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

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

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 Srebo
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

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 $\bar{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$,

Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods Joseph Keshet and Samy Bengio