AUTOMATIC SPEECH AND SPEAKER RECOGNITION

AUTOMATIC SPEECH AND SPEAKER RECOGNITION

EDITED BY

Chin-Hui Lee
Frank K. Soong
AT&T Bell Laboratories
Murray Hill, New Jersey, USA

■

Kuldip K. Paliwal
School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia

KLUWER ACADEMIC PUBLISHERS
Boston/London/Dordrecht

iv

i

CONTENTS

CONTRIBUTORS

Alejandro Acero
Microsoft Corporation
Redmond, WA 98052, USA

F. Alleva
Microsoft Corporation
Redmond, WA 98052, USA

Ted H. Applebaum
Speech Technology Laboratory
Santa Barbara, CA 93105, USA

Xavier Aubert
Philips GmbH Forschungslaboratorien
D-52066 Aachen, Germany

Lalit R. Bahl
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Larry Bahler
ITT ACD
San Diego, CA 92131, USA

Jerome R. Bellegarda
Apple Computer
Cupertino, CA 95014, USA

Hervé Bourlard
ICSI
Berkeley, CA 94704, USA

Wu Chou
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

S. K. Das
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Sadaoki Furui
NTT Human Interface Laboratories
Tokyo, 180 Japan

Jean-Luc Gauvain
LIMSI/CNRS
Orsay, France

P. S. Gopalakrishnan
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Brian A. Hanson
Speech Technology Laboratory
Santa Barbara, CA 93105, USA

Alan Higgins
ITT ACD
San Diego, CA 92131, USA

Mike Hochberg
Nuance Communications
Menlo Park, CA 94025, USA

Xuedong Huang
Microsoft Corporation
Redmond, WA 98052, USA

Mei-Yuh Hwang
Microsoft Corporation
Redmond, WA 98052, USA

L. Jiang
Microsoft Corporation
Redmond, WA 98052, USA

Biing-Hwang Juang
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Jean-Claude Junqua
Speech Technology Laboratory
Santa Barbara, CA 93105, USA

Chin-Hui Lee
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Fu-Hua Liu
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA

Andrej Ljolje
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

M. Mahajan
Microsoft Corporation
Redmond, WA 98052, USA

John Makhoul
BBN Corporation
Cambridge, MA 02138, USA

Nelson Morgan
ICSI
Berkeley, CA 94704, USA

Hermann Ney
University of Technology
D-52056 Aachen, Germany

Yves Normandin
CRIM
Montréal (Québec) Canada

Long Nguyen
BBN Corporation
Cambridge, MA 02138, USA

Yoshiaki Ohshima
IBM Tokyo Research Center
Tokyo, Japan

Mari Ostendorf
Boston University
Boston, MA, USA

Michael A. Picheny
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

J. Porter
ITT ACD
San Diego, CA 92131, USA

Lawrence R. Rabiner
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Steve Renals
University of Sheffield
Sheffield, S1 4DP, UK.

Michael D. Riley
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Tony Robinson
Cambridge University
Cambridge, CB2 1PZ, U.K.

Richard C. Rose
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Shigeki Sagayama
NTT Human Interface Laboratories
Yokosuka, Japan

Richard Schwartz
BBN Corporation
Cambridge, MA 02138, USA

Richard M. Stern
Carnegie Mellon University
Pittsburgh, PA 15213, USA

PREFACE

The term speech and speaker recognition often refers to the science and technology of developing algorithms and implementing them on machines to recognize the linguistic content in a spoken utterance and to identify the talker who speaks the utterance. Since speech is the most natural means of communication among human beings, it also plays a key role in the development of a natural interface to enhance human-machine communication. In order to realize such an interface, it is essential that the speech communication process becomes well understood so that it can be applied at the design stage.

Research in the field of automatic speech and speaker recognition has made a number of significant advances in the last two decades, spurred on by advances in signal processing, algorithms, architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, the use of the hidden Markov modeling framework to characterize both the spectral and the temporal variations in the speech signal, the use of a rich set of speech utterance examples from a large population of speakers to train the hidden Markov models of some fundamental speech units, the organization of speech and language knowledge sources into a structural finite state network, and the use of dynamic programming based heuristic search methods to find the best word sequence in the lexical network corresponding to the spoken utterance.

Speech and speaker recognition systems have been developed for a wide variety of applications, ranging from small vocabulary keyword recognition over dial-up telephone lines, to medium size vocabulary voice interactive command and control systems on personal computers, to large vocabulary speech dictation, spontaneous speech understanding, and limited-domain speech translation. Although we have learned a great deal about how to build practical and useful speech recognition systems, there remains a whole range of fundamental questions about the technology for which we have no definitive answers. The topics discussed in this book address some of these fundamental questions.

As the title indicates, this book groups together a number of advanced topics on speech and speaker recognition in a single volumn. These topics are of fundamental importance, not yet covered in details in existing textbooks. The book is intended to serve as a reference book for speech researchers and for graduate students who are interested in pursuing potential research topics in speech and speaker recognition. The outline of the book is as follows. Although no explicit partition is given, the book is divided into 5 parts: (1) Chapters 1-2 are devoted to technology overviews; (2) Chapters 3-12 discuss acoustic modeling of fundamental speech units and lexical modeling of words and pronounciations; (3) Chapters 13-15 address the issues related to flexibility and robustness; (4) Chapters 16-18 concern the theoretical and practical issues of search; and (5) Chapters 19-20 give two examples of algorithmic and implementational aspects for recognition system realization. All the chapters are presented in a way that they can be studied independently. When appropriate, an extensive list of references is provided at the end of each chapter which allows the reader to explore related topics in the subject area of the chapter.

In the following, a road map for exploring the book is laid out. Chapter 1 gives a critical overview of the state-of-the-art of the automatic speech recognition technology. It also attempts to summarize the current capabilities and limitations of the technology. In addition, it issues a number of research challenges to expand the capabilities and to overcome some of the limitations. Chapter 2 presents a comprehensive review of the speaker recognition technology. Characterization of speaker individuality and how it can be applied to improve speech recognition are also discussed.

The topic of acoustic modeling has been an active area of research focus in the last two decades. Key advances on this topic have contributed greatly to the recent progress in speech and speaker recognition. This important topic is covered extensively in Part 2 of this book. Assuming the readers have already some knowledge about maximum likelihood estimation of hidden Markov models, it starts with three chapters on alternative HMM estimation algorithms, namely maximum mutual information, maximum a posteriori, and minimum classification error estimation respectively. They are then followed by two chapters on precise acoustic modeling: one on context dependent vector quantization to generate accurate signal labels and the other on hidden Markov network which employs successive state splitting principle to generate acoustic model state network. The next chapter discusses framework to go from frame to segment modeling. It is then followed by a unique nonparametric formulation for modeling speech and speakers. In addition to hidden Markov modeling, artificial neural network approaches have been proposed for acoustic modeling. The next two chapters are devoted to this new computing paradigm, the first on the use of recurrent neural networks and the second on the use of hybrid connectionist models. The last chapter of Part 2 is concerned with automatic generation of pronunciation networks. It also touches briefly on the important issue of lexical modeling.

The three chapters in Part 3 deal with technology limitation on the issues of speaking format flexibility and performance robustness. Since a conventional speech recognition system only accepts speech utterances that follow a rigid speaking format, it has a major problem in recognizing key words and phrases embedded in undesirable out-of-vocabulary signal segments. It also has problem rejecting utterances that do not contain any keyword. Word spotting algorithms can be designed to alleviate some of the difficulties and improve the system flexibility by relaxing these constraints on the speakers. Another major problem that prevents widespread deployment of recognition systems today is the robustness issue. When an acoustic mismatch between training and testing conditions occurs, the performance of a speech recognizer is often degraded. The first robustness chapter discuss one aspect of this issue by incorporating robust signal features such as spectral dynamics which are less affected by the speaking environment than many conventional parameters. The second chapter approaches the problem from a different perspective by feature compensation so as to adaptively make features consistent with the speaking environment.

The other key technology that helps advancing recognition technology is dynamic programming (DP) based heuristic search. Knowledge sources required for speech recognition, including semantic, syntactic, lexical, morphemic, syllabic, phonemic, and acoustic knowledge, are usually organized in a hierarchical network representation. Dynamic programming search is then used to find the most likely word sequence that corresponds to the spoken utterance given the knowledge source constraints. The first chapter in Part 4 systematically introduces the fundamentals of optimal DP search. It is followed by the chapter on fast matching methods that try to maintain admissibility while reducing computation in large vocabulary recognition problems. The last chapter in Part 4 addresses the concept of multi-pass search and the N-best search paradigm. It offers a way to incorporate knowledge sources that are not easily integrated in a finite state network representation. It also provides efficient search heuristics that are important for finding multiple hypotheses in a large search space under competing constraints.

The final part of the book presents two recognition system examples of "how to put everything together and make it work". They both discuss the important issue of usability extensively. In order to make recognition technology easily accessible and usable for practical applications, many system design issues need to be addressed. They include robustness, accuracy, vocabulary addition, rejection, acoustic adaptation, task adaptation, talker dependency, microphone dependency, channel dependency, environment dependency, computing resources, task definition, system integration, human factors, etc. By looking into these system issues and collecting field data for further study, new research directions will emerge and new techniques will be developed to solve some of the new challenging problems.

This book project was initiated while Prof. Paliwal was a visiting researcher at AT&T Bell Laboratories in Murray Hill in 1993. It was immediately clear that there are just too many research topics that are still fast progressing in the recognition field. Instead of being complete, it was decided among the editors to select only a small number of emerging advances and have them presented comprehensively. Distinguished colleagues were then invited to contribute to this volume. The invited contributors were requested to present the chosen topic in depth and in breadth. Due to some constraints, not all the invited colleagues responded to our request. The editors are therefore completely responsible for the selected topics, the order they are presented, the contributors they chose to present the particular topic, and the final outcome of the book.

Finally the editors would like to express their sincere gratitude to all the contributors to this volume. Without their continuing support and their quality work, it would not be possible to have such a quality volume. The editors are also indebted to their colleagues at AT&T Bell Laboratories. This book is a result of many inspiring discussions among our colleagues in Murray Hill. Special thanks is extended to Dr. B.-H. Juang who suggested the idea of editing such a book. Last but not the least, we would like to thank our families for putting up with us and continuously supporting this rewarding project.

CHL, FKS and KKP, December 1995.