

DISTANT SPEECH RECOGNITION

Matthias Wölfel

Universität Karlsruhe (TH), Germany

and

John McDonough

Universität des Saarlandes, Germany



A John Wiley and Sons, Ltd., Publication

DISTANT SPEECH RECOGNITION

DISTANT SPEECH RECOGNITION

Matthias Wölfel

Universität Karlsruhe (TH), Germany

and

John McDonough

Universität des Saarlandes, Germany



A John Wiley and Sons, Ltd., Publication

This edition first published 2009

© 2009 John Wiley & Sons Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Wölfel, Matthias.

Distant speech recognition / Matthias Wölfel, John McDonough.
p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-51704-8 (cloth)

1. Automatic speech recognition. I. McDonough, John (John W.) II. Title.

TK7882.S65W64 2009

006.4'54 – dc22

2008052791

A catalogue record for this book is available from the British Library

ISBN 978-0-470-51704-8 (H/B)

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India

Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire

Contents

Foreword	xiii
Preface	xvii
1 Introduction	1
1.1 Research and Applications in Academia and Industry	1
1.1.1 <i>Intelligent Home and Office Environments</i>	2
1.1.2 <i>Humanoid Robots</i>	3
1.1.3 <i>Automobiles</i>	4
1.1.4 <i>Speech-to-Speech Translation</i>	6
1.2 Challenges in Distant Speech Recognition	7
1.3 System Evaluation	9
1.4 Fields of Speech Recognition	10
1.5 Robust Perception	12
1.5.1 <i>A Priori Knowledge</i>	12
1.5.2 <i>Phonemic Restoration and Reliability</i>	12
1.5.3 <i>Binaural Masking Level Difference</i>	14
1.5.4 <i>Multi-Microphone Processing</i>	14
1.5.5 <i>Multiple Sources by Different Modalities</i>	15
1.6 Organizations, Conferences and Journals	16
1.7 Useful Tools, Data Resources and Evaluation Campaigns	18
1.8 Organization of this Book	18
1.9 Principal Symbols used Throughout the Book	23
1.10 Units used Throughout the Book	25
2 Acoustics	27
2.1 Physical Aspect of Sound	27
2.1.1 <i>Propagation of Sound in Air</i>	28
2.1.2 <i>The Speed of Sound</i>	29
2.1.3 <i>Wave Equation and Velocity Potential</i>	29
2.1.4 <i>Sound Intensity and Acoustic Power</i>	31
2.1.5 <i>Reflections of Plane Waves</i>	32
2.1.6 <i>Reflections of Spherical Waves</i>	33

2.2	Speech Signals	34
2.2.1	<i>Production of Speech Signals</i>	34
2.2.2	<i>Units of Speech Signals</i>	36
2.2.3	<i>Categories of Speech Signals</i>	39
2.2.4	<i>Statistics of Speech Signals</i>	39
2.3	Human Perception of Sound	41
2.3.1	<i>Phase Insensitivity</i>	42
2.3.2	<i>Frequency Range and Spectral Resolution</i>	42
2.3.3	<i>Hearing Level and Speech Intensity</i>	42
2.3.4	<i>Masking</i>	44
2.3.5	<i>Binaural Hearing</i>	45
2.3.6	<i>Weighting Curves</i>	45
2.3.7	<i>Virtual Pitch</i>	46
2.4	The Acoustic Environment	47
2.4.1	<i>Ambient Noise</i>	47
2.4.2	<i>Echo and Reverberation</i>	48
2.4.3	<i>Signal-to-Noise and Signal-to-Reverberation Ratio</i>	51
2.4.4	<i>An Illustrative Comparison between Close and Distant Recordings</i>	52
2.4.5	<i>The Influence of the Acoustic Environment on Speech Production</i>	53
2.4.6	<i>Coloration</i>	54
2.4.7	<i>Head Orientation and Sound Radiation</i>	55
2.4.8	<i>Expected Distances between the Speaker and the Microphone</i>	57
2.5	Recording Techniques and Sensor Configuration	58
2.5.1	<i>Mechanical Classification of Microphones</i>	58
2.5.2	<i>Electrical Classification of Microphones</i>	59
2.5.3	<i>Characteristics of Microphones</i>	60
2.5.4	<i>Microphone Placement</i>	60
2.5.5	<i>Microphone Amplification</i>	62
2.6	Summary and Further Reading	62
2.7	Principal Symbols	63
3	Signal Processing and Filtering Techniques	65
3.1	Linear Time-Invariant Systems	65
3.1.1	<i>Time Domain Analysis</i>	66
3.1.2	<i>Frequency Domain Analysis</i>	69
3.1.3	<i>z-Transform Analysis</i>	72
3.1.4	<i>Sampling Continuous-Time Signals</i>	79
3.2	The Discrete Fourier Transform	82
3.2.1	<i>Realizing LTI Systems with the DFT</i>	85
3.2.2	<i>Overlap-Add Method</i>	86
3.2.3	<i>Overlap-Save Method</i>	87
3.3	Short-Time Fourier Transform	87
3.4	Summary and Further Reading	90
3.5	Principal Symbols	91

4	Bayesian Filters	93
4.1	Sequential Bayesian Estimation	95
4.2	Wiener Filter	98
4.2.1	<i>Time Domain Solution</i>	98
4.2.2	<i>Frequency Domain Solution</i>	99
4.3	Kalman Filter and Variations	101
4.3.1	<i>Kalman Filter</i>	101
4.3.2	<i>Extended Kalman Filter</i>	106
4.3.3	<i>Iterated Extended Kalman Filter</i>	107
4.3.4	<i>Numerical Stability</i>	108
4.3.5	<i>Probabilistic Data Association Filter</i>	110
4.3.6	<i>Joint Probabilistic Data Association Filter</i>	115
4.4	Particle Filters	121
4.4.1	<i>Approximation of Probabilistic Expectations</i>	121
4.4.2	<i>Sequential Monte Carlo Methods</i>	125
4.5	Summary and Further Reading	132
4.6	Principal Symbols	133
5	Speech Feature Extraction	135
5.1	Short-Time Spectral Analysis	136
5.1.1	<i>Speech Windowing and Segmentation</i>	136
5.1.2	<i>The Spectrogram</i>	137
5.2	Perceptually Motivated Representation	138
5.2.1	<i>Spectral Shaping</i>	138
5.2.2	<i>Bark and Mel Filter Banks</i>	139
5.2.3	<i>Warping by Bilinear Transform – Time vs Frequency Domain</i>	142
5.3	Spectral Estimation and Analysis	145
5.3.1	<i>Power Spectrum</i>	145
5.3.2	<i>Spectral Envelopes</i>	146
5.3.3	<i>LP Envelope</i>	147
5.3.4	<i>MVDR Envelope</i>	150
5.3.5	<i>Perceptual LP Envelope</i>	153
5.3.6	<i>Warped LP Envelope</i>	153
5.3.7	<i>Warped MVDR Envelope</i>	156
5.3.8	<i>Warped-Twice MVDR Envelope</i>	157
5.3.9	<i>Comparison of Spectral Estimates</i>	159
5.3.10	<i>Scaling of Envelopes</i>	160
5.4	Cepstral Processing	163
5.4.1	<i>Definition and Characteristics of Cepstral Sequences</i>	163
5.4.2	<i>Homomorphic Deconvolution</i>	166
5.4.3	<i>Calculating Cepstral Coefficients</i>	167
5.5	Comparison between Mel Frequency, Perceptual LP and warped MVDR Cepstral Coefficient Front-Ends	168
5.6	Feature Augmentation	169
5.6.1	<i>Static and Dynamic Parameter Augmentation</i>	169

5.6.2	<i>Feature Augmentation by Temporal Patterns</i>	171
5.7	Feature Reduction	171
5.7.1	<i>Class Separability Measures</i>	172
5.7.2	<i>Linear Discriminant Analysis</i>	173
5.7.3	<i>Heteroscedastic Linear Discriminant Analysis</i>	176
5.8	Feature-Space Minimum Phone Error	178
5.9	Summary and Further Reading	178
5.10	Principal Symbols	179
6	Speech Feature Enhancement	181
6.1	Noise and Reverberation in Various Domains	183
6.1.1	<i>Frequency Domain</i>	183
6.1.2	<i>Power Spectral Domain</i>	185
6.1.3	<i>Logarithmic Spectral Domain</i>	186
6.1.4	<i>Cepstral Domain</i>	187
6.2	Two Principal Approaches	188
6.3	Direct Speech Feature Enhancement	189
6.3.1	<i>Wiener Filter</i>	189
6.3.2	<i>Gaussian and Super-Gaussian MMSE Estimation</i>	191
6.3.3	<i>RASTA Processing</i>	191
6.3.4	<i>Stereo-Based Piecewise Linear Compensation for Environments</i>	192
6.4	Schematics of Indirect Speech Feature Enhancement	193
6.5	Estimating Additive Distortion	194
6.5.1	<i>Voice Activity Detection-Based Noise Estimation</i>	194
6.5.2	<i>Minimum Statistics Noise Estimation</i>	195
6.5.3	<i>Histogram- and Quantile-Based Methods</i>	196
6.5.4	<i>Estimation of the a Posteriori and a Priori Signal-to-Noise Ratio</i>	197
6.6	Estimating Convolutional Distortion	198
6.6.1	<i>Estimating Channel Effects</i>	199
6.6.2	<i>Measuring the Impulse Response</i>	200
6.6.3	<i>Harmful Effects of Room Acoustics</i>	201
6.6.4	<i>Problem in Speech Dereverberation</i>	201
6.6.5	<i>Estimating Late Reflections</i>	202
6.7	Distortion Evolution	204
6.7.1	<i>Random Walk</i>	204
6.7.2	<i>Semi-random Walk by Polyak Averaging and Feedback</i>	205
6.7.3	<i>Predicted Walk by Static Autoregressive Processes</i>	206
6.7.4	<i>Predicted Walk by Dynamic Autoregressive Processes</i>	207
6.7.5	<i>Predicted Walk by Extended Kalman Filters</i>	209
6.7.6	<i>Correlated Prediction Error Covariance Matrix</i>	210
6.8	Distortion Evaluation	211
6.8.1	<i>Likelihood Evaluation</i>	212
6.8.2	<i>Likelihood Evaluation by a Switching Model</i>	213
6.8.3	<i>Incorporating the Phase</i>	214
6.9	Distortion Compensation	215

6.9.1	<i>Spectral Subtraction</i>	215
6.9.2	<i>Compensating for Channel Effects</i>	217
6.9.3	<i>Distortion Compensation for Distributions</i>	218
6.10	Joint Estimation of Additive and Convolutional Distortions	222
6.11	Observation Uncertainty	227
6.12	Summary and Further Reading	228
6.13	Principal Symbols	229
7	Search: Finding the Best Word Hypothesis	231
7.1	Fundamentals of Search	233
7.1.1	<i>Hidden Markov Model: Definition</i>	233
7.1.2	<i>Viterbi Algorithm</i>	235
7.1.3	<i>Word Lattice Generation</i>	238
7.1.4	<i>Word Trace Decoding</i>	240
7.2	Weighted Finite-State Transducers	241
7.2.1	<i>Definitions</i>	241
7.2.2	<i>Weighted Composition</i>	244
7.2.3	<i>Weighted Determinization</i>	246
7.2.4	<i>Weight Pushing</i>	249
7.2.5	<i>Weighted Minimization</i>	251
7.2.6	<i>Epsilon Removal</i>	253
7.3	Knowledge Sources	255
7.3.1	<i>Grammar</i>	256
7.3.2	<i>Pronunciation Lexicon</i>	263
7.3.3	<i>Hidden Markov Model</i>	264
7.3.4	<i>Context Dependency Decision Tree</i>	264
7.3.5	<i>Combination of Knowledge Sources</i>	273
7.3.6	<i>Reducing Search Graph Size</i>	274
7.4	Fast On-the-Fly Composition	275
7.5	Word and Lattice Combination	278
7.6	Summary and Further Reading	279
7.7	Principal Symbols	281
8	Hidden Markov Model Parameter Estimation	283
8.1	Maximum Likelihood Parameter Estimation	284
8.1.1	<i>Gaussian Mixture Model Parameter Estimation</i>	286
8.1.2	<i>Forward–Backward Estimation</i>	290
8.1.3	<i>Speaker-Adapted Training</i>	296
8.1.4	<i>Optimal Regression Class Estimation</i>	300
8.1.5	<i>Viterbi and Label Training</i>	301
8.2	Discriminative Parameter Estimation	302
8.2.1	<i>Conventional Maximum Mutual Information Estimation</i> <i>Formulae</i>	302
8.2.2	<i>Maximum Mutual Information Training on Word Lattices</i>	306
8.2.3	<i>Minimum Word and Phone Error Training</i>	308

8.2.4	<i>Maximum Mutual Information Speaker-Adapted Training</i>	310
8.3	Summary and Further Reading	313
8.4	Principal Symbols	315
9	Feature and Model Transformation	317
9.1	Feature Transformation Techniques	318
9.1.1	<i>Vocal Tract Length Normalization</i>	318
9.1.2	<i>Constrained Maximum Likelihood Linear Regression</i>	319
9.2	Model Transformation Techniques	320
9.2.1	<i>Maximum Likelihood Linear Regression</i>	321
9.2.2	<i>All-Pass Transform Adaptation</i>	322
9.3	Acoustic Model Combination	332
9.3.1	<i>Combination of Gaussians in the Logarithmic Domain</i>	333
9.4	Summary and Further Reading	334
9.5	Principal Symbols	336
10	Speaker Localization and Tracking	337
10.1	Conventional Techniques	338
10.1.1	<i>Spherical Intersection Estimator</i>	339
10.1.2	<i>Spherical Interpolation Estimator</i>	341
10.1.3	<i>Linear Intersection Estimator</i>	342
10.2	Speaker Tracking with the Kalman Filter	345
10.2.1	<i>Implementation Based on the Cholesky Decomposition</i>	348
10.3	Tracking Multiple Simultaneous Speakers	351
10.4	Audio-Visual Speaker Tracking	352
10.5	Speaker Tracking with the Particle Filter	354
10.5.1	<i>Localization Based on Time Delays of Arrival</i>	356
10.5.2	<i>Localization Based on Steered Beamformer Response Power</i>	356
10.6	Summary and Further Reading	357
10.7	Principal Symbols	358
11	Digital Filter Banks	359
11.1	Uniform Discrete Fourier Transform Filter Banks	360
11.2	Polyphase Implementation	364
11.3	Decimation and Expansion	365
11.4	Noble Identities	368
11.5	Nyquist(M) Filters	369
11.6	Filter Bank Design of De Haan <i>et al.</i>	371
11.6.1	<i>Analysis Prototype Design</i>	372
11.6.2	<i>Synthesis Prototype Design</i>	375
11.7	Filter Bank Design with the Nyquist(M) Criterion	376
11.7.1	<i>Analysis Prototype Design</i>	376
11.7.2	<i>Synthesis Prototype Design</i>	377
11.7.3	<i>Alternative Design</i>	378

11.8	Quality Assessment of Filter Bank Prototypes	379
11.9	Summary and Further Reading	384
11.10	Principal Symbols	384
12	Blind Source Separation	387
12.1	Channel Quality and Selection	388
12.2	Independent Component Analysis	390
	12.2.1 <i>Definition of ICA</i>	390
	12.2.2 <i>Statistical Independence and its Implications</i>	392
	12.2.3 <i>ICA Optimization Criteria</i>	396
	12.2.4 <i>Parameter Update Strategies</i>	403
12.3	BSS Algorithms based on Second-Order Statistics	404
12.4	Summary and Further Reading	407
12.5	Principal Symbols	408
13	Beamforming	409
13.1	Beamforming Fundamentals	411
	13.1.1 <i>Sound Propagation and Array Geometry</i>	411
	13.1.2 <i>Beam Patterns</i>	415
	13.1.3 <i>Delay-and-Sum Beamformer</i>	416
	13.1.4 <i>Beam Steering</i>	421
13.2	Beamforming Performance Measures	426
	13.2.1 <i>Directivity</i>	426
	13.2.2 <i>Array Gain</i>	428
13.3	Conventional Beamforming Algorithms	430
	13.3.1 <i>Minimum Variance Distortionless Response Beamformer</i>	430
	13.3.2 <i>Array Gain of the MVDR Beamformer</i>	433
	13.3.3 <i>MVDR Beamformer Performance with Plane Wave Interference</i>	433
	13.3.4 <i>Superdirective Beamformers</i>	437
	13.3.5 <i>Minimum Mean Square Error Beamformer</i>	439
	13.3.6 <i>Maximum Signal-to-Noise Ratio Beamformer</i>	441
	13.3.7 <i>Generalized Sidelobe Canceler</i>	441
	13.3.8 <i>Diagonal Loading</i>	445
13.4	Recursive Algorithms	447
	13.4.1 <i>Gradient Descent Algorithms</i>	448
	13.4.2 <i>Least Mean Square Error Estimation</i>	450
	13.4.3 <i>Recursive Least Squares Estimation</i>	455
	13.4.4 <i>Square-Root Implementation of the RLS Beamformer</i>	461
13.5	Nonconventional Beamforming Algorithms	465
	13.5.1 <i>Maximum Likelihood Beamforming</i>	466
	13.5.2 <i>Maximum Negentropy Beamforming</i>	471
	13.5.3 <i>Hidden Markov Model Maximum Negentropy Beamforming</i>	477
	13.5.4 <i>Minimum Mutual Information Beamforming</i>	480
	13.5.5 <i>Geometric Source Separation</i>	487

13.6	Array Shape Calibration	488
13.7	Summary and Further Reading	489
13.8	Principal Symbols	491
14	Hands On	493
14.1	Example Room Configurations	494
14.2	Automatic Speech Recognition Engines	496
14.3	Word Error Rate	498
14.4	Single-Channel Feature Enhancement Experiments	499
14.5	Acoustic Speaker-Tracking Experiments	501
14.6	Audio-Video Speaker-Tracking Experiments	503
14.7	Speaker-Tracking Performance vs Word Error Rate	504
14.8	Single-Speaker Beamforming Experiments	505
14.9	Speech Separation Experiments	507
14.10	Filter Bank Experiments	508
14.11	Summary and Further Reading	509
	Appendices	511
A	List of Abbreviations	513
B	Useful Background	517
B.1	Discrete Cosine Transform	517
B.2	Matrix Inversion Lemma	518
B.3	Cholesky Decomposition	519
B.4	Distance Measures	519
B.5	Super-Gaussian Probability Density Functions	521
	<i>B.5.1 Generalized Gaussian pdf</i>	521
	<i>B.5.2 Super-Gaussian pdfs with the Meier G-function</i>	523
B.6	Entropy	528
B.7	Relative Entropy	529
B.8	Transformation Law of Probabilities	529
B.9	Cascade of Warping Stages	530
B.10	Taylor Series	530
B.11	Correlation and Covariance	531
B.12	Bessel Functions	531
B.13	Proof of the Nyquist–Shannon Sampling Theorem	532
B.14	Proof of Equations (11.31–11.32)	532
B.15	Givens Rotations	534
B.16	Derivatives with Respect to Complex Vectors	537
B.17	Perpendicular Projection Operators	540
	Bibliography	541
	Index	561

Foreword

As the authors of *Distant Speech Recognition* note, automatic speech recognition is the key enabling technology that will permit natural interaction between humans and intelligent machines. Core speech recognition technology has developed over the past decade in domains such as office dictation and interactive voice response systems to the point that it is now commonplace for customers to encounter automated speech-based intelligent agents that handle at least the initial part of a user query for airline flight information, technical support, ticketing services, *etc.* While these limited-domain applications have been reasonably successful in reducing the costs associated with handling telephone inquiries, their fragility with respect to acoustical variability is illustrated by the difficulties that are experienced when users interact with the systems using speakerphone input. As time goes by, we will come to expect the range of natural human-machine dialog to grow to include seamless and productive interactions in contexts such as humanoid robotic butlers in our living rooms, information kiosks in large and reverberant public spaces, as well as intelligent agents in automobiles while traveling at highway speeds in the presence of multiple sources of noise. Nevertheless, this vision cannot be fulfilled until we are able to overcome the shortcomings of present speech recognition technology that are observed when speech is recorded at a distance from the speaker.

While we have made great progress over the past two decades in core speech recognition technologies, the failure to develop techniques that overcome the effects of acoustical variability in homes, classrooms, and public spaces is the major reason why automated speech technologies are not generally available for use in these venues. Consequently, much of the current research in speech processing is directed toward improving robustness to acoustical variability of all types. Two of the major forms of environmental degradation are produced by additive noise of various forms and the effects of linear convolution. Research directed toward compensating for these problems has been in progress for more than three decades, beginning with the pioneering work in the late 1970s of Steven Boll in noise cancellation and Thomas Stockham in homomorphic deconvolution.

Additive noise arises naturally from sound sources that are present in the environment in addition to the desired speech source. As the speech-to-noise ratio (SNR) decreases, it is to be expected that speech recognition will become more difficult. In addition, the impact of noise on speech recognition accuracy depends as much on the type of noise source as on the SNR. While a number of statistical techniques are known to be reasonably effective in dealing with the effects of quasi-stationary broadband additive noise of arbitrary spectral coloration, compensation becomes much more difficult when the noise is highly transient

in nature, as is the case with many types of impulsive machine noise on factory floors and gunshots in military environments. Interference by sources such as background music or background speech is especially difficult to handle, as it is both highly transient in nature and easily confused with the desired speech signal.

Reverberation is also a natural part of virtually all acoustical environments indoors, and it is a factor in many outdoor settings with reflective surfaces as well. The presence of even a relatively small amount of reverberation destroys the temporal structure of speech waveforms. This has a very adverse impact on the recognition accuracy that is obtained from speech systems that are deployed in public spaces, homes, and offices for virtually any application in which the user does not use a head-mounted microphone. It is presently more difficult to ameliorate the effects of common room reverberation than it has been to render speech systems robust to the effects of additive noise, even at fairly low SNRs. Researchers have begun to make progress on this problem only recently, and the results of work from groups around the world have not yet congealed into a clear picture of how to cope with the problem of reverberation effectively and efficiently.

Distant Speech Recognition by Matthias Wölfel and John McDonough provides an extraordinarily comprehensive exposition of the most up-to-date techniques that enable robust distant speech recognition, along with very useful and detailed explanations of the underlying science and technology upon which these techniques are based. The book includes substantial discussions of the major sources of difficulties along with approaches that are taken toward their resolution, summarizing scholarly work and practical experience around the world that has accumulated over decades. Considering both single-microphone and multiple-microphone techniques, the authors address a broad array of approaches at all levels of the system, including methods that enhance the waveforms that are input to the system, methods that increase the effectiveness of features that are input to speech recognition systems, as well as methods that render the internal models that are used to characterize speech sounds more robust to environmental variability.

This book will be of great interest to several types of readers. First (and most obviously), readers who are unfamiliar with the field of distant speech recognition can learn in this volume all of the technical background needed to construct and integrate a complete distant speech recognition system. In addition, the discussions in this volume are presented in self-contained chapters that enable technically literate readers in all fields to acquire a deep level of knowledge about relevant disciplines that are complementary to their own primary fields of expertise. Computer scientists can profit from the discussions on signal processing that begin with elementary signal representation and transformation and lead to advanced topics such as optimal Bayesian filtering, multirate digital signal processing, blind source separation, and speaker tracking. Classically-trained engineers will benefit from the detailed discussion of the theory and implementation of computer speech recognition systems including the extraction and enhancement of features representing speech sounds, statistical modeling of speech and language, along with the optimal search for the best available match between the incoming utterance and the internally-stored statistical representations of speech. Both of these groups will benefit from the treatments of physical acoustics, speech production, and auditory perception that are too frequently omitted from books of this type. Finally, the detailed contemporary exposition will serve to bring experienced practitioners who have been in the field for some time up to date on the most current approaches to robust recognition for language spoken from a distance.

Doctors Wölfel and McDonough have provided a resource to scientists and engineers that will serve as a valuable tutorial exposition and practical reference for all aspects associated with robust speech recognition in practical environments as well as for speech recognition in general. I am very pleased that this information is now available so easily and conveniently in one location. I fully expect that the publication of *Distant Speech Recognition* will serve as a significant accelerant to future work in the field, bringing us closer to the day in which transparent speech-based human-machine interfaces will become a practical reality in our daily lives everywhere.

Richard M. Stern
Pittsburgh, PA, USA

Preface

Our primary purpose in writing this book has been to cover a broad body of techniques and diverse disciplines required to enable reliable and natural verbal interaction between humans and computers. In the early nineties, many claimed that automatic speech recognition (ASR) was a “solved problem” as the word error rate (WER) had dropped below the 5% level for professionally trained speakers such as in the Wall Street Journal (WSJ) corpus. This perception changed, however, when the Switchboard Corpus, the first corpus of spontaneous speech recorded over a telephone channel, became available. In 1993, the first reported error rates on Switchboard, obtained largely with ASR systems trained on WSJ data, were over 60%, which represented a twelve-fold degradation in accuracy. Today the ASR field stands at the threshold of another radical change. WERs on telephony speech corpora such as the Switchboard Corpus have dropped below 10%, prompting many to once more claim that ASR is a solved problem. But such a claim is credible only if one ignores the fact that such WERs are obtained with *close-talking microphones*, such as those in telephones, and when only a single person is speaking. One of the primary hindrances to the widespread acceptance of ASR as the man-machine interface of first choice is the necessity of wearing a head-mounted microphone. This necessity is dictated by the fact that, under the current state of the art, WERs with microphones located a meter or more away from the speaker’s mouth can catastrophically increase, making most applications impractical. The interest in developing techniques for overcoming such practical limitations is growing rapidly within the research community. This change, like so many others in the past, is being driven by the availability of new corpora, namely, speech corpora recorded with far-field sensors. Examples of such include the meeting corpora which have been recorded at various sites including the International Computer Science Institute in Berkeley, California, Carnegie Mellon University in Pittsburgh, Pennsylvania and the National Institute of Standards and Technologies (NIST) near Washington, D.C., USA. In 2005, conversational speech corpora that had been collected with *microphone arrays* became available for the first time, after being released by the European Union projects *Computers in the Human Interaction Loop* (CHIL) and *Augmented Multiparty Interaction* (AMI). Data collected by both projects was subsequently shared with NIST for use in the semi-annual Rich Transcription evaluations it sponsors. In 2006 Mike Lincoln at Edinburgh University in Scotland collected the first corpus of *overlapping speech* captured with microphone arrays. This data collection effort involved real speakers who read sentences from the 5,000 word WSJ task.

In the view of the current authors, ground breaking progress in the field of distant speech recognition can only be achieved if the mainstream ASR community adopts methodologies and techniques that have heretofore been confined to the fringes. Such technologies include speaker tracking for determining a speaker's position in a room, beamforming for combining the signals from an array of microphones so as to concentrate on a desired speaker's speech and suppress noise and reverberation, and source separation for effective recognition of overlapping speech. Terms like filter bank, generalized sidelobe canceller, and diffuse noise field must become household words within the ASR community. At the same time researchers in the fields of acoustic array processing and source separation must become more knowledgeable about the current state of the art in the ASR field. This community must learn to speak the language of word lattices, semi-tied covariance matrices, and weighted finite-state transducers. For too long, the two research communities have been content to effectively ignore one another. With a few notable exceptions, the ASR community has behaved as if a speech signal does not exist before it has been converted to cepstral coefficients. The array processing community, on the other hand, continues to publish experimental results obtained on artificial data, with ASR systems that are nowhere near the state of the art, and on tasks that have long since ceased to be of any research interest in the mainstream ASR world. It is only if each community adopts the best practices of the other that they can together meet the challenge posed by distant speech recognition. We hope with our book to make a step in this direction.

Acknowledgments

We wish to thank the many colleagues who have reviewed parts of this book and provided very useful feedback for improving its quality and correctness. In particular we would like to thank the following people: Elisa Barney Smith, Friedrich Faubel, Sadaoki Furui, Reinhold Häb-Umbach, Kenichi Kumatani, Armin Sehr, Antske Fokkens, Richard Stern, Piergiorgio Svaizer, Helmut Wölfel, Najib Hadir, Hassan El-soumsoumani, and Barbara Rauch. Furthermore we would like to thank Tiina Ruonamaa, Sarah Hinton, Anna Smart, Sarah Tilley, and Brett Wells at Wiley who have supported us in writing this book and provided useful insights into the process of producing a book, not to mention having demonstrated the patience of saints through many delays and deadline extensions. We would also like to thank the university library at Universität Karlsruhe (TH) for providing us with a great deal of scholarly material, either online or in books.

We would also like to thank the people who have supported us during our careers in speech recognition. First of all thanks is due to our Ph.D. supervisors Alex Waibel, Bill Byrne, and Frederick Jelinek who have fostered our interest in the field of automatic speech recognition. Satoshi Nakamura, Mari Ostendorf, Dietrich Klakow, Mike Savic, Gerasimos (Makis) Potamianos, and Richard Stern always proved more than willing to listen to our ideas and scientific interests, for which we are grateful. We would furthermore like to thank IEEE and ISCA for providing platforms for exchange, publications and for hosting various conferences. We are indebted to Jim Flanagan and Harry Van Trees, who were among the great pioneers in the array processing field. We are also much obliged to the tireless employees at NIST, including Vince Stanford, Jon Fiscus and John Garofolo, for providing us with our first real microphone array, the Mark III, and hosting the annual evaluation campaigns which have provided a tremendous impetus for advancing

the entire field. Thanks is due also to Cedrick Rochét for having built the Mark III while at NIST, and having improved it while at Universität Karlsruhe (TH). In the latter effort, Maurizio Omologo and his coworkers at ITC-irst in Trento, Italy were particularly helpful. We would also like to thank Kristian Kroschel at Universität Karlsruhe (TH) for having fostered our initial interest in microphone arrays and agreeing to collaborate in teaching a course on the subject. Thanks is due also to Mike Riley and Mehryar Mohri for inspiring our interest in weighted finite-state transducers. Emilian Stoimenov was an important contributor to many of the finite-state transducer techniques described here. And of course, the list of those to whom we are indebted would not be complete if we failed to mention the undergraduates and graduate students at Universität Karlsruhe (TH) who helped us to build an instrumented seminar room for the CHIL project, and thereafter collect the audio and video data used for many of the experiments described in the final chapter of this work. These include Tobias Gehrig, Uwe Mayer, Fabian Jakobs, Keni Bernardin, Kai Nickel, Hazim Kemal Ekenel, Florian Kraft, and Sebastian Stüker. We are also naturally grateful to the funding agencies who made the research described in this book possible: the European Commission, the American Defense Advanced Research Projects Agency, and the Deutsche Forschungsgemeinschaft.

Most important of all, our thanks goes to our families. In particular, we would like to thank Matthias' wife Irina Wölfel, without whose support during the many evenings, holidays and weekends devoted to writing this book, we would have had to survive only on cold pizza and Diet Coke. Thanks is also due to Helmut and Doris Wölfel, John McDonough, Sr. and Christopher McDonough, without whose support through life's many trials, this book would not have been possible. Finally, we fondly remember Kathleen McDonough.

Matthias Wölfel
Karlsruhe, Germany

John McDonough
Saarbrücken, Germany



1

Introduction

For humans, speech is the quickest and most natural form of communication. Beginning in the late 19th century, verbal communication has been systematically extended through technologies such as radio broadcast, telephony, TV, CD and MP3 players, mobile phones and the Internet by voice over IP. In addition to these examples of one and two way verbal human–human interaction, in the last decades, a great deal of research has been devoted to extending our capacity of verbal communication with computers through *automatic speech recognition* (ASR) and speech synthesis. The goal of this research effort has been and remains to enable simple and natural *human–computer interaction* (HCI). Achieving this goal is of paramount importance, as verbal communication is not only fast and convenient, but also the only feasible means of HCI in a broad variety of circumstances. For example, while driving, it is much safer to simply ask a car navigation system for directions, and to receive them verbally, than to use a keyboard for tactile input and a screen for visual feedback. Moreover, hands-free computing is also accessible for disabled users.

1.1 Research and Applications in Academia and Industry

Hands-free computing, much like hands-free speech processing, refers to computer interface configurations which allow an interaction between the human user and computer without the use of the hands. Specifically, this implies that no close-talking microphone is required. Hands-free computing is important because it is useful in a broad variety of applications where the use of other common interface devices, such as a mouse or keyboard, are impractical or impossible. Examples of some currently available hands-free computing devices are camera-based head location and orientation-tracking systems, as well as gesture-tracking systems. Of the various hands-free input modalities, however, *distant speech recognition* (DSR) systems provide by far the most flexibility. When used in combination with other hands-free modalities, they provide for a broad variety of HCI possibilities. For example, in combination with a pointing gesture system it would become possible to turn on a particular light in the room by pointing at it while saying, “Turn on this light.”

The remainder of this section describes a variety of applications where speech recognition technology is currently under development or already available commercially. The

application areas include intelligent home and office environments, humanoid robots, automobiles, and speech-to-speech translation.

1.1.1 *Intelligent Home and Office Environments*

A great deal of research effort is directed towards equipping household and office devices – such as appliances, entertainment centers, personal digital assistants and computers, phones or lights – with more user friendly interfaces. These devices should be unobtrusive and should not require any special attention from the user. Ideally such devices should know the mental state of the user and act accordingly, gradually relieving household inhabitants and office workers from the chore of manual control of the environment. This is possible only through the application of sophisticated algorithms such as speech and speaker recognition applied to data captured with far-field sensors.

In addition to applications centered on HCI, computers are gradually gaining the capacity of acting as mediators for human–human interaction. The goal of the research in this area is to build a computer that will serve human users in their interactions with other human users; instead of requiring that users concentrate on their interactions with the machine itself, the machine will provide ancillary services enabling users to attend exclusively to their interactions with other people. Based on a detailed understanding of human perceptual context, intelligent rooms will be able to provide active assistance without any explicit request from the users, thereby requiring a minimum of attention from and creating no interruptions for their human users. In addition to speech recognition, such services need qualitative human analysis and human factors, natural scene analysis, multimodal structure and content analysis, and HCI. All of these capabilities must also be integrated into a single system.

Such interaction scenarios have been addressed by the recent projects *Computers in the Human Interaction Loop* (CHIL), *Augmented Multi-party Interaction* (AMI), as well as the successor of the latter *Augmented Multi-party Interaction with Distance Access* (AMIDA), all of which were sponsored by the European Union. To provide such services requires technology that models human users, their activities, and intentions. Automatically recognizing and understanding human speech plays a fundamental role in developing such technology. Therefore, all of the projects mentioned above have sought to develop technology for automatic transcription using speech data captured with distant microphones, determining who spoke when and where, and providing other useful services such as the summarizations of verbal dialogues. Similarly, the *Cognitive Assistant that Learns and Organizes* (CALO) project sponsored by the US *Defense Advanced Research Project Agency* (DARPA), takes as its goal the extraction of information from audio data captured during group interactions.

A typical meeting scenario as addressed by the AMIDA project is shown in Figure 1.1. Note the three microphone arrays placed at various locations on the table, which are intended to capture far-field speech for speaker tracking, beamforming, and DSR experiments. Although not shown in the photograph, the meeting participants typically also wear close-talking microphones to provide the best possible sound capture as a reference against which to judge the performance of the DSR system.

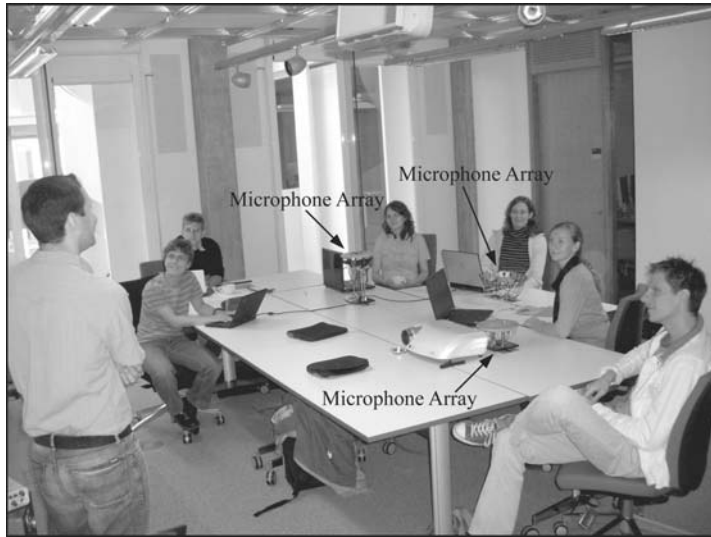


Figure 1.1 A typical AMIDA interaction. (© Photo reproduced by permission of the University of Edinburgh)

1.1.2 Humanoid Robots

If humanoid robots are ever to be accepted as full ‘partners’ by their human users, they must eventually develop perceptual capabilities similar to those possessed by humans, as well as the capacity of performing a diverse collection of tasks, including learning, reasoning, communicating and forming goals through interaction with both users and instructors. To provide for such capabilities, ASR is essential, because, as mentioned previously, spoken communication is the most common and flexible form of communication between people. To provide a natural interaction between a human and a humanoid robot requires not only the development of speech recognition systems capable of functioning reliably on data captured with far-field sensors, but also natural language capabilities including a sense of social interrelations and hierarchies.

In recent years, humanoid robots, albeit with very limited capabilities, have become commonplace. They are, for example, deployed as entertainment or information systems. Figure 1.2 shows an example of such a robot, namely, the humanoid tour guide robot *TPR-Robina*¹ developed by Toyota. The robot is able to escort visitors around the Toyota Kaikan Exhibition Hall and to interact with them through a combination of verbal communication and gestures.

While humanoid robots programmed for a limited range of tasks are already in widespread use, such systems lack the capability of learning and adapting to new environments. The development of such a capacity is essential for humanoid robots to become helpful in everyday life. The *Cognitive Systems for Cognitive Assistants (COSY)* project, financed by the European Union, has the objective to develop two kinds of robots providing such advanced capabilities. The first robot will find its way around a

¹ ROBINA stands for ROBOT as INtelligent Assistant.



Figure 1.2 Humanoid tour guide robot TPR-Robina by Toyota which escort visitors around Toyota Kaikan Exhibition Hall in Toyota City, Aichi Prefecture, Japan. (© Photo reproduced by permission of Toyota Motor Corporation)

complex building, showing others where to go and answering questions about routes and locations. The second will be able to manipulate structured objects on a table top. A photograph of the second COSY robot during an interaction session is shown in Figure 1.3.

1.1.3 Automobiles

There is a growing trend in the automotive industry towards increasing both the number and the complexity of the features available in high end models. Such features include entertainment, navigation, and telematics systems, all of which compete for the driver's visual and auditory attention, and can increase his cognitive load. ASR in such automobile environments would promote the "Eyes on the road, hands on the wheel" philosophy. This would not only provide more convenience for the driver, but would in addition actually

- out of vocabulary, 263
- overfitting, 171
- overlap-add, 85
- overlap-save, 85, 87
- oversampling, 360
- overtraining, 310
- parametric
 - JPDAF, 120
- Parseval's theorem, 79
- partial fraction expansion, 76
- partials, 35
- particle, 126
 - effective, 129
 - filter, 121, 183
 - weight, 95
- passband, 78, 373
- path, 243
 - successful, 243
- pdf
 - K_0 , 472
 - Γ , 472
 - Gaussian, 472
 - generalized Gaussian, 472
 - importance, 124
 - Laplace, 472
 - proposal, 124
- pentaphone model, 264
- perception
 - human, 12
 - loudness, 43
- perfect reconstruction, 359
- perplexity, 9, 259
- perseverative, 38
- phase
 - insensitivity, 42
 - shift, 70
- phon, 43
- phone, 36
- phoneme, 34, 36
- phonemic restoration, 13
- pitch, 34
 - human perception, 42
 - virtual, 46
- plosive, 37
- Poisson model, 114
- poles, 75
 - simple, 75
- Polyak average, 205
- polyphase representation, 364
 - Type 1, 364
 - Type 2, 364
- positive-definite, 105
- power
 - pattern, 426
 - spectral density, 477, 517
 - spectrum, 136, 145
- power-normalized LMS algorithm, 455
- pre-emphasis filter, 138
- prediction, 94, 96, 103, 106, 193
 - code-excited, 147
 - dynamic autoregressive process, 207
 - extended Kalman filter, 209
 - linear, 147
 - multi-step, 203
 - Polyak averaging and feedback, 205
 - step-size, 203
- pressure, 28
- principle of superposition, 67
- probability
 - backoff, 259
 - mass function, 113
- process, 95
 - stochastic, 233
- product rule, 101
- projection equation, 353
- proposal pdf, 124
- prototype, 360, 371
 - analysis-synthesis, 377
- proxemics, 57
- pulmonic, 37
- quantile, 197
- quasi-stationary, 136
- radiation, 55
- random walk, 204
- rarefaction, 28
- rate, 263
- Rayleigh quotient, 174
- recognition, 235
- reconstruction, 359

- recovery, 89
- reducing variance, 16
- reference, 498
- reflections
 - early, 201
 - late, 201
- reflectogram, 200
- region
 - convergence, 73
 - virtual, 420
 - visible, 420
- relative gradient, 404
- resampling, 128
 - semi-deterministic, 130
- residue, 46
- reverberation, 47, 48, 52
 - radius, 50
 - time, 50
- reversibility, 102
- Riccati equation, 104
- Rich Transcription, 18
- ROVER, 278
- sample, 123
 - effective size, 128
 - variance, 208
- sampling, 79
 - frequency, 80
 - importance, 124
 - interval, 80
 - Nyquist, 81
 - rate, 80, 82
 - subband, 90
 - theorem
 - Nyquist, 81
 - spatial, 61
- scale
 - Bark, 140
 - bilinear transform, 142
 - factor, 69
 - mel, 140
- scatter
 - matrix
 - between-class, 172
 - total, 173
 - within-class, 172
- score function, 403
- search, 231
 - error, 238
 - graph, 235
- second-order statistics, 388
- self-loop, 233
- semiring, 241
 - log-probability, 242
 - string, 242
 - tropical, 242
- semivowel, 38
- sensor
 - combination, 16
 - integration, 16
- septaphone model, 264
- sequence
 - periodic, 83
- sequential, 246
- series expansion, 73
- set partitioning algorithm, 251
- shape factor, 395
- shift, 361
- shifting property, 67
- sibilant, 38
- sidelobe, 72, 136, 361, 363, 420
- signal
 - cancellation problem, 447
 - noise ratio, 51
 - quality, 181
 - reconstruction, 79
 - reverberation ratio, 52
 - sampling, 79
- signal-to-noise ratio, 10
 - a posteriori, 197
 - a priori, 198
- simultaneous
 - masking, 44
- skewness, 40
- smoothing, 96
- soft-decision, 195
- sonogram, 138
- sound
 - capture, 7
 - distance, 45
 - intensity, 31
 - inverse square law, 31

- pressure, 28, 29
- pressure level, 42
- properties, 27
- speed, 29
- source
 - geometric separation, 466, 487
 - image, 33
 - point, 30
 - simple, 30
 - virtual, 33
- spatial
 - aliasing, 61, 424
 - correlation coefficient, 434
 - sampling theorem, 61
 - spectral matrix, 428
 - normalized, 429
- speaker
 - adaptation, 317
 - adapted training, 299
 - localization, 337
 - tracking, 337
- specific gas constant, 28
- spectra
 - analysis, 145
 - comparison, 159
 - envelope, 146
 - estimation, 145, 224
 - nonparametric, 145
 - parametric, 145
 - floor, 217
 - leakage, 136
 - linear prediction, 147
 - matrix, 438
 - minimum variance distortionless
 - response, 150
 - warped, 156
 - warped-twice, 157
 - power, 136, 145, 146
 - relationship, 152
 - relative, 191
 - resolution, 42, 143
 - scaling, 160
 - shaping, 138
 - subtraction, 215
 - non-linear, 217
 - tilt, 144, 187, 479
- spectral
 - subtraction, 182
- spectrogram, 137
- speech, 35
 - enhancement, 182
 - feature
 - fields in recognition, 10
 - parameters, 34
 - power, 43
 - statistical properties, 39
 - unvoiced, 35, 36, 136
 - voiced, 35, 136
- speed of sound, 29
- spherical
 - coordinates, 30
 - interpolation, 341
 - intersection, 339
 - waves, 33
- SPLICE, 192
- stability, 77
- stable, 68
- stacking, 170
- state
 - clustering, 265
 - equation, 28, 95
 - final, 233
 - hidden Markov model, 233
 - initial, 233
- static features, 169
- stationary, 28, 47
- statistical
 - inference, 221
 - language model, 256
- statistically
 - uncorrelated, 392
- statistics, 39
 - higher order, 40, 387, 390
 - second-order, 388
- steering, 422
 - error, 445
 - function, 159
- step size, 448
- stereo data, 192, 193
- stiffness, 31
- stop, 37
- stopband, 78, 361

- string-to-weight subsequential transducer, 247
- strong law of large numbers, 124
- sub-Gaussian, 521
- subband, 89, 90, 359
 - domain, 90
 - domain snapshot, 428
 - samples, 90
- successful path, 243
- super-Gaussian, 521
- superdirective, 437
- switching
 - dynamic system model, 206
 - model, 211, 213
- syllable, 37
- synthesis, 83, 371
- systems
 - discrete-time, 66
 - function, 73
 - inverse, 76
 - linear, 66
 - linear time-invariant, 65
 - minimum phase, 76
 - model, 95
 - time invariant, 67
- systematic resampling, 130
- Taylor series, 531
 - vector, 218
- telephone, 58
 - channel, 47
- temperature, 28
- temporal patterns, 171
- testing data, 8
- time
 - arrival, 45
 - delay, 66
 - delay of arrival, 338, 413
 - invariant, 67
- token, 236
- trace back, 237, 301
- track detection indicator, 117
- tracking, 97
- training, 235
 - data, 8
 - discriminative, 302
 - label, 301, 302
 - Viterbi, 301, 302
- transducer
 - finite-state
 - weight, 243
 - sequential, 274
 - string-to-weight, 247
- transfer function, 73, 198
- transform
 - bilinear, 142, 324
 - discrete cosine, 517
 - Fourier
 - discrete, 82, 85
 - discrete-time, 69
 - fast, 82
 - inverse, 69
 - short-time, 72, 87, 89
 - time-dependent, 87
 - Householder, 537
 - operator, 66
 - pair, 70
 - phase, 339
 - sine-log all-pass, 329
- transition, 242
 - band, 78
- TRAPS, 171
- trigram, 258
- triphone model, 264
- Type 2 DCT matrix, 517
- uncertainty
 - observation, 227
 - processing, 13
- unidirectional, 59
- unit
 - circle, 73, 77
 - impulse, 67
- update, 96, 104, 193
- validation
 - matrix, 116
 - observation, 110
 - region, 110
- virtual
 - pitch, 46