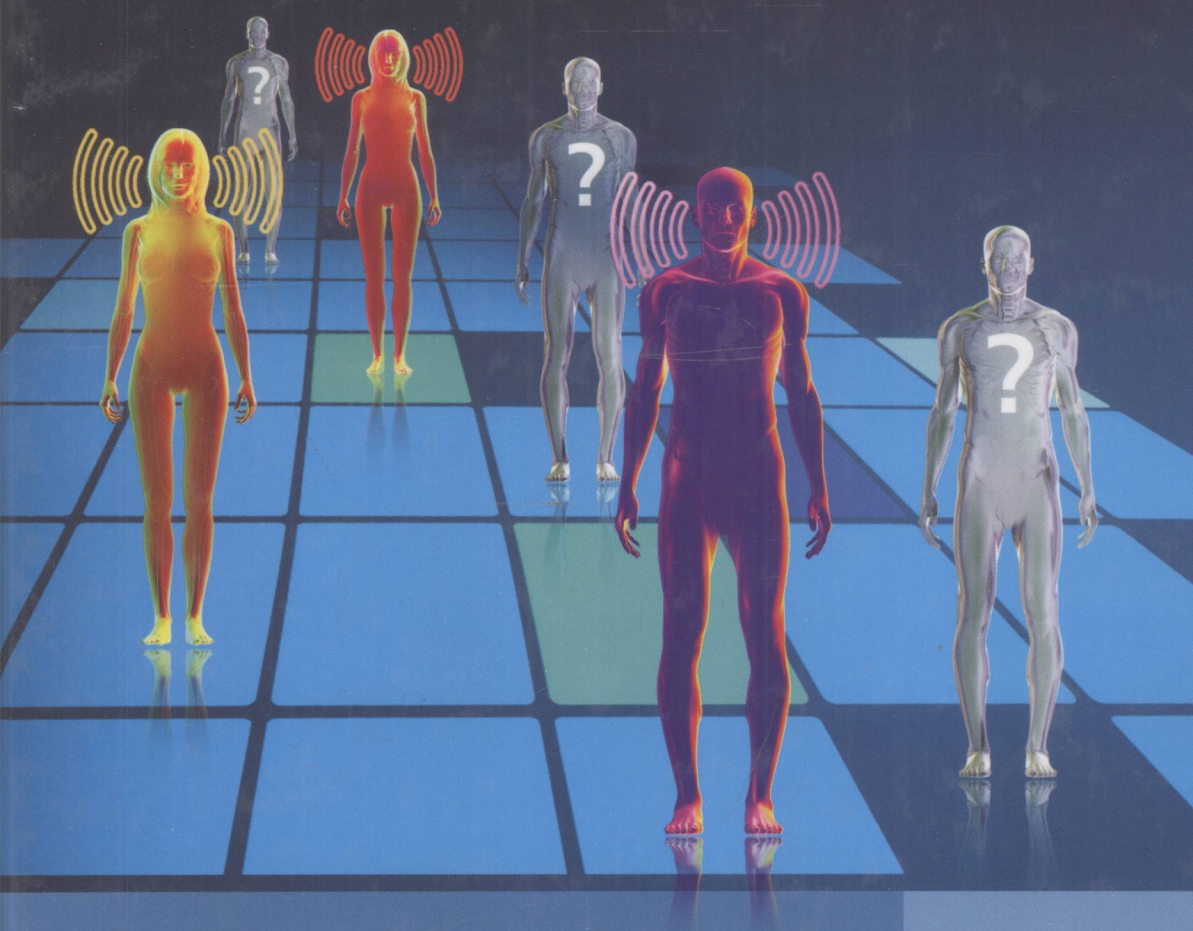
MATTHIAS WÖLFEL AND JOHN MCDONOUGH

DISTANT SPEECH RECOGNITION



WILEY

TN912.3
W854

DISTANT SPEECH RECOGNITION

Matthias Wölfel
Universität Karlsruhe (TH), Germany

and

John McDonough
Universität des Saarlandes, Germany

WILEY
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 CPI Antony Rowe, Chippenham, Wiltshire

DISTANT SPEECH RECOGNITION

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

此为试读,需要完整PDF请访问: www.ertongbook.com

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 noteable 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

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

此为试读,需要完整PDF请访问: www.ertongbook.com