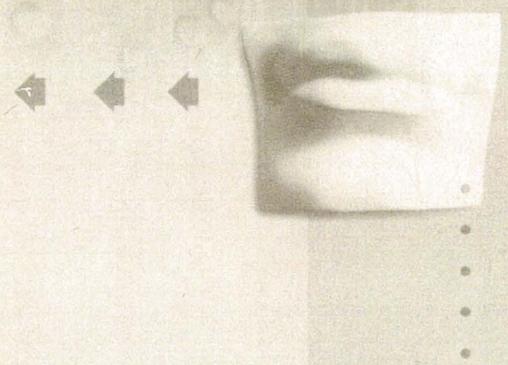
XUEDONG HUANG | ALEX ACERO | HSIAO-WUEN HON

SPOKEN LANGUAGE PROCESSING

A Guide to Theory, Algorithm, and System Development



Foreword by Dr. Raj Reddy
Carnegie Mellon University

Spoken Language Processing

Spoken Language Processing

A Guide to Theory, Algorithm, and System Development

Xuedong Huang
Alex Acero
Hsiao-Wuen Hon

Microsoft Research

Prentice Hall PTR
Upper Saddle River, New Jersey 07458
www.phptr.com

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
1. 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*

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.

Printed in the United States of America

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*

To Yingzhi, Angela, Christina, and Derek

To Donna and Nicolas

To Phen, Stephanie, and Jacqueline

Contents

PART III: SPEECH RECOGNITION

PART IV: TEXT-TO-SPEECH SYSTEMS

Foreword

Recognition and understanding of spontaneous unrehearsed speech remains an elusive goal. To understand speech, a human considers not only the specific information conveyed to the ear, but also the context in which the information is being discussed. For this reason, people can understand spoken language even when the speech signal is corrupted by noise. However, understanding the context of speech is, in turn, based on a broad knowledge of the world. And this has been the source of the difficulty and over forty years of research.

It is difficult to develop computer programs that are sufficiently sophisticated to understand continuous speech by a random speaker. Only when programmers simplify the problem—by isolating words, limiting the vocabulary or number of speakers, or constraining the way in which sentences may be formed—is speech recognition by computer possible.

Since the early 1970s, researchers at AT&T, BBN, CMU, IBM, Lincoln Labs, MIT, and SRI have made major contributions in Spoken Language Understanding Research. In 1971, the Defense Advanced Research Projects Agency (DARPA) initiated an ambitious five-year, $15 million, multisite effort to develop speech understanding systems. The goals were to develop systems that would accept continuous speech from many speakers, with minimal speaker adaptation, and operate on a 1000-word vocabulary, artificial syntax, and a

constrained task domain. Two of the systems, Harpy and Hearsay-II, both developed at Carnegie Mellon University, achieved the original goals and in some instances surpassed them.

During the last three decades I have been at Carnegie Mellon, I have been very fortunate to be able to work with many brilliant students and researchers. Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon were arguably among the outstanding researchers in the speech group at CMU. Since then, they have moved to Microsoft and have put together a world-class team at Microsoft Research. Over the years, they have contributed standards for building spoken language understanding systems with Microsoft's SAPI/SDK family of products and pushed the technologies forward with the rest of the community. Today, they continue to play a premier leadership role in both the research community and in industry.

This new book, *Spoken Language Processing*, represents a welcome addition to the technical literature on this increasingly important emerging area of Information Technology. As we move from desktop PCs to personal digital assistants (PDAs), wearable computers, and Internet cell phones, speech becomes a central, if not the only, means of communication between the human and machine! Huang, Acero, and Hon have undertaken a commendable task of creating a comprehensive reference that covers theoretical, algorithmic, and systems aspects of the spoken language tasks of recognition, synthesis, and understanding.

The task of spoken language communication requires a system to recognize, interpret, execute, and respond to a spoken query. This task is complicated by the fact that the speech signal is corrupted by many sources: noise in the background, characteristics of the microphone, vocal tract characteristics of the speakers, and differences in pronunciation. In addition, the system has to cope with non-grammaticality of spoken communication and ambiguity of language. An effective system must strive to utilize all the available sources of knowledge—acoustics, phonetics and phonology, lexical, syntactic, and semantic structure of language, and task-specific context-dependent information.

Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics. While the human vocal apparatus can produce an almost infinite number of articulatory gestures, the number of phonemes is limited. English as spoken in the United States, for example, contains 16 vowel and 24 consonant sounds. Each phoneme has distinguishable acoustic characteristics and, in combination with other phonemes, forms larger units such as syllables and words. Knowledge about the acoustic differences among these sound units is essential to distinguish one word from another, say, *bit* from *pit*.

When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment because of the interaction among various anatomical structures (such as the tongue, lips, and vocal chords) and their different degrees of sluggishness. The result is an overlap of phonemic information in the acoustic signal from one segment to the other. For example, the same underlying phoneme *t* can have drastically different acoustic characteristics in different words, say, in *tea*, *tree*, *city*, *beaten*, and *steep*. This effect, known as coarticulation, can occur within a given word or across a word boundary. Thus, the word *this* will have very different acoustic properties in phrases such as *this car* and *this ship*.

This book is self-contained for those who wish to familiarize themselves with the current state of spoken language systems technology. However, a researcher or a professional in the field will benefit from a thorough grounding in a number of disciplines, including:

- *Signal processing*: Fourier Transforms, DFT, and FFT
- *Acoustics*: physics of sounds and speech, models of vocal tract
- *Pattern recognition*: clustering and pattern matching techniques
- *Artificial intelligence*: knowledge representation and search, natural language processing
- *Computer science*: hardware, parallel systems, algorithm optimization
- *Statistics*: probability theory, hidden Markov models, dynamic programming
- *Linguistics*: acoustic phonetics, lexical representation, syntax, and semantics

A newcomer to this field, easily overwhelmed by the vast number of different algorithms scattered across many conference proceedings, can find in this book a set of techniques that Huang, Acero, and Hon have found to work well in practice. This book is unique in that it includes both the theory and implementation details necessary to build spoken language systems. If you were able to assemble all the individual material that is covered in the book and put it on a shelf, it would be several times larger than this volume and yet you would be missing vital information. You would not have the material that is in this book that threads it all into one story, one context. If you need additional resources, the authors include extensive references to get that additional detail. *Spoken Language Processing* is very appealing both as a textbook and as a reference book for practicing engineers. Some readers familiar with a specific topic may decide to skip a few chapters; others may want to focus in other chapters. This is not a book that you will pick up and read once from cover to cover, but one you will keep near you for reference as long as you work in this field.

Raj Reddy
Dean, School of Computer Science
Carnegie Mellon University

Preface

Our primary motivation in writing this book is to share our working experience to bridge the gap between the knowledge of industry gurus and newcomers to the spoken language processing community. Many powerful techniques hide in conference proceedings and academic papers for years before becoming widely recognized by the research community or the industry. We spent many years pursuing spoken language technology research at Carnegie Mellon University before we started spoken language R&D at Microsoft. We fully understand that it is by no means a small undertaking to transfer a state-of-the-art spoken language research system into a commercially viable product that can truly help people improve their productivity. Our experience in both industry and academia is reflected in the context of this book, which presents a contemporary and comprehensive description of both theoretic and practical issues in spoken language processing. This book is intended for people of diverse academic and practical backgrounds. Speech scientists, computer scientists, linguists, engineers, physicists, and psychologists all have a unique perspective on spoken language processing. This book will be useful to all of these special interest groups.

Spoken language processing is a diverse subject that relies on knowledge of many levels, including acoustics, phonology, phonetics, linguistics, semantics, pragmatics, and discourse. The diverse nature of spoken language processing requires knowledge in computer science, electrical engineering, mathematics, syntax, and psychology. There are a number of excellent books on the subfields of spoken language processing, including speech recognition, text-to-speech conversion, and spoken language understanding, but there is no single book that covers both theoretical and practical aspects of these subfields and spoken language interface design. We devote many chapters systematically introducing fundamental

theories needed to understand how speech recognition, text-to-speech synthesis, and spoken language understanding work. Even more important is the fact that the book highlights what works well in practice, which is invaluable if you want to build a practical speech recognizer, a practical text-to-speech synthesizer, or a practical spoken language system. Using numerous real examples in developing Microsoft's spoken language systems, we concentrate on showing how the fundamental theories can be applied to solve real problems in spoken language processing.

We would like to thank many people who helped us during our spoken language processing R&D careers. We are particularly indebted to Professor Raj Reddy at the School of Computer Science, Carnegie Mellon University. Under his leadership, Carnegie Mellon University has become a center of research excellence on spoken language processing. Today's computer industry and academia benefit tremendously from his leadership and contributions.

Special thanks are due to Microsoft for its encouragement of spoken language R&D. The management team at Microsoft has been extremely generous to the speech technology group. We are particularly grateful to Bill Gates, Nathan Myhrvold, Rick Rashid, Dan Ling, and Jack Breese for the great environment they have created for us at Microsoft Research. We would also like to thank Bob Muglia and Kai-Fu Lee for their leadership role in Microsoft's speech product development.

Scott Meredith helped us write a number of chapters in this book and deserves to be a co-author. His insight and experience in text-to-speech synthesis enriched this book a great deal. We also owe gratitude to many colleagues we worked with in the speech technology group of Microsoft Research. In alphabetic order, Jim Adcock, Bruno Alabiso, Fil Alleva, Eric Bidstrup, Antonio Bigazzi, Ciprian Chelba, Li Deng, James Droppo, Doug Duchene, Joshua Goodman, Mei-Yuh Hwang, Larry Israel, Derek Jacoby, Li Jiang, Yun-Cheng Ju, David Larson, Kevin Larson, Jingsong Liu, Ricky Loynd, Milind Mahajan, Peter Mau, John Merrill, Yunus Mohammed, Salman Mughal, Mike Plumpe, Scott Quinn, Bill Rockenbeck, Mike Rozak, Kevin Schofield, Roxana Teodorescu, Gina Venolia, Kuansan Wang, Ye-Yi Wang, and Shenzhi Zhang.

In addition, we want to thank Les Atlas, Jeff Bilmes, Alan Black, David Caulton, Eric Chang, Phil Chou, Dinei Florencio, Allen Gersho, Francisco Gimenez-Galanes, Hynek Hermansky, Henrique Malvar, Julian Odell, Mari Ostendorf, Joseph Pentheroudakis, Tandy Trower, and Charles Wayne. They provided us with many wonderful comments to refine this book. Tim Moore, Russ Hall, and Jane Bonnell at Prentice Hall helped us finish this book in a finite amount of time.

Finally, writing this book was a marathon that could not have been finished without the support of our spouses, Yingzhi, Donna, and Phen, during the many evenings and weekends we spent on this project.

Xuedong Huang
Alex Acero
Hsiao-Wuen Hon
Redmond, WA