Editors
Tuomas Virtanen | Rita Singh | Bhiksha Raj

Techniques for Noise Robustness in Automatic Speech Recognition

WILEY

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

A John Wiley & Sons, Ltd., Publication

This edition first published 2013

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

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.S65V57 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

Part Four MODEL ENHANCEMENT

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 Virtanen[1], Rita Singh[2], Bhiksha Raj[2]
[1]*Tampere University of Technology, Finland*
[2]*Carnegie 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.
 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
$a, b, c, \ldots$	Scalar variables
$A, B, C, \ldots$	Constants
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	Vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	Matrices
$\otimes$	Convolution
$\mathcal{N}$	Normal distribution
$\mathcal{E}\{x\}$	Expected value of x
$\mathbf{A}^{\mathsf{T}}$	Transpose of matrix $\mathbf{A}$
$x_{i:j}$	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 t
$\boldsymbol{\mu}$	Mean vector
$\boldsymbol{\Theta}, \boldsymbol{\Sigma}$	Covariance matrix
P, p	Probability

Part One
Foundations

2

The Basics of Automatic Speech Recognition

Rita Singh[1], Bhiksha Raj[1], Tuomas Virtanen[2]
[1]*Carnegie Mellon University, USA*
[2]*Tampere 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.
 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 $\mathbf{X}$. The Bayesian classification paradigm chooses the explanation C_i such that

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

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 $\mathbf{X}$, the sequence of words $\hat{w}_1, \hat{w}_2, \cdots$ that were spoken is estimated as

$$\hat{w}_1, \hat{w}_2, \cdots = \underset{w_1, w_2, \cdots}{\operatorname{argmax}} P(w_1, w_2, \cdots|\mathbf{X}, \Lambda). \tag{2.2}$$

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 $\mathbf{X}$ and all other evidence represented by Λ.

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

$$\hat{w}_1, \hat{w}_2, \cdots = \underset{w_1, w_2, \cdots}{\operatorname{argmax}} P(\mathbf{X}|w_1, w_2, \cdots)P(w_1, w_2, \cdots|\Lambda). \tag{2.3}$$

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*.