Tuomas Virtanen | Rita Singh | Bhiksha Raj Techniques for Noise Robustness in Automatic Speech Recognition



TECHNIQUES FOR NOISE ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION

TECHNIQUES FOR NOISE ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION

Editors

Tuomas Virtanen *Tampere University of Technology, Finland*

Rita Singh Carnegie Mellon University, USA

Bhiksha Raj Carnegie Mellon University, USA



This edition first published 2013 © 2013 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.

Library of Congress Cataloging-in-Publication Data

Virtanen, Tuomas.

Techniques for noise robustness in automatic speech recognition / Tuomas Virtanen, Rita Singh, Bhiksha Raj. p. cm. Includes bibliographical references and index.

ISBN 978-1-119-97088-0 (cloth)

1. Automatic speech recognition. I. Singh, Rita. II. Raj, Bhiksha. III. Title. TK7882.865V57 2012

006.4'54-dc23

2012015742

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

ISBN: 978-0-470-97409-4

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

Contents

List of Contributors			
Ackn	owledgm	ents	xvii
1	Introdu Tuomas	i ction Virtanen, Rita Singh, Bhiksha Raj	1
1.1 1.2 1.3	Scope o Outline Notation	f the Book	1 2 4
Part	One F	OUNDATIONS	
2	The Ba <i>Rita Sin</i>	sics of Automatic Speech Recognition gh, Bhiksha Raj, Tuomas Virtanen	9
2.1	Introduc	ction	9
2.2	Speech	Recognition Viewed as Bayes Classification	10
2.3	Hidden	Markov Models	11
	2.3.1	Computing Probabilities with HMMs	12
	2.3.2	Determining the State Sequence	17
	2.3.3	Learning HMM Parameters	19
	2.3.4	Additional Issues Relating to Speech Recognition Systems	20
2.4	HMM-H	Based Speech Recognition	24
	2.4.1	Representing the Signal	24
	2.4.2	The HMM for a Word Sequence	25
	2.4.3	Searching through all Word Sequences	26
	Referen	ces	29
3	The Pro <i>Bhiksha</i>	bblem of Robustness in Automatic Speech Recognition Raj, Tuomas Virtanen, Rita Singh	31
3.1	Errors in	n Bayes Classification	31
	3.1.1	Type 1 Condition: Mismatch Error	33
	3.1.2	Type 2 Condition: Increased Bayes Error	34
3.2	Bayes C	Classification and ASR	35
	3.2.1	All We Have is a Model: A Type 1 Condition	35

	3.2.2	Intrinsic Interferences—Signal Components that are Unrelated to	
		the Message: A Type 2 Condition	36
	3.2.3	External Interferences—The Data are Noisy: Type 1 and	
		Type 2 Conditions	36
3.3	Externa	al Influences on Speech Recordings	36
	3.3.1	Signal Capture	37
	3.3.2	Additive Corruptions	41
	3.3.3	Reverberation	42
	3.3.4	A Simplified Model of Signal Capture	43
3.4	The Eff	fect of External Influences on Recognition	44
3.5	Improving Recognition under Adverse Conditions		46
	3.5.1	Handling the Model Mismatch Error	46
	3.5.2	Dealing with Intrinsic Variations in the Data	47
	3.5.3	Dealing with Extrinsic Variations	47
	Referen	nces	50

Part Two SIGNAL ENHANCEMENT

4	Voice A	Voice Activity Detection, Noise Estimation, and Adaptive Filters for			
	Acoust	ic Signal Enhancement	53		
	Rainer	Martin, Dorothea Kolossa			
4.1	Introdu	ction	53		
4.2	Signal	Analysis and Synthesis	55		
	4.2.1	DFT-Based Analysis Synthesis with Perfect Reconstruction	55		
	4.2.2	Probability Distributions for Speech and Noise DFT Coefficients	57		
4.3	Voice A	Activity Detection	58		
	4.3.1	VAD Design Principles	58		
	4.3.2	Evaluation of VAD Performance	62		
	4.3.3	Evaluation in the Context of ASR	62		
4.4	Noise Power Spectrum Estimation		65		
	4.4.1	Smoothing Techniques	65		
	4.4.2	Histogram and GMM Noise Estimation Methods	67		
	4.4.3	Minimum Statistics Noise Power Estimation	67		
	4.4.4	MMSE Noise Power Estimation	68		
	4.4.5	Estimation of the A Priori Signal-to-Noise Ratio	69		
4.5	Adaptiv	ve Filters for Signal Enhancement	71		
	4.5.1	Spectral Subtraction	71		
	4.5.2	Nonlinear Spectral Subtraction	73		
	4.5.3	Wiener Filtering	74		
	4.5.4	The ETSI Advanced Front End	75		
	4.5.5	Nonlinear MMSE Estimators	75		
4.6	ASR Pe	erformance	80		
4.7	Conclusions		81		
	References		82		

5	Extraction of Speech from Mixture Signals Paris Smaragdis	87	
5.1	The Problem with Mixtures	87	
5.2	Multichannel Mixtures	88	
0.2	5.2.1 Basic Problem Formulation	88	
	5.2.2 Convolutive Mixtures	92	
53	Single-Channel Mixtures	98	
0.0	5.3.1 Problem Formulation	98	
	5.3.2 Learning Sound Models	100	
	5.3.3 Separation by Spectrogram Factorization	101	
	5.3.4 Dealing with Unknown Sounds	105	
54	Variations and Extensions	107	
5 5	Conclusions	107	
0.0	References	107	
6	Microphone Arrays	109	
	John McDonough, Kenichi Kumatani		
6.1	Speaker Tracking	110	
6.2	Conventional Microphone Arrays	113	
6.3	Conventional Adaptive Beamforming Algorithms	120	
	6.3.1 Minimum Variance Distortionless Response Beamformer	120	
	6.3.2 Noise Field Models	122	
	6.3.3 Subband Analysis and Synthesis	123	
	6.3.4 Beamforming Performance Criteria	126	
	6.3.5 Generalized Sidelobe Canceller Implementation	129	
	6.3.6 Recursive Implementation of the GSC	130	
	6.3.7 Other Conventional GSC Beamformers	131	
	6.3.8 Beamforming based on Higher Order Statistics	132	
	6.3.9 Online Implementation	136	
	6.3.10 Speech-Recognition Experiments	140	
6.4	Spherical Microphone Arrays	142	
6.5	Spherical Adaptive Algorithms	148	
6.6	Comparative Studies	149	
6.7	Comparison of Linear and Spherical Arrays for DSR	152	
6.8	Conclusions and Further Reading	154	
	References	155	
Part	Three FEATURE ENHANCEMENT		
7	From Signals to Speech Features by Digital Signal Processing Matthias Wölfel	161	
7.1	Introduction	161	
	7.1.1 About this Chapter	162	
7.2	The Speech Signal 16		

7.3	Spectra	l Processing	163
	7.3.1	Windowing	163
	7.3.2	Power Spectrum	165
	7.3.3	Spectral Envelopes	166
	7.3.4	LP Envelope	166
	7.3.5	MVDR Envelope	169
	7.3.6	Warping the Frequency Axis	171
	7.3.7	Warped LP Envelope	175
	7.3.8	Warped MVDR Envelope	176
	7.3.9	Comparison of Spectral Estimates	177
	7.3.10	The Spectrogram	179
7.4	Cepstra	l Processing	179
	7.4.1	Definition and Calculation of Cepstral Coefficients	180
	7.4.2	Characteristics of Cepstral Sequences	181
7.5	Influen	ce of Distortions on Different Speech Features	182
	7.5.1	Objective Functions	182
	7.5.2	Robustness against Noise	185
	7.5.3	Robustness against Echo and Reverberation	187
	7.5.4	Robustness against Changes in Fundamental Frequency	189
7.6	Summa	ry and Further Reading	191
	Referen	nces	191
8	Featur	es Based on Auditory Physiology and Perception	193
Ū	Richard	l M. Stern, Nelson Morgan	170
8.1	Introdu	ction	193
8.2	Some A	Attributes of Auditory Physiology and Perception	194
	8.2.1	Peripheral Processing	194
	8.2.2	Processing at more Central Levels	200
	8.2.3	Psychoacoustical Correlates of Physiological Observations	202
	8.2.4	The Impact of Auditory Processing on Conventional	
		Feature Extraction	206
	8.2.5	Summary	208
8.3	"Classi	c" Auditory Representations	208
8.4	Current	t Trends in Auditory Feature Analysis	213
8.5	Summa	ſŸ	221
	Acknow	vledgments	222
	Referen	nces	222
9	Featur	e Compensation	229
	Jasha L	Droppo	
9.1	Life in	an Ideal World	229
	9.1.1	Noise Robustness Tasks	229
	9.1.2	Probabilistic Feature Enhancement	230
	9.1.3	Gaussian Mixture Models	231

9.2	MMSE-SPLICE		232
	9.2.1 Paran	neter Estimation	233
	9.2.2 Resul	ts	236
9.3	Discriminative S	SPLICE	237
	9.3.1 The M	IMI Objective Function	238
	9.3.2 Train	ing the Front-End Parameters	239
	9.3.3 The R	Pprop Algorithm	240
	9.3.4 Resul	ts	241
9.4	Model-Based Fe	eature Enhancement	242
	9.4.1 The A	dditive Noise-Mixing Equation	243
	9.4.2 The J	oint Probability Model	244
	9.4.3 Vector	r Taylor Series Approximation	246
	9.4.4 Estim	ating Clean Speech	247
	9.4.5 Resul	ts	247
9.5	Switching Linea	r Dynamic System	248
9.6	Conclusion		249
	References		249
10	Reverberant Sp	peech Recognition	251
	Reinhold Haeb-	Umbach, Alexander Krueger	
10.1	Introduction		251
10.2	The Effect of Re	everberation	252
	10.2.1 What	is Reverberation?	252
	10.2.2 The R	elationship between Clean and Reverberant	
	Speec	h Features	254
	10.2.3 The E	ffect of Reverberation on ASR Performance	258
10.3	Approaches to R	Reverberant Speech Recognition	258
	10.3.1 Signa	l-Based Techniques	259
	10.3.2 Front	-End Techniques	260
	10.3.3 Back-	End Techniques	262
	10.3.4 Conci	luding Remarks	265
10.4	Feature Domain	Model of the Acoustic Impulse Response	265
10.5	Bayesian Featur	e Enhancement	267
	10.5.1 Basic	Approach	268
	10.5.2 Meas	urement Update	269
	10.5.3 Time	Update	270
	10.5.4 Infere	ence	271
10.6	Experimental Re	esults	272
	10.6.1 Datab	bases	272
	10.6.2 Overv	view of the Tested Methods	273
	10.6.3 Recog	gnition Results on Reverberant Speech	274
	10.6.4 Recog	gnition Results on Noisy Reverberant Speech	276
10.7	Conclusions		277
	Acknowledgmen	nt	278
	•		
10.7	Conclusions Acknowledgmei	nt	

Part Four MODEL ENHANCEMENT

11	Adapta Yannick	tion and Discriminative Training of Acoustic Models Estève, Paul Deléglise	285
11.1	Introduc	ction	285
	11.1.1	Acoustic Models	286
	11.1.2	Maximum Likelihood Estimation	287
11.2	Acousti	c Model Adaptation and Noise Robustness	288
	11.2.1	Static (or Offline) Adaptation	289
	11.2.2	Dynamic (or Online) Adaptation	289
11.3	Maximu	um A Posteriori Reestimation	290
11.4	Maximu	Im Likelihood Linear Regression	293
	11.4.1	Class Regression Tree	294
	11.4.2	Constrained Maximum Likelihood Linear Regression	297
	11.4.3	CMLLR Implementation	297
	11.4.4	Speaker Adaptive Training	298
11.5	Discrim	inative Training	299
	11.5.1	MMI Discriminative Training Criterion	301
	11.5.2	MPE Discriminative Training Criterion	302
	11.5.3	I-smoothing	303
	11.5.4	MPE Implementation	304
11.6	Conclus	ion	307
	Referen	ces	308
12	Factoria	al Models for Noise Robust Speech Recognition	311
	John K.	Hersney, Steven J. Kennie, Jonathan Le Roux	
12.1	Introduc	ction	311
12.2	The Mo	del-Based Approach	313
12.3	Signal F	Feature Domains	314
12.4	Interacti	ion Models	317
	12.4.1	Exact Interaction Model	318
	12.4.2	Max Model	320
	12.4.3	Log-Sum Model	321
	12.4.4	Mel Interaction Model	321
12.5	Inferenc	re Methods	322
	12.5.1	Max Model Inference	322
	12.5.2	Parallel Model Combination	324
	12.5.3	Vector Taylor Series Approaches	326
	12.5.4	SNR-Dependent Approaches	331
12.6	Efficien	t Likelihood Evaluation in Factorial Models	332
	12.6.1	Efficient Inference using the Max Model	332
	12.6.2	Efficient Vector-Taylor Series Approaches	334
	12.6.3	Band Quantization	335
12.7	Current	Directions	337
	12.7.1	Dynamic Noise Models for Robust ASR	338

	12.7.2 12.7.3	Multi-Talker Speech Recognition using Graphical Models Noise Robust ASR using Non-Negative	339
		Basis Representations	340
	Referen	ces	341
13	Acousti Michael	c Model Training for Robust Speech Recognition	347
13.1	Introduc	ction	347
13.2	Tradition	nal Training Methods for Robust Speech Recognition	348
13.3	A Brief	Overview of Speaker Adaptive Training	349
13.4	Feature-	Space Noise Adaptive Training	351
	13.4.1	Experiments using fNAT	352
13.5	Model-S	Space Noise Adaptive Training	353
13.6	Noise A	daptive Training using VTS Adaptation	355
	13.6.1	Vector Taylor Series HMM Adaptation	355
	13.6.2	Updating the Acoustic Model Parameters	357
	13.6.3	Updating the Environmental Parameters	360
	13.6.4	Implementation Details	360
	13.6.5	Experiments using NAT	361
13.7	Discussi	ion	364
	13.7.1	Comparison of Training Algorithms	364
	13.7.2	Comparison to Speaker Adaptive Training	364
	13.7.3	Related Adaptive Training Methods	365
13.8	Conclus	ion	366
	Referen	ces	366

Part Five COMPENSATION FOR INFORMATION LOSS

14	Missing Jon Bar	-Data Techniques: Recognition with Incomplete Spectrograms <i>ker</i>	371
14.1	Introduc	tion	371
14.2	Classification with Incomplete Data		373
	14.2.1	A Simple Missing Data Scenario	374
	14.2.2	Missing Data Theory	376
	14.2.3	Validity of the MAR Assumption	378
	14.2.4	Marginalising Acoustic Models	379
14.3	Energetic Masking		381
	14.3.1	The Max Approximation	381
	14.3.2	Bounded Marginalisation	382
	14.3.3	Missing Data ASR in the Cepstral Domain	384
	14.3.4	Missing Data ASR with Dynamic Features	386
14.4	Meta-M	issing Data: Dealing with Mask Uncertainty	388
	14.4.1	Missing Data with Soft Masks	388

	14.4.2	Sub-band Combination Approaches	391
	14.4.3	Speech Fragment Decoding	393
14.5	Some Pe	erspectives on Performance	395
	Referen	ces	396
15	Missing	g-Data Techniques: Feature Reconstruction	399
	Jort Flo	rent Gemmeke, Ulpu Remes	
15.1	Introduc	ction	399
15.2	Missing	-Data Techniques	401
15.3	Correlat	tion-Based Imputation	402
	15.3.1	Fundamentals	402
	15.3.2	Implementation	404
15.4	Cluster-	Based Imputation	406
	15.4.1	Fundamentals	406
	15.4.2	Implementation	408
	15.4.3	Advances	409
15.5	Class-C	onditioned Imputation	411
	15.5.1	Fundamentals	411
	15.5.2	Implementation	412
	15.5.3	Advances	413
15.6	Sparse I	mputation	414
	15.6.1	Fundamentals	414
	15.6.2	Implementation	416
	15.6.3	Advances	418
15.7	Other Fe	eature-Reconstruction Methods	420
	15.7.1	Parametric Approaches	420
	15.7.2	Nonparametric Approaches	421
15.8	Experin	nental Results	421
	15.8.1	Feature-Reconstruction Methods	422
	15.8.2	Comparison with Other Methods	424
	15.8.3	Advances	426
	15.8.4	Combination with Other Methods	427
15.9	Discussi	ion and Conclusion	428
	Acknow	ledgments	429
	Referen	ces	430
16	Compu	tational Auditory Scene Analysis and Automatic	
	Speech	Recognition	433
	Arun Na	arayanan, DeLiang Wang	
16.1	Introduc	ction	433
16.2	Auditor	y Scene Analysis	434
16.3	Comput	ational Auditory Scene Analysis	435
	16.3.1	Ideal Binary Mask	435
	16.3.2	Typical CASA Architecture	438

16.4	CASA Strategies	440
	16.4.1 IBM Estimation Based on Local SNR Estimates	440
	16.4.2 IBM Estimation using ASA Cues	442
	16.4.3 IBM Estimation as Binary Classification	448
	16.4.4 Binaural Mask Estimation Strategies	451
16.5	Integrating CASA with ASR	452
	16.5.1 Uncertainty Transform Model	454
16.6	Concluding Remarks	458
	Acknowledgment	458
	References	458
17	Uncertainty Decoding	463
	Hank Liao	
17.1	Introduction	463
17.2	Observation Uncertainty	465
17.3	Uncertainty Decoding	466
17.4	Feature-Based Uncertainty Decoding	468
	17.4.1 SPLICE with Uncertainty	470
	17.4.2 Front-End Joint Uncertainty Decoding	471
	17.4.3 Issues with Feature-Based Uncertainty Decoding	472
17.5	Model-Based Joint Uncertainty Decoding	473
	17.5.1 Parameter Estimation	475
	17.5.2 Comparisons with Other Methods	476
17.6	Noisy CMLLR	477
17.7	Uncertainty and Adaptive Training	480
	17.7.1 Gradient-Based Methods	481
	17.7.2 Factor Analysis Approaches	482
17.8	In Combination with Other Techniques	483
17.9	Conclusions	484
	References	485

Index

487

List of Contributors

Jon Barker University of Sheffield, UK

Paul Deléglise University of Le Mans, France

Jasha Droppo Microsoft Research, USA

Yannick Estève University of Le Mans, France

Jort Florent Gemmeke KU Leuven, Belgium

Reinhold Haeb-Umbach University of Paderborn, Germany

John R. Hershey Mitsubishi Electric Research Laboratories, USA

Dorothea Kolossa Ruhr-Universität Bochum, Germany

Alexander Krueger University of Paderborn, Germany

Kenichi Kumatani Disney Research, USA

Jonathan Le Roux Mitsubishi Electric Research Laboratories, USA Hank Liao Google Inc., USA

Rainer Martin Ruhr-Universität Bochum, Germany

John McDonough Carnegie Mellon University, USA

Nelson Morgan International Computer Science Institute and the University of California, Berkeley, USA

Arun Narayanan The Ohio State University, USA

Bhiksha Raj Carnegie Mellon University, USA

Ulpu Remes Aalto University School of Science, Finland

Steven J. Rennie IBM Thomas J. Watson Research Center, USA

Michael L. Seltzer Microsoft Research, USA

Rita Singh Carnegie Mellon University, USA

Paris Smaragdis University of Illinois at Urbana-Champaign, USA

Richard Stern Carnegie Mellon University, USA

Tuomas Virtanen Tampere University of Technology, Finland

DeLiang Wang The Ohio State University, USA

Matthias Wölfel Pforzheim University, Germany

Acknowledgments

The editors would like to thank Jort Gemmeke, Joonas Nikunen, Pasi Pertilä, Janne Pylkkönen, Ulpu Remes, Rahim Saeidi, Michael Wohlmayr, Elina Helander, Kalle Palomäki, and Katariina Mahkonen, who have have assisted by providing constructive comments about individual chapters of the book.

Tuomas Virtanen would like to thank the Academy of Finland for financial support; and Professors Moncef Gabbouj, Sourish Chaudhuri, Mark Harvilla, and Ari Visa for supporting his position in the Department of Signal Processing, which has allowed for his editing this book.

1

Introduction

Tuomas Virtanen¹, Rita Singh², Bhiksha Raj² ¹Tampere University of Technology, Finland ²Carnegie Mellon University, USA

1.1 Scope of the Book

The term "computer speech recognition" conjures up visions of the science-fiction capabilities of HAL2000 in 2001, A Space Odessey, or "Data," the anthropoid robot in Star Trek, who can communicate through speech with as much ease as a human being. However, our real-life encounters with automatic speech recognition are usually rather less impressive, comprising often-annoying exchanges with interactive voice response, dictation, and transcription systems that make many mistakes, frequently misrecognizing what is spoken in a way that humans rarely would. The reasons for these mistakes are many. Some of the reasons have to do with fundamental limitations of the mathematical framework employed, and inadequate awareness or representation of context, world knowledge, and language. But other equally important sources of error are distortions introduced into the recorded audio during recording, transmission, and storage.

As automatic speech-recognition—or ASR—systems find increasing use in everyday life, the speech they must recognize is being recorded over a wider variety of conditions than ever before. It may be recorded over a variety of *channels*, including landline and cellular phones, the internet, etc. using different kinds of microphones, which may be placed close to the mouth such as in head-mounted microphones or telephone handsets, or at a distance from the speaker, such as desktop microphones. It may be corrupted by a wide variety of *noises*, such as sounds from various devices in the vicinity of the speaker, general background sounds such as those in a moving car or background babble in crowded places, or even competing speakers. It may also be affected by *reverberation*, caused by sound reflections in the recording environment. And, of course, all of the above may occur concurrently in myriad combinations and, just to make matters more interesting, may change unpredictably over time.

Techniques for Noise Robustness in Automatic Speech Recognition, First Edition.

Edited by Tuomas Virtanen, Rita Singh, and Bhiksha Raj.

© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.

For speech-recognition systems to perform acceptably, they must be *robust* to the distorting influences. This book deals with techniques that impart such robustness to ASR systems. We present a collection of articles from experts in the field, which describe an array of strategies that operate at various stages of processing in an ASR system. They range from techniques for minimizing the effect of external noises at the point of signal capture, to methods of deriving features from the signal that are fundamentally robust to signal degradation, techniques for attenuating the effect of external noises on the signal, and methods for modifying the recognition system itself to recognize degraded speech better.

The selection of techniques described in this book is intended to cover the range of approaches that are currently considered state of the art. Many of these approaches continue to evolve, nevertheless we believe that for a practitioner of the field to follow these developments, he must be familiar with the fundamental principles involved. The articles in this book are designed and edited to adequately present these fundamental principles. They are intended to be easy to understand, and sufficiently tutorial for the reader to be able to implement the described techniques.

1.2 Outline

Robustnesss techniques for ASR fall into a number of different categories. This book is divided into five parts, each focusing on a specific category of approaches. A clear understanding of robustness techniques for ASR requires a clear understanding of the principles behind automatic speech recognition and the robustness issues that affect them. These foundations are briefly discussed in Part One of the book. Chapter 2 gives a short introduction to the fundamentals of automatic speech recognition. Chapter 3 describes various distortions that affect speech signals, and analyzes their effect on ASR.

Part Two discusses techniques that are aimed at minimizing the distortions in the speech signal itself.

Chapter 4 presents methods for *voice-activity detection* (VAD), *noise estimation*, and *noise-suppression* techniques based on filtering. A VAD analyzes which signal segments correspond to speech and which to noise, so that an ASR system does not mistakenly interpret noise as speech. VAD can also provide an estimate of the noise during periods of speech inactivity. The chapter also reviews methods that are able to track noise characteristics even during speech activity. Noise estimates are required by many other techniques presented in the book.

Chapter 5 presents two approaches for separating speech from noises. The first one uses multiple microphones and an assumption that speech and noise signals are statistically independent of each other. The method does not use *a priori* information about the source signals, and is therefore termed *blind source separation*. Statistically independent signals are separated using an algorithm called *independent component analysis*. The second approach requires only a single microphone, but it is based on *a priori* information about speech or noise signals. The presented method is based on factoring the spectrogram of noisy speech into speech and noise using *nonnegative matrix factorization*.

Chapter 6 discusses methods that apply multiple microphones to selectively enhance speech while suppressing noise. They assume that the speech and noise sources are located in spatially different positions. By suitably combining the signals recorded by each microphone they are able to perform *beamforming*, which can selectively enhance signals from the location of the

speech source. The chapter first presents the fundamentals of conventional linear microphone arrays, then reviews different criteria that can be used to design them, and then presents methods that can be used in the case of *spherical microphone arrays*.

Part Three of the book discusses methods that attempt to minimize the effect of distortions on *acoustic features* that are used to represent the speech signal.

Chapter 7 reviews conventional feature extraction methods that typically parameterize the envelope of the spectrum. Both methods based on *linear prediction* and *cepstral* processing are covered. The chapter then discusses *minimum variance distortionless response* or *warping* techniques that can be applied to make the envelope estimates more reliable for purposes of speech recognition. The chapter also studies the effect of distortions on the features.

Chapter 8 approaches the noise robustness problem from the point of view of human speech perception. It first presents a series of auditory measurements that illustrate selected properties of the human auditory system, and then discusses principles that make the human auditory system less sensitive to external influences. Finally, it presents several computational *auditory models* that mimic human auditory processes to extract noise robust features from the speech signal.

Chapter 9 presents methods that reduce the effect of distortions on features derived from speech. These *feature-enhancement* techniques can be trained to map noisy features to clean ones using training examples of clean and noisy speech. The mapping can include a criterion which makes the enhanced features more *discriminative*, i.e., makes them more effective for speech recognition. The chapter also presents methods that use an explicit model for additive noises.

Chapter 10 focuses on the recognition of reverberant speech. It first analyzes the effect of reverberation on speech and the features derived from it. It gives a review of different approaches that can be used to perform recognition of reverberant speech and presents methods for enhancing features derived from reverberant speech based on a model of reverberation.

Part Four discusses methods which modify the statistical parameters employed by the recognizer to improve recognition of corrupted speech.

Chapter 11 presents adaptation methods which change the parameters of the recognizer without assuming a specific kind of distortion. These *model-adaptation* techniques are frequently used to adapt a recognizer to a specific speaker, but can equally effectively be used to adapt it to distorted signals. The chapter also presents training criteria that makes the statistical models in the recognizer more *discriminative*, to improve the recognition performance that can be obtained with them.

Chapter 12 focuses on compensating for the effect of interfering sound sources on the recognizer. Based on a model of interfering noises and a model of the interaction process between speech and noise, these *model-compensation* techniques can be used to derive a statistical model for noisy speech. In order to find a mapping between the models for clean and noisy speech, the techniques use various approximations of the interaction process.

Chapter 13 discusses a methodology that can be used to find the parameters of an ASR system to make it more robust, given any signal or feature enhancement method. These *noise-adaptive-training* techniques are applied in the training stage, where the parameters the ASR system are tuned to optimize the recognition accuracy.

Part Five presents techniques which address the issue that some information in the speech signal may be lost because of noise. We now have a problem of *missing data* that must be dealt with.

Chapter 14 first discusses the general taxonomy of different missing-data problems. It then discusses the conditions under which speech features can be considered reliable, and when they may be assumed to be missing. Finally, it presents methods that can be used to perform robust ASR when there is uncertainty about which parts of the signal are missing.

Chapter 15 presents methods that produce an estimate of missing features (i.e., *feature reconstruction*) using reliable features. Reconstruction methods based on a Gaussian mixture model utilize local correlations between missing and reliable features. The reconstruction can also be done separately for each state of the ASR system. *Sparse representation* methods model the noisy observation as a linear combination of a small number of atomic units taken from a larger dictionary, and the weights of the atomic units are determined using reliable features only.

Chapter 16 discusses methods that estimate which parts of a speech signal are missing and which ones are reliable. The estimation can be based either on the signal-to-noise ratio in each time-frequency component, or on more perceptually motivated cues derived from the signal, or using a binary classification approach.

Chapter 17 presents approaches which enable the modeling of the *uncertainty* caused by noise in the recognition system. It first discusses feature-based uncertainty, which enables modeling of the uncertainty in enhanced signals or features obtained through algorithms discussed in the previous chapters of the book. Model-based *uncertainty decoding*, on the other hand, enables us to account for uncertainties in model compensation or adaptation techniques. The chapter also discusses the use of uncertainties with noise-adaptive training techniques.

We also revisit the contents of the book in the end of Chapter 3, once we have analyzed the types of errors encountered in automatic speech recognition.

1.3 Notation

The table below lists the most commonly used symbols in the book. Some of the chapters deviate from the definitions below, but in such cases the used symbols are explicitly defined.

Symbol	Definition
$\overline{a, b, c, \ldots}$	Scalar variables
A, B, C, \ldots	Constants
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$	Vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$	Matrices
\otimes	Convolution
\mathcal{N}	Normal distribution
$\mathcal{E}\{x\}$	Expected value of <i>x</i>
\mathbf{A}^{T}	Transpose of matrix A
$x_{i:j}$	Set $x_i, x_{i+1}, \ldots, x_j$
S	Speech signal
n	Additive noise signal
x	Noisy speech signal
h	Response from speaker to microphone
t	Time index

Symbol	Definition
f	Frequency index
\mathbf{x}_t	Observation vector of noisy speech in frame t
q	State variable
q_t	State at time <i>t</i>
μ	Mean vector
$\boldsymbol{\Theta}, \boldsymbol{\Sigma}$	Covariance matrix
P, p	Probability

Part One Foundations

The Basics of Automatic Speech Recognition

Rita Singh¹, Bhiksha Raj¹, Tuomas Virtanen² ¹Carnegie Mellon University, USA ²Tampere University of Technology, Finland

2.1 Introduction

In order to understand the techniques described later in this book, it is important to understand how automatic speech-recognition (ASR) systems function. This chapter briefly outlines the framework employed by ASR systems based on hidden Markov models (HMMs).

Most mainstream ASR systems are designed as probabilistic Bayes classifiers that identify the most likely word sequence that explains a given recorded acoustic signal. To do so, they use an estimate of the probabilities of possible word sequences in the language, and the probability distributions of the acoustic signals for each word sequence. Both the probability distributions of word sequences, and those of the acoustic signals for any word sequence, are represented through parametric *models*. Probabilities of word sequences are modeled by various forms of grammars or *N*-gram models. The probabilities of the acoustic signals are modeled by HMMs.

In the rest of this chapter, we will briefly describe the components and process of ASR as outlined above, as a prelude to explaining the circumstances under which it may perform poorly, and how that relates to the remaining chapters of this book. Since this book primarily addresses factors that affect the *acoustic* signal, we will only pay cursory attention to the manner in which word-sequence probabilities are modeled, and elaborate mainly on the modeling of the acoustic signal.

In Section 2.2, we outline Bayes classification, as applied to speech recognition. The fundamentals of HMMs—how to calculate probabilities with them, how to find the most likely explanation for an observation, and how to estimate their parameters—are given in Section 2.3. Section 2.4 describes how HMMs are used in practical ASR systems. Several issues related to practical implementation are addressed. Recognition is not performed with

Techniques for Noise Robustness in Automatic Speech Recognition, First Edition.

Edited by Tuomas Virtanen, Rita Singh, and Bhiksha Raj.

^{© 2013} John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.

the speech signal itself, but on *features* derived from it. We give a brief review of the most commonly used features in Section 2.4.1. Feature computation is covered in greater detail in Chapters 7 and 8 of the book. The number of possible word sequences that must be investigated in order to determine the most likely one is potentially extremely large. It is infeasible to explicitly characterize the probability distributions of the acoustics for each and every word sequence. In Sections 2.4.2 and 2.4.3, we explain how we can nevertheless explore all of them by *composing* the HMMs for word sequences from smaller units, and how the set of all possible word sequences can be represented as compact graphs that can be searched.

Before proceeding, we note that although this book largely presents speech recognition and robustness issues related to it from the perspective of HMM-based systems, the fundamental ideas presented here, and many of the algorithms and techniques described both in this chapter and elsewhere in the book, carry over to other formalisms that may be employed for speech recognition as well.

2.2 Speech Recognition Viewed as Bayes Classification

At their core, state-of-art ASR systems are fundamentally *Bayesian classifiers*. The Bayesian classification paradigm follows a rather simple intuition: the best guess for the explanation of any observation (such as a recording of speech) is the most *likely* one, given any other information we have about the problem at hand. Mathematically, it can be stated as follows: let C_1, C_2, C_3, \ldots represent all possible explanations for an observation **X**. The Bayesian classification paradigm chooses the explanation C_i such that

$$P(C_i | \mathbf{X}, \theta) \ge P(C_j | \mathbf{X}, \theta) \quad \forall j \neq i,$$
(2.1)

where $P(C_i|\mathbf{X}, \theta)$ is the conditional probability of class C_i given the observation \mathbf{X} , and θ represents all other evidence, or information known *a priori*. In other words, it chooses the *a posteriori* most probable explanation C_i , given the observation and all prior evidence.

For the ASR problem, the problem is now stated as follows. Given a speech recording X, the sequence of words $\hat{w}_1, \hat{w}_2, \cdots$ that were spoken is estimated as

$$\hat{w}_1, \hat{w}_2, \dots = \operatorname*{argmax}_{w_1, w_2, \dots} P(w_1, w_2, \dots | \mathbf{X}, \Lambda).$$

$$(2.2)$$

Here, Λ represents other evidence that we may have about what was spoken. Equation (2.2) states that the "best guess" word sequence $\hat{w}_1, \hat{w}_2 \cdots$ is the word sequence that is *a posteriori* most probable, after consideration of both the recording **X** and all other evidence represented by Λ .

In order to implement Equation (2.2) computationally, the problem is refactored using Bayes' rule as follows:

$$\hat{w}_1, \hat{w}_2, \dots = \operatorname*{argmax}_{w_1, w_2, \dots} P(\mathbf{X}|w_1, w_2, \dots) P(w_1, w_2, \dots | \Lambda).$$
 (2.3)

In the term $P(\mathbf{X}|w_1, w_2, \cdots)$, we assume that the speech signal \mathbf{X} becomes independent of all other factors, once the sequence of words is given. The *true* distribution of \mathbf{X} for any word sequence is not known. Instead it is typically modeled by a *hidden Markov model* (HMM) [2]. Since the term $P(\mathbf{X}|w_1, w_2, \cdots)$ models the properties of the acoustic speech signal, is it termed an *acoustic model*.