

XUEDONG HUANG

ALEX ACERO

HSIAO-WUEN HON

S P O K E N

LANGUAGE PROCESSING

A Guide to Theory, Algorithm, and System Development

Foreword by Dr. Raj Reddy
Carnegie Mellon University

Harman Ex. 1012
Harman v. ST CasesTech
IPR2024-01302
Page 00001

Library of Congress Cataloging-in-Publication Data

Huang, Xuedong.

Spoken language processing: a guide to theory, algorithm, and system development/

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon.

p. cm.

Includes bibliographical references and index.

ISBN 0-13-022616-5

I. Natural language processing (Computer science) I. Acero, Alex. II. Hon, Hsiao-Wuen. III. Title.

QA76.9.N38 H83 2001

006.3'5—dc21

00-050196

Editorial/production supervision: *Jane Bonnell*

Cover design director: *Jerry Votta*

Cover design: *Anthony Gemmellaro*

Manufacturing buyer: *Maura Zaldivar*

Development editor: *Russ Hall*

Acquisitions editor: *Tim Moore*

Editorial assistant: *Allyson Kloss*

Marketing manager: *Debby van Dijk*



© 2001 by Prentice Hall PTR

Prentice-Hall, Inc.

Upper Saddle River, New Jersey 07458

Prentice Hall books are widely used by corporations and government agencies for training, marketing, and resale.

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact Corporate Sales Department, Phone: 800-382-3419; FAX: 201-236-7141;

E-mail: corpsales@prenhall.com

Or write: Prentice Hall PTR, Corporate Sales Dept., One Lake Street, Upper Saddle River, NJ 07458.

Company and product names mentioned herein are the trademarks or registered trademarks of their respective owners.

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-022616-5

Prentice-Hall International (UK) Limited, *London*

Prentice-Hall of Australia Pty. Limited, *Sydney*

Prentice-Hall Canada Inc., *Toronto*

Prentice-Hall Hispanoamericana, S.A., *Mexico*

Prentice-Hall of India Private Limited, *New Delhi*

Prentice-Hall of Japan, Inc., *Tokyo*

Pearson Education Asia Pte. Ltd.

Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

speaker and rejects signals coming from other angles. Reverberation, as discussed in Section 10.1.2, can be combated with these techniques. Blind source separation techniques are another family of statistical techniques that typically do not use spatial constraints, but rather statistical independence between different sources.

While in this section we describe only linear processing, i.e., the output speech is a linearly filtered version of the microphone signals, we could also combine these techniques with the nonlinear methods of Section 10.5.

10.4.1. Microphone Arrays

The goals of microphone arrays are twofold: finding the position of a sound source in a room, and improving the SNR of the received signal. *Steering* is helpful in videoconferencing, where a camera has to follow the current speaker. Since the speaker is typically far away from the microphone, the received signal likely contains a fair amount of additive noise. Microphone arrays can also be used to increase the SNR.

Let $x[n]$ be the signal at the source S . Microphone i picks up a signal

$$y_i[n] = x[n] * g_i[n] + v_i[n] \quad (10.60)$$

that is a filtered version of the source plus additive noise $v_i[n]$. If we have N such microphones, we can attempt to recover $s[n]$ because all the signals $y_i[n]$ should be correlated.

A typical assumption made is that all the filters $g_i[n]$ are delayed versions of the same filter $g[n]$

$$g_i[n] = g[n - D_i] \quad (10.61)$$

with the delay $D_i = d_i / c$, d_i being the distance between the source S and microphone i , and c the speed of sound in air. We cannot recover signal $x[n]$ without knowledge of $g[n]$ or the signal itself, so the goal is to obtain the filtered signal $y[n]$

$$y[n] = x[n] * g[n] \quad (10.62)$$

so that, combining Eqs. (10.60), (10.61), and (10.62),

$$y_i[n] = y[n - D_i] + v_i[n] \quad (10.63)$$

Assuming $v_i[n]$ are independent and Gaussianly distributed, the optimal estimate of $x[n]$ is given by

$$\tilde{y}[n] = \frac{1}{N} \sum_{i=0}^{N-1} y_i[n + D_i] = y[n] + v[n] \quad (10.64)$$

which is the so-called *delay-and-sum beamformer* [24, 29], where the residual noise $v[n]$

$$v[n] = \frac{1}{N} \sum_{i=0}^{N-1} v_i[n + D_i] \quad (10.65)$$

has a variance that decreases as the number of microphones N increases, since the noises $v_i[n + D_i]$ are uncorrelated.

Equation (10.65) requires estimation of the delays D_i . To attenuate the additive noise $v[n]$, it is not necessary to identify the absolute delays, but rather the delays relative to one reference microphone (for example, the center microphone). It can be shown that the maximum likelihood solution consists in maximizing the energy of $\tilde{y}[n]$ in Eq. (10.64), which is the sum of cross-correlations:

$$D_i = \arg \max_{D_i} \left(\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} R_{ij}[D_i - D_j] \right) \quad 0 \leq i < N \quad (10.66)$$

This approach assumes that we know nothing about the geometry of the microphone placement. In fact, given a point source and assuming no reflections, we can compute the delay based on the distance between the source and the microphone. The use of geometry allows us to reduce the number of parameters to estimate from $(N - 1)$ to a maximum of 3, in case we desire to estimate the exact location. This location is often described in spherical coordinates (φ, θ, ρ) with φ being the direction of arrival, θ the elevation angle, and ρ the distance to the reference microphone, as shown in Figure 10.21.

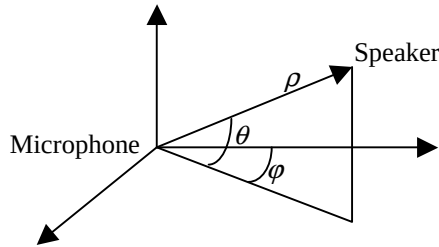


Figure 10.21 Spherical coordinates (φ, θ, ρ) with φ being the direction of arrival, θ the elevation angle, and ρ the distance to the reference microphone.

While 2-D and 3-D microphone configurations can be used, which would allow us to determine not just the steering angle φ , but also distance to the origin ρ and azimuth θ , linear microphone arrays are the most widely used configurations because they are the simplest. In a linear array all the microphones are placed on a line (see Figure 10.22). In this case, we cannot determine the elevation angle θ . To determine both φ and ρ we need at least two microphones in the array.

If the microphones are relatively close to each other compared to the distance to the source, the angle of arrival φ is approximately the same for all signals. With this assumption, the normalized delay $\bar{D}_i$ with respect to the reference microphone is given by

XUEDONG HUANG

ALEX ACERO

HSIAO-WUEN HON

S P O K E N

LANGUAGE PROCESSING

A Guide to Theory, Algorithm, and System Development

■ New advances
in spoken language
processing: *theory
and practice*

■ In-depth coverage of
speech processing,
speech recognition,
speech synthesis, spoken
language understanding,
and speech interface
design

■ Many case studies from
state-of-the-art
systems, including
examples from
Microsoft's advanced
research labs

Spoken Language Processing draws on the latest advances and techniques from multiple fields: computer science, electrical engineering, acoustics, linguistics, mathematics, psychology, and beyond. Starting with the fundamentals, it presents all this and more:

- Essential background on speech production and perception, probability and information theory, and pattern recognition
- Extracting information from the speech signal: useful representations and practical compression solutions
- Modern speech recognition techniques: hidden Markov models, acoustic and language modeling, improving resistance to environmental noises, search algorithms, and large vocabulary speech recognition
- Text-to-speech: analyzing documents, pitch and duration controls, trainable synthesis, and more
- Spoken language understanding: dialog management, spoken language applications, and multimodal interfaces

To illustrate the book's methods, the authors present detailed case studies based on state-of-the-art systems, including Microsoft's Whisper speech recognizer, Whistler text-to-speech system, Dr. Who dialog system, and the MPad handheld device. Whether you're planning, designing, building, or purchasing spoken language technology, this is the state of the art—from algorithms through business productivity.

About the Authors

XUEDONG HUANG is founder and head of the Speech Technology Group at Microsoft Research. He received his Ph.D. from the University of Edinburgh. He is an IEEE Fellow.

ALEX ACERO and HSIAO-WUEN HON are Senior Researchers at Microsoft Research and Senior Members of IEEE. Both received doctorates from Carnegie Mellon University.

PRENTICE HALL
Upper Saddle River, NJ 07458
www.pphtr.com

ISBN 0-13-022616-5



9 780130 226167