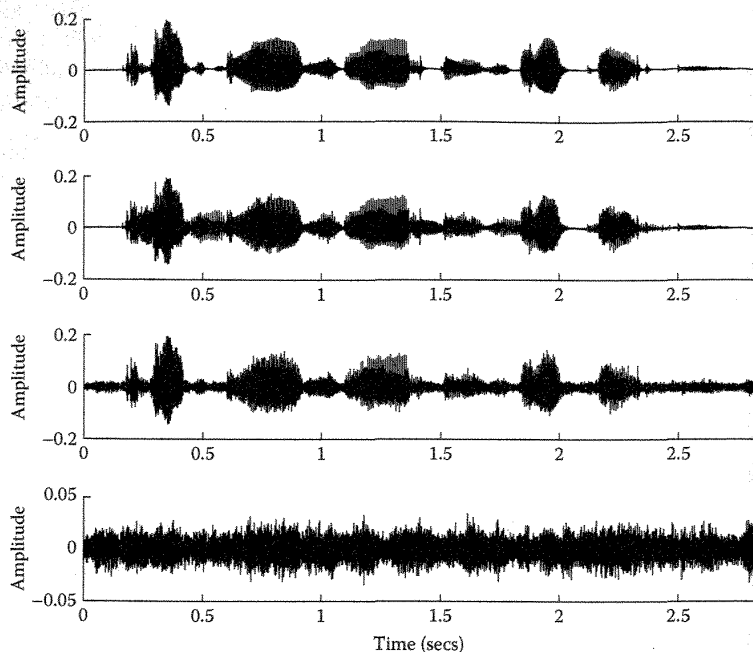


FIGURE 4.1 Time waveforms of the sentence “The birch canoe slid on the smooth planks” corrupted by a single-talker interferer (second panel from top) and 8-talker babble (third panel from top) at + 5 dB S/N. Top panel shows the sentence in quiet and bottom panel shows the time-waveform of 8-talker babble.

are not able to take advantage of any silent periods in the masker to hear out the target signal.

As demonstrated earlier, normal-hearing listeners benefit a great deal from the temporal fluctuations present in the interfering signal. In contrast, hearing-impaired listeners receive little benefit. Whereas normal-hearing listeners gain about 6–10 dB when steady-state noise is replaced by interfering speech or modulated noise, hearing impaired listeners gain no more than 0–2 dB [5]. This reduced release of masking is partly responsible for the difficulty hearing-impaired listeners experience in understanding speech in noisy situations.

So far, we have considered intelligibility of speech in the presence of multiple talkers when presented monaurally, as for instance, through telephones that is, listeners had to rely exclusively on monaural cues to recognize the target signal. The situation becomes more complicated when speech is presented binaurally, as the spatial locations of the signal and noise sources, among other factors, can significantly affect speech recognition. The latter situation is more realistic as it reflects everyday listening situations.

4.1.2 EFFECT OF SOURCE SPATIAL LOCATION: BINAURAL HEARING

In real-world listening scenarios, competing voices (along with other types of noise) typically originate from different locations in space. Consequently, each ear receives a slightly different signal. The signals emanating from various sources reach the two ears at different times and with different amplitudes owing to the attenuation from the mass of the head, often called *head shadow*. The differences in amplitude level due to head shadow are called interaural level differences (ILDs), and the differences in time of arrival are called interaural time differences (ITDs). Both ILDs and ITDs provide cues to sound localization and also contribute to binaural intelligibility of speech under noisy conditions.

It is reasonable to expect that binaural speech intelligibility is affected by the spatial distribution of the noise sources (i.e., their location with respect to the listener) and the number of sources. Several studies demonstrated large intelligibility benefit when the noise source was moved spatially away from the target along the horizontal plane [11–13]. In these studies, speech was presented from the front (0° azimuth), whereas the noise (single interfering noise source) was presented from various angles (azimuths) in the horizontal plane. Results indicated that the largest benefit (10–12 dB) in terms of SRT improvement occurred when the interfering sound source was located at 110 – 120° azimuth. This binaural benefit — often called *spatial release of masking* — is attributed to two components: one caused by head shadow, i.e., access to a “better” ear with more favorable SNR, and the other caused by interaural time delays (ITDs) [14]. To better understand the head shadow advantage, consider, for instance, the situation where the noise source is at 90° azimuth and the target signals come from the front of the listener. The ear contralateral to the noise source will receive a better SNR signal owing to head shadow, thus making it easier for the listener to identify the words spoken by the target speaker. The difference between the SRT obtained using the better ear (i.e., ear contralateral to the noise source) vs. the worse ear (ear closest to the noise source) can be as large as 8 dB when there is only a single masker present and the target signal comes from the front [11]. This difference, however, gradually disappears as the number of noise sources increases and the noise sources get distributed around the head [11].

When multiple noise sources are present, the spatial configuration of the sources has a large impact on speech intelligibility. More specifically, the spatial separation of the sources and the distribution of the sources around the listener play a critical role in speech intelligibility [11,12,15,16].

To illustrate this, we will consider the data collected in [12], in which target sentences were presented from the front and one to three interfering speakers were presented from different locations. To evaluate the benefit incurred with listening with two ears (binaural hearing) vs. listening with one ear (monaural hearing), data were collected for both binaural and monaural listening conditions. The various spatial configurations considered were simulated by convolving anechoic head-related impulse responses [17] (for different positions and the left and right ears) with the speech material. The “virtual” stimuli were presented

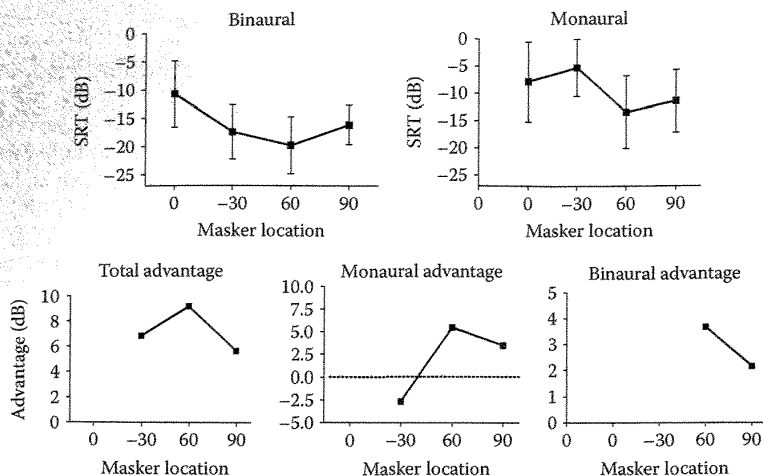


FIGURE 4.2 Binaural and monaural speech intelligibility (top two panels) as a function of masker (single talker) position expressed in degrees. Target signal was coming from the front of the listener, i.e., from 0°. Masker location of 0° indicates the front direction, positive angles indicate directions from the right of the listener, and negative angles indicate directions from the left of the listener. Masker was a single talker. Bottom panels show the advantage (in dB) gained when listening with two ears (see text for explanation).

binaurally to the listeners through headphones. The results are plotted in Figure 4.2 for a single interfering speaker and in Figure 4.3 for three interfering speakers. These results are expressed in terms of SRT (dB), for recognition of IEEE sentences [18]. A 10-dB SRT improvement was obtained simply by moving the noise sources (three talkers) at 90° azimuth. A smaller improvement (5 dB) was obtained when the three interfering speakers were located at (30°, 60°, 90°). Data from monaural listening are shown in panel b for comparison. In the monaural condition, the left ear was used as it was considered to be the “better” ear for the majority of the noise configurations (see Figure 4.3). Clearly, better performance is obtained with binaural than with monaural hearing, even when listening through the “better” ear — the ear farthest from the noise sources. As expected, the configuration (−30°, 60°, 90°) is least favorable for monaural hearing as the noise sources (interfering talkers) are distributed on both sides (left and right) of the listener.

The bottom panels in Figure 4.3 and Figure 4.2 summarize the results in terms of the total advantage, gained by the mere fact that the noise sources are spatially separated (total spatial release of masking), and its two additive components: monaural advantage and binaural advantage (also called *binaural squelch*). The monaural advantage is defined as the difference in SRT between each monaural

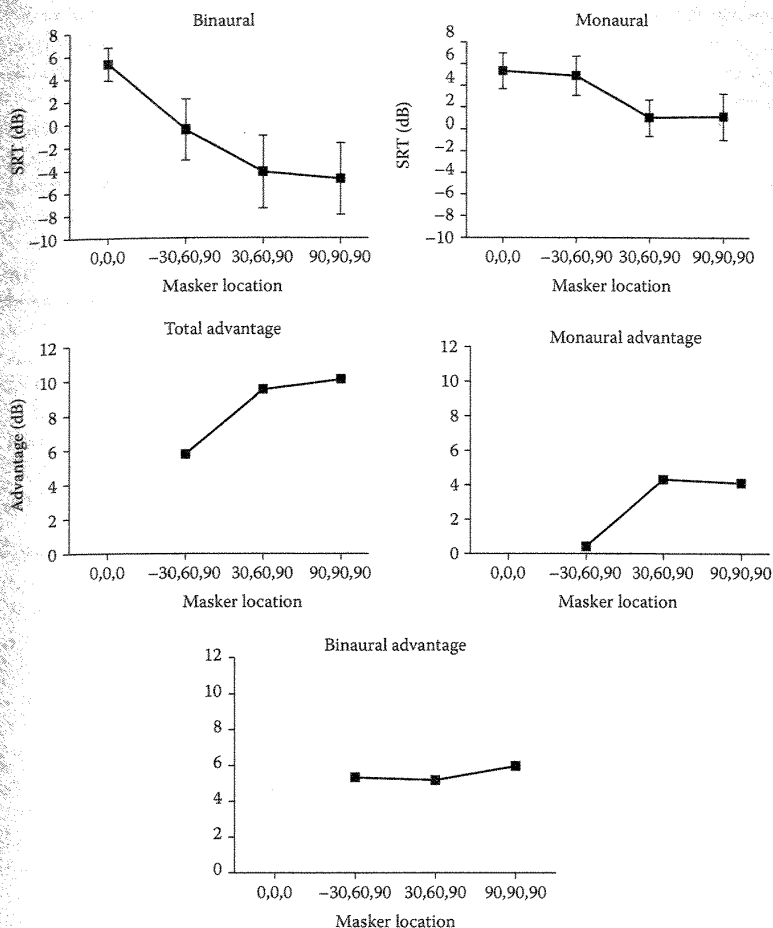


FIGURE 4.3 Binaural and monaural speech intelligibility (top two panels) as a function of the position of three different interfering talkers (maskers). The spatial configuration (30°, 60°, 90°) for instance indicates that the three interfering talkers were located at 30°, 60°, and 90°, respectively. Target signal was coming from the front of the listener, i.e., from 0°. Bottom panels show the advantage (in dB) gained when listening with two ears (see text for explanation).

(best-ear listening) condition where the noise and target are spatially separated and the unseparated condition (both speech and noise come from the front). The binaural advantage is defined as the part of the total spatial release advantage that is not accounted for by the monaural advantage. The binaural advantage reflects the true

benefit obtained when listening with two ears over listening with just the better monaural ear. The total advantage is greatly affected by the noise spatial configuration, with the largest benefit obtained at (90°, 90°, 90°) where the noise sources were locally isolated.

The monaural advantage is also affected by the spatial configuration and disappears once the interfering sources are spatially distributed on the right and left of the listener. This finding suggests that the head shadow plays a minor role in realistic listening environments when competing noise sources (e.g., speakers) surround the listener. Unlike the monaural advantage, the binaural advantage is more robust to spatial configurations. The data in [12] showed that the binaural advantage increased as the number of interfering talkers increased from 1 to 3, varying from a low of 3 dB (one interfering talker) to 7 dB (three interfering talkers) improvement in SRT. The temporal properties of the interfering sound can affect the binaural advantage, with larger benefits obtained with multiple speech maskers [12,13,19] than with multiple noiselike maskers (e.g., continuous, speech-modulated). For the latter maskers, a limited 3 dB binaural advantage was maintained in [11] independent of the spatial configuration (symmetric vs. asymmetric) and number of speech-modulated noise sources.

As indicated earlier, when multiple noise sources are present, the importance of spatial configuration increases. In fact, the effect of spatial configuration is greater than that of adding or removing one source; that is, one can compensate for an additional interfering noise source simply by choosing a more favorable configuration [15]. In the study by Yost et al. [16], the scores for correct identification of words, letters, and numbers presented simultaneously from three different sources increased by 30% when the angle between adjacent sources increased from 30° to 90°. In contrast, the scores increased only by 10% when the number of noise sources was reduced from 3 to 2. The study concluded that binaural hearing does provide significant benefit when there are more than two noise sources present. It should be noted that a "divided attention" task was used in the preceding study, in that listeners were asked to identify all active sound sources rather than identify a particular source, as is typically done in most studies. The latter task is referred to as *selective attention* task because the listener attends to one source while ignoring other competing sources. Some argue that the selective attention task is more reflective of "cocktail party" situations, where we attend to one voice while ignoring other voices.

As mentioned earlier, access to both ILD and ITD cues contributes to binaural speech intelligibility. However, the contributions of ITD and ILD cues are not additive and depend on the spectral content of the speech material as well as the spatial configuration of the noise sources [20,21]. The benefits of ITD and ILD cues are known to be frequency dependent, with the ITD cues being more useful in the low frequencies and ILD cues being more useful in the high frequencies. This dependence is due to the fact that low frequencies are diffracted around the head with little attenuation, while high frequencies (>4 kHz) are heavily attenuated.

Several studies investigated the individual contributions of ILD and ITD cues by creating stimuli that contained either ILD cues alone, ITD cues alone, or both [20,21]. Results from [21] indicated that the use of ILD cues is dominated by best-ear listening (see Figure 4.4). This was based on the fact that intelligibility of ILD-only stimuli was the same in the $(30^\circ, 60^\circ, 90^\circ)$ configuration, where the noise sources were spatially separated from speech, as it was in the $(0^\circ, 0^\circ, 0^\circ)$ configuration, where the noise sources were coincident with speech. From Figure 4.4, it is clear that the use of ILD cues is heavily influenced by the location of the noise sources. Performance in the $(-30^\circ, 60^\circ, 90^\circ)$ configuration, for instance, was significantly worse than performance in other configurations because the noise sources were distributed across the two hemifields. In contrast, the use of ITD cues is relatively more robust (see Figure 4.4) to the location of noise sources, suggesting that the binaural system is able to exploit ITDs even in complex listening situations where multiple speakers might be present.

Considerable effort has been made to find ways to predict the binaural benefit from spatial release of masking [14,22]. One simple formula that was found to work quite well in predicting binaural intelligibility data is the one proposed in [14]. For speech presented from the front, the release of masking due to spatial

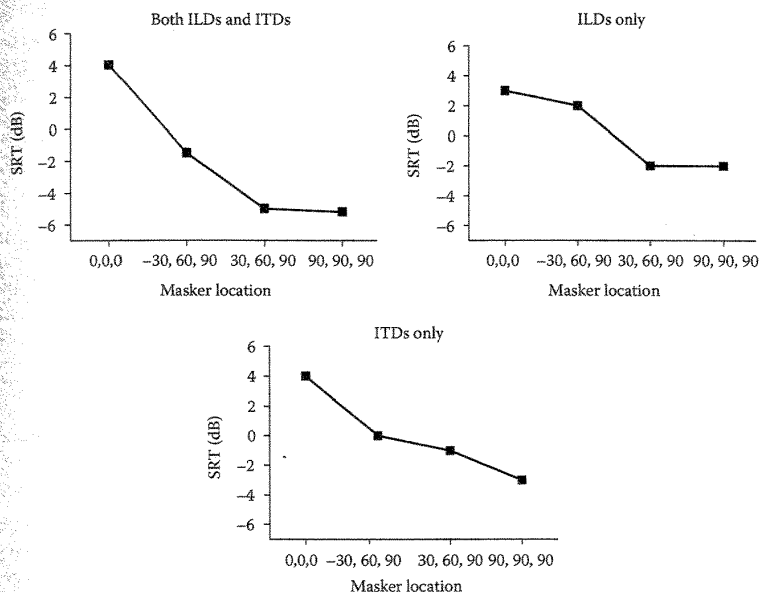


FIGURE 4.4 Binaural speech intelligibility as a function of the position of three different interfering talkers (maskers). Signals were processed so as to contain both ITD and ILD cues (left panel), or ILD cues only (right panel), or ITD cues only (bottom panel). Target signal was coming from the front of the listener, i.e., from 0° .

separation (i.e., the difference in SRT for speech and noise presented from the front, vs. speech and noise being spatially separated) is given by:

$$R = C \left[\alpha \left(1 - \frac{1}{N} \sum_{i=1}^N \cos \theta_i \right) + \beta \frac{1}{N} \left| \sum_{i=1}^N \sin \theta_i \right| \right] \quad (4.1)$$

where R is the predicted release from masking, C is a constant ($C = 0.85 - 1$) used to account for differences in speech material, α and β are constants, N is the number of noise sources, and θ_i is the azimuth of the i th noise source. Good fit with experimental data were found with $\alpha = 1.38$ and $\beta = 8.02$. The cosine term in Equation 4.1 estimates the average displacement contribution between the target (coming from the front) and each of the N noise sources, and the sine term estimates the symmetry contribution of the noise sources. The sine term assumes a smaller value when the noise sources are symmetrically located.

It should be noted that all the data discussed so far were obtained in anechoic conditions, i.e., in the absence of reverberation. In realistic environments there might be reverberation in the room, and speech intelligibility will deteriorate because: (1) the contribution of head shadow will be reduced, causing in turn a reduction in binaural release of masking, and (2) there will be a smaller gain from temporal fluctuations of the noise sources. A small number of studies were conducted in realistic listening conditions where reverberation might be present (e.g., [23] and [24]).

4.2 ACOUSTIC PROPERTIES OF SPEECH CONTRIBUTING TO ROBUSTNESS

The speech signal itself possesses a number of properties that shield it from external distortion. These properties are described next.

4.2.1 SHAPE OF THE SPEECH SPECTRUM

The shape of the long-term average speech spectrum can tell us a great deal about which frequency regions of the spectrum are most likely to be affected by distortion and which ones are most likely to be affected the least. Figure 4.5 shows the long-term average speech spectrum (LTASS) obtained from a large number of native speakers of 12 different languages [25]. The LTASS was estimated by filtering speech through 25 1/3-octave-wide filters spanning the range of 63–16000 Hz, and computing the root-mean-square (rms) level within each band. The 1/3-octave analysis was performed using 125-msec segments of speech and averaged over a 1-min recorded passage. Figure 4.5 shows separately the LTASS spectra obtained for male and female speakers. As one would expect, the difference between the male and female LTASS is most prominent in the very-low-frequency region, reflecting the lower fundamental frequency of male speakers.

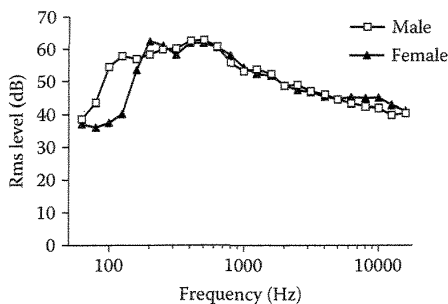


FIGURE 4.5 The long-term average speech spectrum from 10 adult males (open squares symbols) and 10 adult females (filled triangles symbols) for 12 different languages.

Most of the speech energy lies below 1 kHz, with a peak at 500 Hz. Speech energy gradually decreases starting at 500 Hz.

It is clear from Figure 4.5 that the LTASS has a low-frequency dominance. This low-frequency emphasis of the LTASS shields speech information from distortion and noise for several reasons:

1. With decreasing S/N levels, the low-frequency region is the last to be masked by noise. Generally, the low-frequency region will be masked to a lesser degree than the high-frequency region (see example in Figure 4.6). This holds for both white and pink-type of input noise. The effect, however, is more evident with white noise.
2. As a direct consequence of the above, listeners will have access to reliable F1 and partial F2 information [26]. Such information is critically important for vowel identification and stop consonant perception.
3. Speech harmonics falling in the low frequencies will be affected less than the high-frequency harmonics. Listeners will therefore have access to relatively reliable F0 cues needed for segregating speech from competing voices.
4. Neural phase-locking provides an effective means of temporally coding information pertaining to the lower formant frequencies [27], and is known to be preserved at the auditory nerve level up to 4–5 kHz.
5. Auditory frequency selectivity is sharpest in low frequencies and decreases at high frequencies [28]. This enables listeners to make fine F0 distinctions, important for segregating speech in multitalker environments.

Each of the preceding factors contributes to the robustness of speech information in noisy conditions.

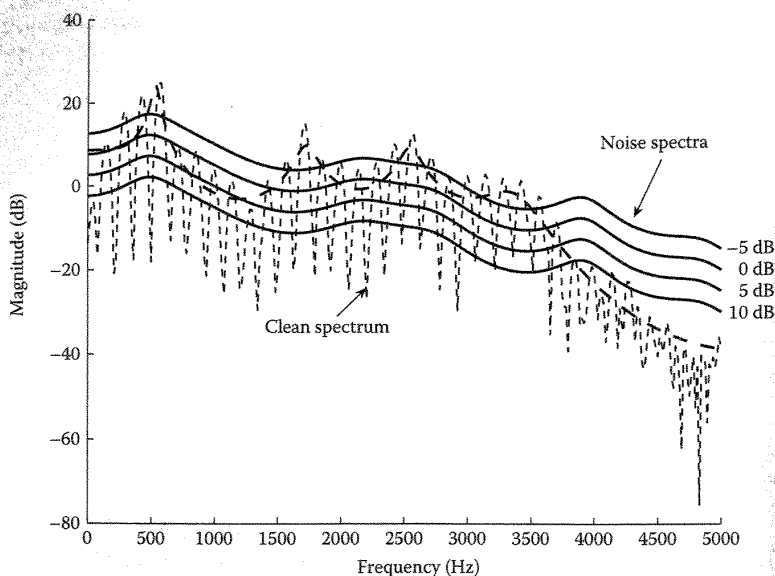


FIGURE 4.6 LPC spectra of the vowel /eh/ (male speaker, $F_0 = 144$ Hz) overlaid with the LPC spectra of multitalker babble scaled at different S/N ratios. The LPC spectra of the noise are shown prior to addition of the clean vowel at the indicated S/N ratios. Dashed lines show the clean FFT and LPC magnitude spectra.

4.2.2 SPECTRAL PEAKS

The low-frequency spectral peaks are more likely to be preserved in noise than the spectral valleys [26]. This is illustrated in Figure 4.7 which shows the linear predictive coding (LPC) spectra of the vowel /eh/ (as in “head”) corrupted in 5 to 10 dB multitalker babble. As can be seen, the frequency locations of the first three peaks, corresponding to the lower three formants (F_1 – F_3), do not change by much after adding noise. For this particular example, the F_1 peak location appears to shift from 691 Hz in quiet to 633 Hz in 0 dB, whereas the F_2 peak location appears to change from 1879 Hz in quiet to 1726 Hz in 0 dB. As one might expect, the effect of noise on the location of the spectral peaks will depend on the spectral characteristics of the noise, and this is demonstrated in Figure 4.8, which shows the FFT and LPC spectra of the vowel /eh/ corrupted in white noise at +5 and 0 dB S/N. As can be seen, white noise reduced the dynamic range and spectral contrast (peak-to-valley ratio) of the vowel, possibly to a greater extent compared to babble. However, the frequency location of the first peak corresponding to F_1 did not change by much even at 0 dB.

The frequencies of the first three formant (F_1 – F_3) peaks as well as their trajectory over time (i.e., the formant transitions) provide valuable cues to listeners

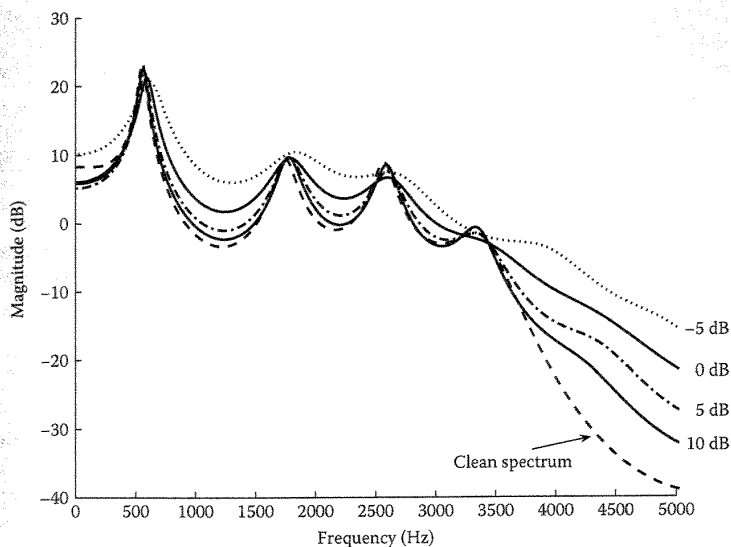


FIGURE 4.7 LPC spectra of the vowel /eh/ (same as in Figure 4.6) corrupted by multi-talker babble at -5 to 10 dB S/N.

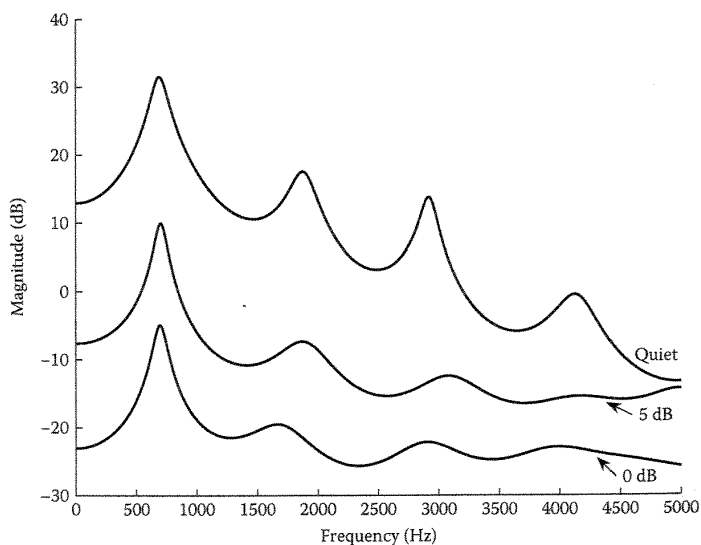


FIGURE 4.8 LPC spectra of the vowel /eh/ corrupted by white noise at 0 and 5 dB S/N. The noisy spectra were shifted down for clarity.

for vowel [29], glide [30], and stop-consonant perception [31]. As the examples in Figure 4.7 and Figure 4.8 demonstrated, the lowest two formant peaks are not affected in noise as much as the higher formant peaks. But to what extent is there reliable F1 and F2 information in noise?

The study in [26] showed that F1 is preserved to a certain degree in noise. Figure 4.9 plots the percent of frames for which F1 and F2 were reliably detected in babble as a function of S/N ratio. This data was based on acoustic analysis of

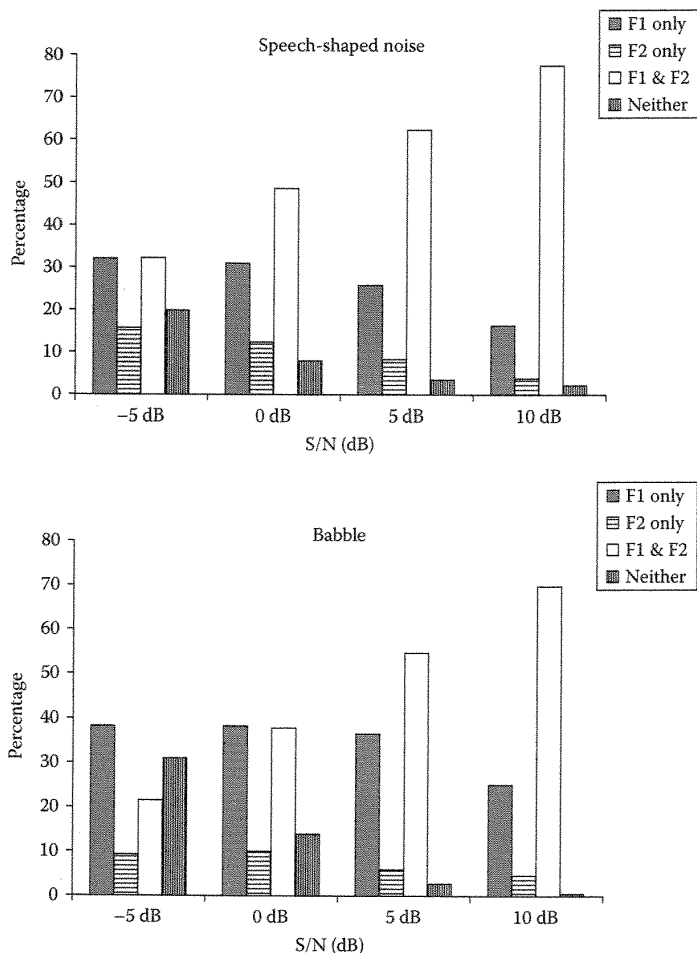


FIGURE 4.9 Percentage of frames for which F1 and F2 were reliably detected in multi-talker babble and speech-shaped noise as a function of S/N ratio. (From Parikh, G. and Loizou, P. (2005), *J. Acoust. Soc. Am.*, 118(6), 3874–3888. With permission.)

11 vowels produced in /hVd/ context by 23 different female speakers and 20 male speakers. Acoustic analysis indicated that F1 was detected more reliably in noise than F2. In multitalker babble at -5 dB S/N, for instance, F1 was identified 60% of the time whereas F2 was identified only 30% of the time (these values were obtained by summing the percentages of "F1 only" and "F1&F2" in Figure 4.9). The results from [26] suggest that in extremely noisy conditions, listeners are able to identify vowels based on a relatively accurate F1 representation and partial F2 representation.

In summary, noise may distort the shape of the spectrum, change its slope, and reduce the spectral dynamic range (see Figure 4.8), but the frequency locations of the lower formant peaks are preserved to some degree (see example in Figure 4.7).

4.2.3 PERIODICITY

As described in Chapter 3, the opening and closing of the vocal folds during voicing produces periodic waveforms (voiced segments of speech). The time duration of one cycle of the vocal folds' opening or closing is known as the fundamental period, and the reciprocal of the fundamental period is known as the vocal pitch or fundamental frequency (F0). The fundamental frequency F0 varies from a low of around 80 Hz for male speakers to a high of 280 Hz for children [29]. The presence or absence of periodicity signifies the distinction between voiced and unvoiced sounds (e.g., between [d] and [t]). The F0 periodicity is also responsible for the perception of vocal pitch, intonation, prosody, and perception of lexical tone in tonal languages.

The voiced segments of speech (e.g., vowels) are quasi-periodic in the time domain and harmonic in the frequency domain. The periodicity of speech is broadly distributed across frequency and time and is robust in the presence of noise. Figure 4.10 shows the FFT spectra of the vowel /eh/ in quiet and in noise. It is clear that the lower harmonics are preserved in noise, at least up to 1 kHz. This suggests that listeners have access to relatively accurate F0 information in noise. Such information, as we will discuss later (Section 4.3.3), is important for understanding speech in situations where two or more people are speaking simultaneously [32].

4.2.4 RAPID SPECTRAL CHANGES SIGNALING CONSONANTS

Unlike vowels, consonants are short in duration and have low intensity. Consequently, they are more vulnerable to noise or distortion compared to vowels. The duration of the /b/ burst for instance can be as brief as 5–10 msec. Vowels on the other hand can last as long as 300 msec [33].

Formant transitions associated with vowel or diphthong production are slow and gradual. In contrast, consonants (particularly stop consonants) are associated with rapid spectral changes and rapid formant transitions. In noise, these rapid spectral changes are preserved to some degree and serve as landmarks signaling the presence of consonants. Figure 4.11 shows example spectrograms of a sentence

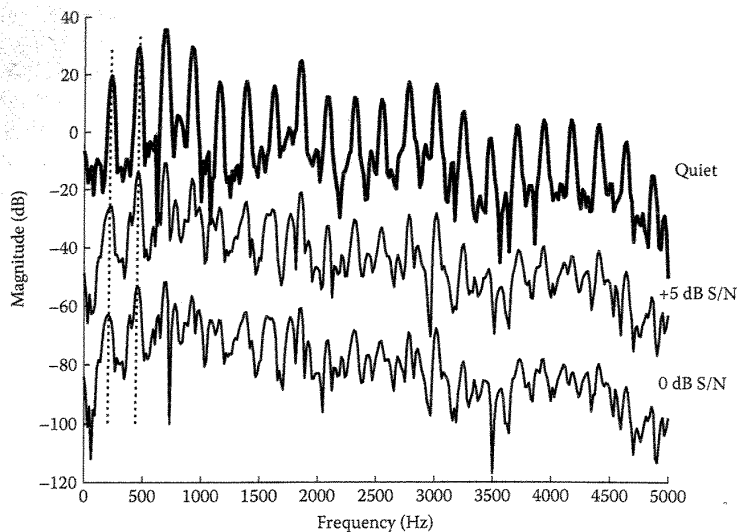


FIGURE 4.10 Example FFT-magnitude spectra of the vowel /eh/ corrupted by multitalker babble at 0 and 5 dB S/N. The noisy magnitude spectra were shifted down for better clarity.

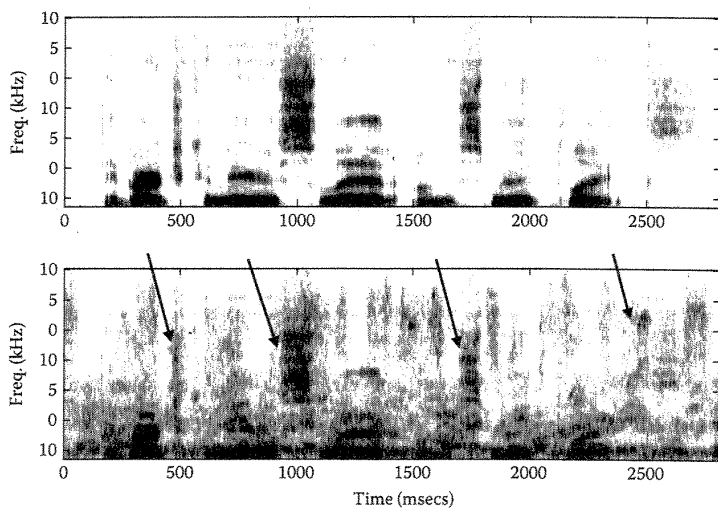


FIGURE 4.11 Spectrogram of the sentence "The birch canoe slid on the smooth planks" in quiet (top panel) and in +5 dB babble (bottom panel). Arrows indicate the onsets of some of the high-frequency consonants.

embedded in +5 dB babble noise. Note that the low-frequency (and intense) vowels alternate frequently with the high-frequency (and weak) consonants, resulting in sudden spectral changes. For instance, the low-frequency spectral dominance at near 500, 1000, and 1700 msec is followed by a sudden dispersion of spectral energy across the spectrum. These sudden spectral changes coincide with the onsets of consonants. Although for the most part the high-frequency information is smeared and heavily masked by the noise, the onsets of most of the consonants are preserved (see arrows in Figure 4.11). There is evidence from neurophysiologic studies (e.g., [34]), suggesting that peripheral or central adaptation mechanisms exist that improve the prominence of spectral components with sudden onsets. The response of auditory nerve fibers to tone bursts is maximal at the onset of the stimulus, and gradually decays to its steady-state value.

4.3 PERCEPTUAL STRATEGIES FOR LISTENING IN NOISE

In a previous section (Section 4.1), we described how listeners are able to carry a conversation in the most adverse conditions, which might include highly reverberant rooms with multiple speakers in the background. But, how can the auditory system achieve such a feat? For one, the speech signal itself is shielded in many ways from external distortion (see Section 4.2). Furthermore, human listeners utilize several perceptual strategies that facilitate speech segregation. These listening strategies are discussed next.

4.3.1 AUDITORY STREAMING

Consider a mixture of sounds consisting of speech and other sources, including, say, the sound of the wind blowing through the window, cars passing by, a telephone ringing, etc. Bregman [2] proposed that humans can analyze such an auditory scene and are able to segregate speech from the other sources by grouping together components of a sound mixture according to their sources. He referred to the components belonging to the same source as *auditory streams*.

Two different types of groupings were suggested as necessary to segregate speech by: (1) linking events that follow one another in time (sequential grouping), and (2) integrating components that occur at the same time in different regions of the spectrum (simultaneous grouping). Sequential grouping can be done by tracking characteristics of the speech signal that vary slowly over time. For instance, listeners might exploit trajectories of formant peaks that vary slowly over time and frequency to track a voice in background noise. Simultaneous grouping can be done by exploiting differences in pitch (F_0) between competing voices, differences in the shape of the spectrum envelope, and differences in the spatial location of the sources. In running speech, pitch (F_0) is always changing, and two voices are not likely to follow exactly the same pitch contour at the same time. Segregation based on F_0 can be done, for instance, by looking for one or

more harmonics that can account for as many of the frequency components in the spectrum as possible, and then allocating all the components belonging to the same fundamental frequency (F_0) to the same auditory stream. This suggests a *harmonic sieve* mechanism that excludes harmonic components whose frequencies do not correspond to integer multiples of a given fundamental frequency. Such a mechanism was proposed as a plausible model of pitch perception [35].

According to Bregman [2], the foregoing two grouping mechanisms are governed by primitive auditory constraints such as the constraint that most speech characteristics (e.g., F_0 , formants) tend to be continuous and change slowly over time. They are called primitive in that they probably exist from infancy, i.e., they are innate, and do not rely on higher cognitive functions. Primitive auditory constraints utilize a bottom-up approach to auditory scene analysis. Bregman [2, Chapter 4] further proposed that auditory scene analysis is governed by learned constraints, which he called schema-based constraints. These constraints utilize a top-down approach to auditory scene analysis. The schema-driven segregation process is presumed to involve the activation of stored (learned) knowledge of familiar patterns in the acoustic environment (e.g., sounds of musical instruments, bells ringing, etc.), including learned knowledge of the language (e.g., language structure, semantics, grammar). Evidence of schema-based segregation is provided by studies of sine-wave speech [36]. When the formant trajectories of a sentence are replaced by sine waves following the formant frequencies, listeners can frequently recognize the words in the sentence, despite the fact that the synthesized sentence sounds unnatural, synthetic, and robotic. Listeners must be utilizing speech-specific knowledge about production to be able to fuse the perceived formant tracks into a single auditory stream. Such knowledge is acquired from extended exposure to the acoustic properties of speech.

4.3.2 LISTENING IN THE GAPS AND GLIMPSES

As mentioned earlier, there is evidence from several studies confirming that it is more difficult to identify words in the presence of steady-state noise than in the presence of a single competing speaker. The difference in SRT (50% intelligibility level) can be large, ranging from a low of 6 dB [5] to a high of 10 dB [4]. This difference is attributed to the fact that listeners are exploiting the silent gaps or waveform "valleys" in the competing signal to recognize the words in the target sentence. These gaps or amplitude valleys can potentially enable listeners to "glimpse" entire syllables or words of the target voice. During the gaps in the competing signal, the local (segmental) SNR is quite favorable. Clearly, the local SNR will be fluctuating over time, yet the listeners are able to exploit the local high-SNR segments of the noisy signal and extract (glimpse) useful information. Note that a similar situation occurs when the noise and speech signals are spatially separated. Listeners are able to utilize the "better" ear, i.e., the ear with the highest SNR, to recognize the target voice. Humans possess a remarkable ability to exploit segments of noisy speech with highest SNR in both monaural and binaural listening situations.

The listening-in-the-gaps account of speech segregation falls apart, however, when there are large numbers (more than 4) of competing voices present because the masker waveform becomes nearly continuous, leaving no silent gaps in the waveform (see Figure 4.1). In such situations, different listening strategies are probably employed.

A different view of glimpsing was proposed by Cooke [37,38], extending and generalizing the aforementioned idea of listening in the gaps. This new view was based on a different definition of what constitutes a *glimpse*: "a time-frequency region which contains a reasonably undistorted 'view' of local signal properties" [38]. Useful signal properties may include signal energy or presence of reliable F0 and formant frequency information. Glimpses of speech in background noise may, for instance, comprise all time-frequency areas or slices having a local SNR exceeding a certain threshold value (e.g., 3 dB). The assumption is that the listeners are able to detect "useful" glimpses of speech, which are possibly occurring at different times and occupying different regions of the spectrum, and are able to integrate all these glimpses (as different pieces of the puzzle) to recognize the target speech. Computational models of glimpsing were developed for robust automatic speech recognition by modifying the recognition process to allow for the possibility of "missing data" [39,40]. Despite the attractive appeal of glimpsing as a means of speech separation, there still remain several issues to be resolved. Foremost among those issues is the question of what constitutes a "useful" glimpse and whether glimpses contain sufficient information to support identification of the target signal.

4.3.3 USE OF F0 DIFFERENCES

It is generally accepted that it is easier to recognize competing speech when the target and masker voices are of opposite sex. The difference in F0 between male and female voices could be as large as 100 Hz [29]. Several studies have demonstrated that the difference in fundamental frequencies of the masker and target voices is critically important for recognizing speech in competing voices. Brokx and Nootboom [32] first demonstrated this by artificially modifying the speech of two talkers to have a monotonous (i.e., flat F0 contour) pitch and then synthesizing the monotonous sentences with an LPC vocoder. They varied the F0 difference between the target and masker voices from 0 to 12 semitones (1 octave). Results indicated that word identification accuracy was the lowest when the target and masker had the same F0 and gradually improved with increasing differences in F0 (see Figure 4.12). A mere difference of 3 semitones (F0 = 100 Hz vs. F0 = 120 Hz) was sufficient to produce a significant increase in the number of words identified from 40% correct to 60% correct. Identification accuracy was low when the pitch of the two voices was exactly one octave apart, which is the condition where every second harmonic of the higher-pitched voice coincides with a harmonic of the lower pitched voice. These findings were replicated and extended by Bird and Darwin [41], who showed

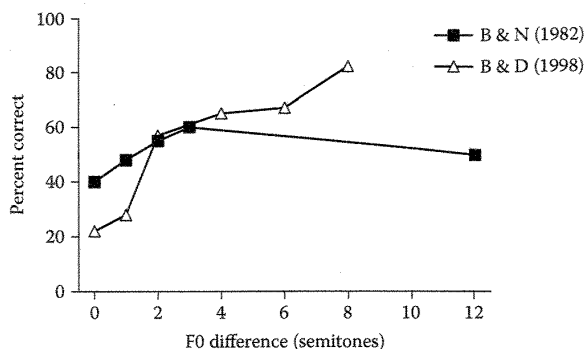


FIGURE 4.12 Results from two studies (B&N [32] and B&D [41]) on intelligibility of sentences corrupted by a single interfering talker as a function of the F0 difference of the target and interfering talkers. The F0 difference is expressed in semitones with 12 semitones equaling one octave.

that identification accuracy continues to improve up to an F0 difference of 8 semitones (see Figure 4.12).

A similar outcome has emerged from studies examining the contribution of F0 in the perception of “double vowels” [42] — pairs of vowels played simultaneously on different fundamentals (F0). The listeners’ task is to identify both vowels. Several studies demonstrated that performance is above chance level even when the two vowels have the same F0 [42–44]. Identification improves significantly, however, as the F0 difference is increased from 0 to 2 semitones, but does not improve further beyond 2 semitones. Overall, the contribution of F0 is greater and more evident for “double sentences” [32,41] than for “double vowels” [43].

The separation of two voices on the basis of fundamental frequency is believed to be mediated by simultaneous grouping mechanisms. Harmonics corresponding to the same F0 are presumably grouped together to form a single auditory stream. Such a grouping, for example, can potentially aid the listeners to identify and track (over time) formant peaks belonging to the target speaker. A number of algorithms have used this principle for automatic separation of two voices [45–47].

4.3.4 USE OF LINGUISTIC KNOWLEDGE

Intelligibility of speech in adverse conditions is greatly influenced by the ability of listeners to exploit linguistic knowledge, particularly when the cues contained in the acoustic signal are weak. The presence of semantic context, in particular, can lead to higher intelligibility of sentences as it restricts the range of plausible words [48].

Early experiments by Treisman [49] demonstrated that the linguistic content of the interfering message can greatly influence speech recognition. Competing

speech in the same language and similar in content as the target speech was found to be the most disruptive. Least disruptive was competing speech spoken in an unfamiliar foreign language. Access to learned knowledge about the language provides linguistic constraints useful in a schema-based perceptual segregation process [2].

4.3.5 USE OF SPATIAL AND VISUAL CUES

In real-world environments (e.g., in a crowded restaurant), competing voices originate from different locations in space. The acoustic signals arriving at the two ears of the listener will have different intensities (ILDs) and different times of arrival (ITDs). The ILD and ITD cues are used by listeners in spatial segregation of voices. The ILD cues alone, by the virtue of providing access to a "better ear," provide a large benefit in speech intelligibility provided the noise sources are not symmetrically located around the listener [20]. The ITD cues alone are effective cues for tracking a particular sound over time, but are not effective for segregating the speech of a single talker from competing talkers [50]. In concert, the ILD and ITD cues provide a large binaural advantage in situations where the speech and noise sources are spatially separated [12].

In real-world environments, and particularly in face-to-face communication settings, listeners also have access to visual cues. The visual cues provide synchronization between the acoustic signal and the movements of the articulators (e.g., tongue, lip). Having access to such visual information can aid in discrimination of certain confusable consonants (e.g., /p/ vs. /t/, /m/ vs. /n/) differing in place of articulation. Discrimination of /t/ vs. /p/, for instance, can be difficult in noise in the absence of visual cues, but it becomes extremely easy if the listener watches the lips of the target speaker. It is relatively easy to identify the labial consonants (e.g., /p/, /b/) because these consonants are produced with the lips closed.

Several studies have demonstrated the intelligibility benefit gained by integrating acoustic and visual information [51]. In [51], sentence recognition in noise was assessed in both auditory-only and audio-visual conditions, using sentences ranging in lip reading difficulty from easy to hard. Results indicated that the overall benefit conferred by visual cues was equivalent to an effective improvement in the SNR of 11 dB. As expected, the better lip readers benefited more than the poor lip readers. Other studies also demonstrated a significant benefit in intelligibility even when limited amounts of acoustic information supplemented speech reading [52].

4.4 SUMMARY

The present chapter provided a brief overview of auditory scene analysis by human listeners. Human listeners (with normal hearing) are able to carry on a conversation in the most adverse conditions for two main reasons. First, the information in speech is shielded from distortion in several ways (see Section 4.2).

More precisely, the speech signal possesses a number of properties (e.g., low-pass shape of speech spectrum, spectral peaks) that protect the information contained in the signal from external distortion and interference. In the context of noise reduction algorithms, this suggests that these properties (e.g., spectral peaks, F0 information) need to be preserved to avoid any degradation in speech intelligibility. Second, multiple listening strategies are used by listeners to accommodate various listening scenarios where not all cues might necessarily be present (Section 4.3). In the absence of visual and spatial cues (e.g., listening via the telephone), for instance, listeners must be using alternative listening strategies (e.g., F0 differences between the target speech and competing talkers, linguistic constraints, glimpsing, etc.) to communicate in noisy environments.

The area of auditory scene analysis is vast and remains an active topic of research not only for scientists interested in cognition but also for people interested in developing computational models for automatic speech separation. This chapter has barely scratched the surface of this topic. For more information on auditory scene analysis, the interested reader is referred to reviews [3,14,53–56] as well as books [2,57].

REFERENCES

1. Cherry, E. (1953), Some experiments on the recognition of speech with one and with two ears, *J. Acoust. Soc. Am.*, 25, 975–979.
2. Bregman, A. (1990), *Auditory Scene Analysis*, Cambridge, MA: MIT Press.
3. Assmann, P. and Summerfield, Q. (2004), The perception of speech under adverse conditions, in Greenberg, S., Ainsworth, W., Popper, A., and Fay, R. (Eds.), *Speech Processing in the Auditory System*, New York: Springer Verlag, pp. 231–308.
4. Miller, G. (1947), The masking of speech, *Psychol. Bull.*, 44(2), 105–129.
5. Festen, J. and Plomp, R. (1990), Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing, *J. Acoust. Soc. Am.*, 88, 1725–1736.
6. Levitt, H. and Rabiner, L. (1967), Binaural release from masking for speech and gain in intelligibility, *J. Acoust. Soc. Am.*, 42(3), 601–608.
7. Carhart, R., Tillman, T., and Greetis, E. (1969), Perceptual masking in multiple sound backgrounds, *J. Acoust. Soc. Am.*, 45(3), 694–703.
8. Duquesnoy, A. (1983), Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons, *J. Acoust. Soc. Am.*, 74(3), 739–743.
9. Drullman, R. and Bronkhorst, A. (2000), Multichannel speech intelligibility and speaker recognition using monaural, binaural, and 3D auditory presentation, *J. Acoust. Soc. Am.*, 107(4), 2224–2235.
10. Brungart, D. (2001), Informational and energetic masking effects in the perception of two simultaneous talkers, *J. Acoust. Soc. Am.*, 109(3), 1101–1109.
11. Bronkhorst, A. and Plomp, R. (1992), Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing, *J. Acoust. Soc. Am.*, 92(6), 3132–3139.
12. Hawley, M., Litovsky, R., and Culling, J. (2004), The benefit of binaural hearing in a cocktail party: effect of location and type of interferer, *J. Acoust. Soc. Am.*, 115(2), 833–843.

13. Peissig, J. and Kollmeier, B. (1997), Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners, *J. Acoust. Soc. Am.*, 101(3), 1660–1670.
14. Bronkhorst, A. (2000), The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions, *Acustica*, 86, 117–128.
15. Hawley, M., Litovsky, R., and Colburn, H. (1999), Intelligibility and localization of speech signals in a multi-source environment, *J. Acoust. Soc. Am.*, 105, 3436–3448.
16. Yost, W., Dye, R., and Sheft, S. (1996), A simulated “cocktail party” with up to three sound sources, *Percept. Psychophys.*, 58(7), 1026–1036.
17. Blauert, J., Brueggen, M., Bronkhorst, A., Drullman, R., Reynaud, G., Pellieux, L., Krebber, W., and Sottek, R. (1998), The AUDIS catalog of human HRTFs, *J. Acoust. Soc. Am.*, 103, 3082.
18. IEEE Subcommittee (1969), IEEE Recommended Practice for Speech Quality Measurements, *IEEE Trans. Audio Electroacoust.*, AU-17(3), 225–246.
19. Noble, W. and Perrett, S. (2002), Hearing speech against spatially separate competing speech versus competing voice, *Percept. Psychophys.*, 64(8), 1325–1336.
20. Bronkhorst, A. and Plomp, R. (1988), The effect of head-induced interaural time and level differences on speech intelligibility in noise, *J. Acoust. Soc. Am.*, 83(4), 1508–1516.
21. Culling, J., Hawley, M., and Litovsky, R. (2004), The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources, *J. Acoust. Soc. Am.*, 116(2), 1057–1065.
22. Levitt, H. and Rabiner, L. (1967), Predicting binaural gain in intelligibility and release from masking for speech, *J. Acoust. Soc. Am.*, 42(4), 820–829.
23. Nabelek, A. and Robinson, P. (1982), Monaural and binaural speech perception in reverberation for listeners of various ages, *J. Acoust. Soc. Am.*, 71, 1242–1248.
24. Moncur, J. and Dirks, D. (1967), Binaural and monaural speech intelligibility in reverberation, *J. Speech Hear. Res.*, 10, 186–195.
25. Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Nakatishi, Y., Oyer, H., Powell, R., Stepfus, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G., Westerman, S., and Ludvigsen, C. (1994), An international comparison of long-term average speech spectra, *J. Acoust. Soc. Am.*, 96, 2108–2120.
26. Parikh, G. and Loizou, P. (2005), The influence of noise on vowel and consonant cues, *J. Acoust. Soc. Am.*, 118(6), 3874–3888.
27. Young, E. and Sachs, M. (1979), Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers, *J. Acoust. Soc. Am.*, 66(5), 1381–1403.
28. Patterson, R. and Moore, B. (1986), Auditory filters and excitation patterns as representations of auditory frequency selectivity, in Moore, B., *Frequency Selectivity in Hearing*, London: Academic Press.
29. Peterson, G. and Barney, H. (1952), Control methods used in a study of the vowels, *J. Acoust. Soc. Am.*, 24(2), 175–184.
30. O'Connor, J., Gerstman, L., Liberman, A., Delattre, P., and Cooper, F. (1957), Acoustic cues for the perception of initial /w, j, r, l/ in English, *Word*, 13, 24–43.
31. Delattre, P., Liberman, A., and Cooper, F. (1955), Acoustic loci and transitional cues for consonants, *J. Acoust. Soc. Am.*, 27, 769–774.
32. Brokx, J. and Nootboom, S. (1982), Intonation and perception of simultaneous voices, *J. Phonetics*, 10, 23–26.

33. Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995), Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.*, 97(5), 3099–3111.
34. Delgutte, B. (1980), Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers, *J. Acoust. Soc. Am.*, 68(3), 843–857.
35. Duifhuis, H., Willems, L., and Sluyter, R. (1982), Measurement of pitch on speech: an implementation of Goldstein's theory of pitch perception, *J. Acoust. Soc. Am.*, 71, 1568–1580.
36. Remez, R., Rubin, P., Pisoni, D., and Carrell T., (1981), Speech perception without traditional cues, *Science*, 212, pp. 947–950.
37. Cooke, M. (2003), Glimpsing speech, *J. Phonetics*, 31, 579–584.
38. Cooke, M. (2005), Making sense of everyday speech: a glimpsing account, in Divenyi, P. (Ed.), *Speech Separation by Humans and Machines*, Kluwer Academic Publishers, pp. 305–314.
39. Cooke, M., Green, P., and Crawford, M. (1994), Handling missing data in speech recognition, *Proc. Int. Conf. Spoken Lang.*, pp. 1555–1558.
40. Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001), Robust automatic speech recognition with missing and uncertain acoustic data, *Speech Commun.*, 34, 267–285.
41. Bird, J. and Darwin, C. (1998), Effects of a difference in fundamental frequency in separating two sentences, in Palmer, A., Rees, A., Summerfield, Q., and Meddis, R. (Eds.), *Psychophysical and Physiological Advances in Hearing*, London: Whurr.
42. Scheffers, M. (1983), Sifting Vowels: Auditory Pitch Analysis and Sound Segregation, Ph.D. thesis, Rijksuniversiteit te Groningen, Netherlands.
43. Assmann, P. and Summerfield, Q. (1990), Modeling the perception of concurrent vowels: vowels with different fundamental frequencies, *J. Acoust. Soc. Am.*, 88, 680–697.
44. Culling, J. and Darwin, C. (1993), Perceptual separation of simultaneous vowels: within and across-formant grouping by F0, *J. Acoust. Soc. Am.*, 93, 3454–3467.
45. Parsons, T. (1976), Separation of speech from interfering speech by means of harmonic selection, *J. Acoust. Soc. Am.*, 60, 656–660.
46. Stubbs, R. and Summerfield, Q. (1988), Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.*, 84(4), 1238–1249.
47. Denbigh, P. and Zhao, J. (1992), Pitch extraction and separation of overlapping speech, *Speech Commun.*, 11, 119–126.
48. Miller, G., Heise, G., and Lichten, W. (1951), The intelligibility of speech as a function of the context of the test materials, *J. Exper. Psych.*, 41, 329–335.
49. Treisman, A. (1964), Verbal cues, language, and meaning in selective attention, *Am. J. Psychol.*, 77(2), 206–219.
50. Culling, J. and Summerfield, Q. (1995), Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay, *J. Acoust. Soc. Am.*, 98, 785–797.
51. MacLeod, A. and Summerfield, Q. (1987), Quantifying the contribution of vision to speech perception in noise, *Br. J. Audiol.*, 21, 131–141.
52. Breeuwer, M. and Plomp, R. (1984), Speechreading supplemented with frequency-selective sound pressure information, *J. Acoust. Soc. Am.*, 76(3), 686–691.
53. Cooke, M. and Ellis, D. (2001), The auditory organization of speech and other sources in listeners and computational models, *Speech Commun.*, 35, 141–177.

54. Darwin, C. and Carlyon, R. (1995), Auditory grouping, in Moore, B. (Ed.), *Hearing*, San Diego, CA: Academic Press, pp. 387–424.
55. Moore, B. and Gockel, H. (2002), Factors influencing sequential stream segregation, *Acustica*, 88, 320–332.
56. Darwin, C. (1997), Auditory grouping, *Trends in Cognitive Science*, 1, 327–333.
57. Divenyi, P. (2005), *Speech Separation by Humans and Machines*, Norwell, MA: Kluwer Academic Publishers.

Electrical Engineering

SPEECH ENHANCEMENT

Theory and Practice

Philipos C. Loizou

The first book to provide comprehensive and up-to-date coverage of all major speech enhancement algorithms proposed in the last two decades, *Speech Enhancement: Theory and Practice* is a valuable resource for experts and newcomers in the field. The book covers traditional speech enhancement algorithms, such as spectral subtraction and Wiener filtering algorithms as well as state-of-the-art algorithms including minimum mean-squared error algorithms that incorporate signal-presence uncertainty and subspace algorithms that incorporate psychoacoustic models. The coverage includes objective and subjective measures used to evaluate speech quality and intelligibility.

Providing clear and concise coverage of the subject, the author brings together a large body of knowledge about how human listeners compensate for acoustic noise when in noisy environments. This book is a valuable resource not only for engineers who want to implement the latest speech enhancement algorithms but also for speech practitioners who want to incorporate some of these algorithms into hearing aid applications for speech intelligibility and quality improvement.

FEATURES

- Supplies up-to-date coverage of all major noise suppression algorithms
- Provides an understanding of the limitations and potential of existing enhancement algorithms
- Covers the fundamentals needed to understand speech enhancement algorithms
- Discusses all major enhancement algorithms as well as noise estimation algorithms
- Presents a description of the evaluation measures used to assess the performance of enhancement algorithms
- Elucidates the evaluation results obtained from a comparison between several algorithms in terms of speech quality and intelligibility
- Includes MATLAB[®] code for the implementation of major speech enhancement algorithms



CRC Press

Taylor & Francis Group
an informa business

www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

DK328X

ISBN 0-8493-5032-8

90000



9 780849 350320

www.crcpress.com