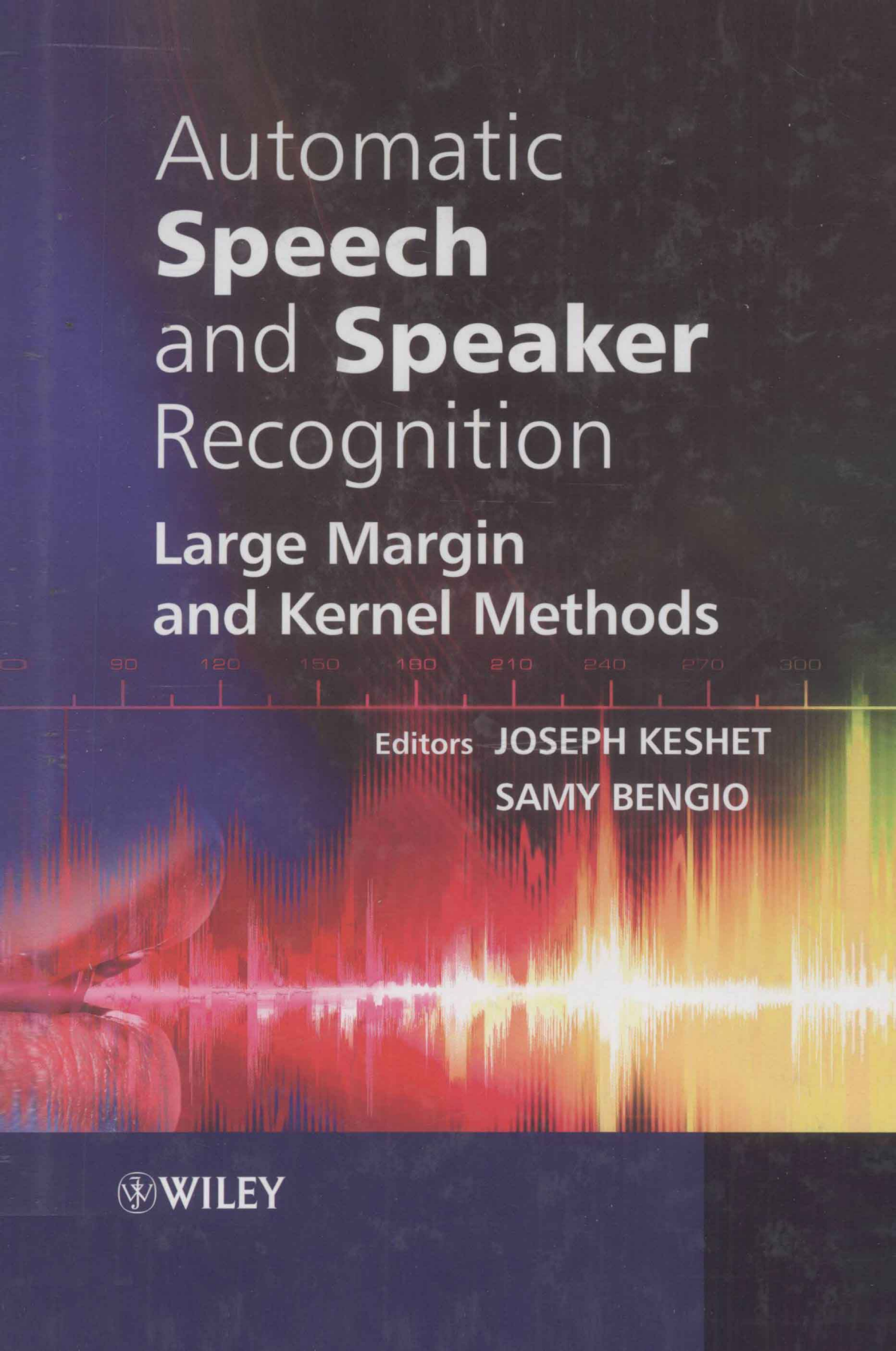


A spectrogram of a speech signal is visible in the background. The horizontal axis represents time in milliseconds, with labels at 90, 120, 150, 180, 210, 240, 270, and 300. The vertical axis represents frequency. The spectrogram shows various formants and harmonic structures, with a prominent horizontal line of energy around 2000 Hz.

Automatic **Speech** and **Speaker** Recognition

Large Margin and Kernel Methods

Editors **JOSEPH KESHET**
SAMY BENGIO

 **WILEY**

Automatic Speech and Speaker Recognition

Large Margin and Kernel Methods

Joseph Keshet

IDIAP Research Institute, Martigny, Switzerland

Samy Bengio

Google Inc., Mountain View, CA, USA



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.

Library of Congress Cataloging-in-Publication Data

Automatic speech and speaker recognition : large margin and kernel methods /
edited by Joseph Keshet, Samy Bengio.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-69683-5 (cloth)

1. Automatic speech recognition. I. Keshet, Joseph. II. Bengio, Samy.

TK7895.S65A983 2009

006.4'54—dc22

2008038551

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

ISBN 9780470696835 (H/B)

Set in 10/12pt Times by Sunrise Setting Ltd, Torquay, UK.
Printed in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire

Automatic Speech and Speaker Recognition

List of Contributors

Yasemin Altun

Department Schölkopf
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
yasemin.altun@tuebingen.mpg.de

Francis R. Bach

INRIA – Willow project
Département d’Informatique
Ecole Normale Supérieure
Paris, France
francis.bach@mines.org

Samy Bengio

Google Research
Google Inc.
Mountain View, CA, USA
bengio@google.com

Dan Chazan

Department of Electrical Engineering
The Technion Institute of Technology
Haifa, Israel
dan_chazan@yahoo.com

Koby Crammer

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, USA
crammer@cis.upenn.edu

Mark J. F. Gales

Department of Engineering
University of Cambridge
Cambridge, UK
mjfg@eng.cam.ac.uk

Yves Grandvalet

Heudiasyc, Unité Mixte 6599
CNRS & Université de Technologie de Compiègne
Compiègne, France
yves.grandvalet@utc.fr

David Grangier

Department of Machine Learning
NEC Laboratories America, Inc.
Princeton, NJ, USA
dgrangier@nec-labs.com

Michael I. Jordan

Computer Science Division and Department of Statistics
University of California
Berkeley, CA, USA
jordan@eecs.berkeley.edu

Joseph Keshet

IDIAP Research Institute
Martigny, Switzerland
jkeshet@idiap.ch

Johnny Mariéthoz

IDIAP Research Institute
Martigny, Switzerland
marietho@idiap.ch

Brian Roark

Center for Spoken Language Understanding
Division of Biomedical Computer Science
Oregon Health & Science University
Portland, Oregon, USA
roark@cslu.ogi.edu

Lawrence K. Saul

Department of Computer Science and Engineering
University of California
San Diego, CA, USA
saul@cs.ucsd.edu

Fei Sha

Computer Science Department
University of Southern California
Los Angeles, CA, USA
feisha@usc.edu

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago
Chicago, USA
shai@tti-c.org

Yoram Singer

Google Research

Google Inc.

Mountain View, CA, USA

singer@google.com

Nathan Srebro

Toyota Technological Institute at Chicago

Chicago, USA

nati@uchicago.edu

Preface

This is the first book dedicated to uniting research related to speech and speaker recognition based on the recent advances in large margin and kernel methods. The first part of the book presents theoretical and practical foundations of large margin and kernel methods, from Support Vector Machines to large margin methods for structured learning. The second part of the book is dedicated to acoustic modeling of continuous speech recognizers, where the grounds for practical large margin sequence learning are set. The third part introduces large margin methods for discriminative language modeling. The last part of the book is dedicated to the application of keyword spotting, speaker verification and spectral clustering.

The book is an important reference to researchers and practitioners in the field of modern speech and speaker recognition. The purpose of the book is twofold: first, to set the theoretical foundation of large margin and kernel methods relevant to the speech recognition domain; second, to propose a practical guide on implementation of these methods to the speech recognition domain. The reader is presumed to have basic knowledge of large margin and kernel methods and of basic algorithms in speech and speaker recognition.

Joseph Keshet

Martigny, Switzerland

Samy Bengio

Mountain View, CA, USA

Contents

List of Contributors	xi
Preface	xv
I Foundations	1
1 Introduction	3
<i>Samy Bengio and Joseph Keshet</i>	
1.1 The Traditional Approach to Speech Processing	3
1.2 Potential Problems of the Probabilistic Approach	5
1.3 Support Vector Machines for Binary Classification	7
1.4 Outline	8
References	9
2 Theory and Practice of Support Vector Machines Optimization	11
<i>Shai Shalev-Shwartz and Nathan Srebro</i>	
2.1 Introduction	11
2.2 SVM and L_2 -regularized Linear Prediction	12
2.2.1 Binary Classification and the Traditional SVM	12
2.2.2 More General Loss Functions	13
2.2.3 Examples	13
2.2.4 Kernels	14
2.2.5 Incorporating a Bias Term	15
2.3 Optimization Accuracy From a Machine Learning Perspective	16
2.4 Stochastic Gradient Descent	18
2.4.1 Sub-gradient Calculus	20
2.4.2 Rate of Convergence and Stopping Criteria	21
2.5 Dual Decomposition Methods	22
2.5.1 Duality	23
2.6 Summary	25
References	26

3	From Binary Classification to Categorical Prediction	27
	<i>Koby Crammer</i>	
3.1	Multi-category Problems	27
3.2	Hypothesis Class	31
3.3	Loss Functions	32
3.3.1	Combinatorial Loss Functions	33
3.4	Hinge Loss Functions	35
3.5	A Generalized Perceptron Algorithm	36
3.6	A Generalized Passive–Aggressive Algorithm	39
3.6.1	Dual Formulation	40
3.7	A Batch Formulation	41
3.8	Concluding Remarks	43
3.9	Appendix. Derivations of the Duals of the Passive–Aggressive Algorithm and the Batch Formulation	44
3.9.1	Derivation of the Dual of the Passive–Aggressive Algorithm	44
3.9.2	Derivation of the Dual of the Batch Formulation	46
	References	48
II	Acoustic Modeling	51
4	A Large Margin Algorithm for Forced Alignment	53
	<i>Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer and Dan Chazan</i>	
4.1	Introduction	54
4.2	Problem Setting	54
4.3	Cost and Risk	55
4.4	A Large Margin Approach for Forced Alignment	56
4.5	An Iterative Algorithm	57
4.6	Efficient Evaluation of the Alignment Function	62
4.7	Base Alignment Functions	64
4.8	Experimental Results	66
4.9	Discussion	67
	References	67
5	A Kernel Wrapper for Phoneme Sequence Recognition	69
	<i>Joseph Keshet and Dan Chazan</i>	
5.1	Introduction	69
5.2	Problem Setting	70
5.3	Frame-based Phoneme Classifier	71
5.4	Kernel-based Iterative Algorithm for Phoneme Recognition	71
5.5	Nonlinear Feature Functions	75
5.5.1	Acoustic Modeling	75
5.5.2	Duration Modeling	77
5.5.3	Transition Modeling	78

5.6	Preliminary Experimental Results	78
5.7	Discussion: Can we Hope for Better Results?	79
	References	80
6	Augmented Statistical Models: Using Dynamic Kernels for Acoustic Models	83
	<i>Mark J. F. Gales</i>	
6.1	Introduction	84
6.2	Temporal Correlation Modeling	84
6.3	Dynamic Kernels	86
6.3.1	Static and Dynamic Kernels	87
6.3.2	Generative Kernels	88
6.3.3	Simple Example	90
6.4	Augmented Statistical Models	92
6.4.1	Generative Augmented Models	92
6.4.2	Conditional Augmented Models	94
6.5	Experimental Results	95
6.6	Conclusions	97
	Acknowledgements	97
	References	98
7	Large Margin Training of Continuous Density Hidden Markov Models	101
	<i>Fei Sha and Lawrence K. Saul</i>	
7.1	Introduction	101
7.2	Background	103
7.2.1	Maximum Likelihood Estimation	103
7.2.2	Conditional Maximum Likelihood	104
7.2.3	Minimum Classification Error	104
7.3	Large Margin Training	105
7.3.1	Discriminant Function	105
7.3.2	Margin Constraints and Hamming Distances	106
7.3.3	Optimization	106
7.3.4	Related Work	107
7.4	Experimental Results	107
7.4.1	Large Margin Training	108
7.4.2	Comparison with CML and MCE	109
7.4.3	Other Variants	109
7.5	Conclusion	112
	References	113
III	Language Modeling	115
8	A Survey of Discriminative Language Modeling Approaches for Large Vocabulary Continuous Speech Recognition	117
	<i>Brian Roark</i>	
8.1	Introduction	117

8.2	General Framework	119
8.2.1	Training Data and the GEN Function	120
8.2.2	Feature Mapping	123
8.2.3	Parameter Estimation	127
8.3	Further Developments	130
8.3.1	Novel Features	130
8.3.2	Novel Objectives	131
8.3.3	Domain Adaptation	132
8.4	Summary and Discussion	133
	References	134
9	Large Margin Methods for Part-of-Speech Tagging	139
	<i>Yasemin Altun</i>	
9.1	Introduction	139
9.2	Modeling Sequence Labeling	140
9.2.1	Feature Representation	141
9.2.2	Empirical Risk Minimization	142
9.2.3	Conditional Random Fields and Sequence Perceptron	143
9.3	Sequence Boosting	144
9.3.1	Objective Function	145
9.3.2	Optimization Method	145
9.4	Hidden Markov Support Vector Machines	149
9.4.1	Objective Function	149
9.4.2	Optimization Method	151
9.4.3	Algorithm	151
9.5	Experiments	153
9.5.1	Data and Features for Part-of-Speech Tagging	153
9.5.2	Results of Sequence AdaBoost	154
9.5.3	Results of Hidden Markov Support Vector Machines	155
9.6	Discussion	156
	References	156
10	A Proposal for a Kernel Based Algorithm for Large Vocabulary Continuous Speech Recognition	159
	<i>Joseph Keshet</i>	
10.1	Introduction	159
10.2	Segment Models and Hidden Markov Models	161
10.3	Kernel Based Model	163
10.4	Large Margin Training	164
10.5	Implementation Details	166
10.5.1	Iterative Algorithm	166
10.5.2	Recognition Feature Functions	167
10.5.3	The Decoder	169
10.5.4	Complexity	169

10.6 Discussion	170
Acknowledgements	170
References	170

IV Applications **173**

11 Discriminative Keyword Spotting **175**

David Grangier, Joseph Keshet and Samy Bengio

11.1 Introduction	175
11.2 Previous Work	177
11.3 Discriminative Keyword Spotting	180
11.3.1 Problem Setting	180
11.3.2 Loss Function and Model Parameterization	182
11.3.3 An Iterative Training Algorithm	184
11.3.4 Analysis	185
11.4 Experiments and Results	188
11.4.1 The TIMIT Experiments	188
11.4.2 The WSJ Experiments	190
11.5 Conclusions	191
Acknowledgements	193
References	193

12 Kernel-based Text-independent Speaker Verification **195**

Johnny Mariéthoz, Samy Bengio and Yves Grandvalet

12.1 Introduction	196
12.2 Generative Approaches	197
12.2.1 Rationale	197
12.2.2 Gaussian Mixture Models	198
12.3 Discriminative Approaches	199
12.3.1 Support Vector Machines	199
12.3.2 Kernels	200
12.4 Benchmarking Methodology	201
12.4.1 Data Splitting for Speaker Verification	201
12.4.2 Performance Measures	202
12.4.3 NIST Data	203
12.4.4 Pre-processing	203
12.5 Kernels for Speaker Verification	203
12.5.1 Mean Operator Sequence Kernels	204
12.5.2 Fisher Kernels	205
12.5.3 Beyond Fisher Kernels	210
12.6 Parameter Sharing	212
12.6.1 Nuisance Attribute Projection	213
12.6.2 Other Approaches	214
12.7 Is the Margin Useful for This Problem?	215
12.8 Comparing all Methods	216

12.9 Conclusion	218
References	219

13 Spectral Clustering for Speech Separation 221

Francis R. Bach and Michael I. Jordan

13.1 Introduction	221
13.2 Spectral Clustering and Normalized Cuts	223
13.2.1 Similarity Matrices	223
13.2.2 Normalized Cuts	223
13.2.3 Spectral Relaxation	225
13.2.4 Rounding	226
13.2.5 Spectral Clustering Algorithms	227
13.2.6 Variational Formulation for the Normalized Cut	229
13.3 Cost Functions for Learning the Similarity Matrix	229
13.3.1 Distance Between Partitions	230
13.3.2 Cost Functions as Upper Bounds	230
13.3.3 Functions of Eigensubspaces	231
13.3.4 Empirical Comparisons Between Cost Functions	233
13.4 Algorithms for Learning the Similarity Matrix	234
13.4.1 Learning Algorithm	236
13.4.2 Related Work	236
13.4.3 Testing Algorithm	236
13.4.4 Handling very Large Similarity Matrices	237
13.4.5 Simulations on Toy Examples	239
13.5 Speech Separation as Spectrogram Segmentation	239
13.5.1 Spectrogram	240
13.5.2 Normalization and Subsampling	241
13.5.3 Generating Training Samples	241
13.5.4 Features and Grouping Cues for Speech Separation	242
13.6 Spectral Clustering for Speech Separation	244
13.6.1 Basis Similarity Matrices	244
13.6.2 Combination of Similarity Matrices	244
13.6.3 Approximations of Similarity Matrices	245
13.6.4 Experiments	245
13.7 Conclusions	247
References	248

Index 251

Part I

Foundations

1

Introduction

Samy Bengio and Joseph Keshet

One of the most natural communication tools used by humans is their voice. It is hence natural that a lot of research has been devoted to analyzing and understanding human uttered speech for various applications. The most obvious one is *automatic speech recognition*, where the goal is to transcribe a recorded speech utterance into its corresponding sequence of words. Other applications include *speaker recognition*, where the goal is to determine either the claimed identity of the speaker (verification) or who is speaking (identification), and speaker segmentation or diarization, where the goal is to segment an acoustic sequence in terms of the underlying speakers (such as during a dialog).

Although an enormous amount of research has been devoted to speech processing, there appears to be some form of local optimum in terms of the fundamental tools used to approach these problems. The aim of this book is to introduce the speech researcher community to radically different approaches based on more recent kernel based machine learning methods. In this introduction, we first briefly review the predominant speech processing approach, based on hidden Markov models, as well as its known problems; we then introduce the most well known kernel based approach, the Support Vector Machine (SVM), and finally outline the various contributions of this book.

1.1 The Traditional Approach to Speech Processing

Most speech processing problems, including speech recognition, speaker verification, speaker segmentation, etc., proceed with basically the same general approach, which is described here in the context of speech recognition, as this is the field that has attracted most of the research in the last 40 years. The approach is based on the following statistical framework.

A sequence of acoustic feature vectors is extracted from a spoken utterance by a front-end signal processor. We denote the sequence of acoustic feature vectors by $\bar{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$,

Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods Joseph Keshet and Samy Bengio
© 2009 John Wiley & Sons, Ltd