Advances in Chinese Spoken Language Processing

Chin-Hui Lee • Haizhou Li
Lin-shan Lee • Ren-Hua Wang • Qiang Huo

Advances in Chinese Spoken Language Processing

This page is intentionally left blank

Advances in Chinese Spoken Language Processing

Chin-Hui Lee
Georgia Institute of Technology, USA

Haizhou Li
Institute for Infocomm Research, Singapore

Lin-shan Lee
National Taiwan University

Ren-Hua Wang
University of Science and Technology of China

Qiang Huo
The University of Hong Kong

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ADVANCES IN CHINESE SPOKEN LANGUAGE PROCESSING

Copyright © 2007 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-981-256-904-2

ISBN-10 981-256-904-9

Printed in Singapore by B & JO Enterprise

PREFACE

It is generally agreed that speech will play a major role in defining next-generation human-machine interfaces because it is the most natural means of communication among humans. To push forward this vision, speech research has enjoyed a long and glorious history spanning the entire twentieth century. As a result in the last three decades we have witnessed an intensive technology progress spurred on by recent advances in speech modeling, coordinated efforts between government funding agencies and speech communities for data collection and benchmark performance evaluation, and easy accesses to fast and affordable computing machineries. In the context of spoken language processing, we consider a collection of technical topics ranging over all aspects of speech communication, including production, perception, recognition, verification, synthesis, coding, analysis, and modeling. We have also seen quite a few spoken language system concepts moving out of research laboratories, and being deployed into real-life services and applications.

To benefit the entire world population, such natural voice user interfaces have to be developed for a large number of languages. For example successful spoken language translation systems are expected to facilitate global communication among people from different corners of the world to converse with each other by simple speaking and listening to their own native languages. In this sense, the particular group of languages that are not extensively studied will not be ready to get incorporated into the future global village of interactive communications. Realizing this desperate urgency, the speech and language researchers in the US and Europe have promoted and benefited tremendously from a series of large-scale research projects sponsored by government agencies and industry in the last three decades. Most of these efforts focused on developing research infrastructures for Indo-European languages, such as English, French, and German. On the other hand the family of spoken Chinese is mostly tonal and analytic in nature. It consists of a wide variety of languages and dialects that are currently being used by a large population spreading over a wide geographical area. Many languages in the Sino-Tibetan family have a strong historic tie with spoken Chinese as well. Studies of Chinese spoken language processing (CSLP) will therefore not only enhance our understanding in Chinese, but also trigger

advances in other similar languages, such as Japanese, Korean and Thai. To support CSLP, the individual government in the four major Chinese-speaking regions, China, Hong Kong, Singapore and Taiwan, has sponsored many related projects to establish language-specific research activities. However coordinated efforts among researchers from and shared research infrastructures across different regions have been rather limited until the later 90's.

Some spotty ideas have been raised by concerned researchers in the mid 90's to resolve the above situation. A real breakthrough came in the summer of 1997 when an ad hoc group was formed with the goal of creating an international forum for exchanging ideas and tools, sharing knowledge and experiences, promoting cross-region collaborations, and establishing a broad Chinese spoken language processing community. This core group included nine members, Profs. Taiyi Huang and Ren-Hua Wang from China, Profs. Chorkin Chan and Pak-Chung Ching from Hong Kong, Prof. Kim-Teng Lua and Dr. Haizhou Li from Singapore, Profs. Lin-shan Lee and Hsiao-Chuan Wang from Taiwan, and Dr. Chin-Hui Lee from USA. After a few months of intensive e-mail discussions, the group members finally met at the National University of Singapore in December, 1997. The result of this gathering was truly ground-breaking. A few special interest groups were formed to address some of the critical issues identified above. Some progress reports were prepared and key events were planned. Nonetheless the most significant outcome was the organization of a biennial meeting, called the International Symposium on Chinese Spoken Language Processing (ISCSLP), specifically devoted to CSLP-related issues. It was designed so that this forum will be hosted, in turns, by one organizing team from the four major Chinese-speaking regions. An ISCSLP Steering Committee, consisting of the above-mentioned nine members, was also established to oversee the related activities. With the enthusiastic support that the Committee received from Prof. Lua and Dr. Li, it was decided that the inaugural meeting, ISCSLP-98, be held in Singapore as a satellite event of the 1998 International Conference on Spoken Language Processing (ICSLP-98) in Sydney, Australia. Since then three follow-up meetings had taken place in Beijing, Taipei and Hong Kong in the years of 2000, 2002 and 2004, respectively. The number of accepted papers has also seen a steady increase from 55 in 1998 to over 100 in 2004, with symposium participants growing to over 200. The amount of support from the CSLP community has been tremendous, and this support also translates into a fast accumulation of knowledge and shared resources in the field of Chinese spoken language processing. To extend the group's impact on the general speech community a special interest group on CSLP (SIG-CSLP) was established within the International Speech Communication Association (ISCA) in 2002 at the

Taipei ISCSLP meeting. Very quickly, this SIG has become one of the most active groups in ISCA.

This year we are coming back to the Lion City after a full cycle of four successful ISCSLP gatherings. At this significant point in history to commemorate the tenth anniversary of the formation of the broad international CSLP community, Prof. Chin-Hui Lee thought it is timely fitting that the community collectively documents recent advances in CSLP. Although there are many general speech processing books available in the market, we have not seen a single volume dedicated solely to CSLP issues. Most of the materials are scattered in different literatures with some of them written in Chinese and not easily accessed by other scholars. With a great endorsement from Dr. Haizhou Li and generous grants from industrial sponsors in Singapore, Prof. Lee contacted one distinguished leader in the field from each Chinese-speaking region, and an international editorial team was then assembled in December 2005. An outline was drafted and a list of potential authoring teams was proposed. Instead of having another general speech book, the team decided to have a quality publication focusing on truly CSLP-related topics with illustrations of key concepts using mainly CSLP examples. Emphases were particularly devoted to highlighting differences between CSLP and general speech processing. Because of this requirement the team has to pass by many top colleagues in the CSLP community because they have not been actively involved in speech research using Chinese spoken language materials. After two months of back-and-forth discussions and revisions, a publication plan was finalized. In late February 2006 invitations were extended to distinguished colleagues who have demonstrated expertise in selected areas with specific guidelines in line with the intended scope of the book which is tutorial and overview in nature. To effectively utilize the limited page allowance, technical details were intentionally omitted and referred to already published references. Special challenges were also issued to encourage cross-region authoring cooperation when addressing common technical issues. The community's responses were overwhelming. We have collected 23 chapters covering a wide range of topics in three broad CSLP subject areas: principles, system integration, and applications. A roadmap to explore the book is outlined as follows.

Part I of the book concentrates on CSLP principles. There are 11 chapters: (1) Chapter 1 is a general production-perception perspective of speech analysis which is intended for all speech researchers and is equally applicable to any spoken language; (2) Chapter 2 presents some background materials on phonetic and phonological aspects of Chinese spoken languages; (3) Chapters 3-5 forms a group of topics addressing the prosodic and tonal aspects of Chinese spoken

languages. Chapter 3 provides an in-depth discussion on prosody analysis. The concept of prosodic phrase grouping which is a key property of Chinese spoken languages is illustrated. Another unique problem for tonal languages is tone modeling. Chapter 4 addresses issues related to tone modeling for speech synthesis. The important subject of Mandarin text-to-speech synthesis is then presented in Chapter 5; (4) Chapters 6-10 contain the group of subjects related to automatic speech recognition. Chapter 6 gives a review on large vocabulary continuous speech recognition of Mandarin Chinese highlighting technology components required to put together Mandarin recognition systems. Details are given in the next four chapters. Chapter 7 concerns with acoustic modeling of fundamental speech units. The syllabic nature of Mandarin, which is another unique property for syllabic languages, can be taken into account to effectively model continuous Mandarin speech. Tone modeling for Mandarin speech recognition, which is usually not considered in recognition of non-tonal languages, is discussed in Chapter 8. Because of the analytic nature of Chinese many single-character syllables are considered as words. Some special considerations are needed in language modeling for continuous Mandarin speech recognition. This is discussed in Chapter 9. Modeling of pronunciation variations in spontaneous Mandarin speech is key for improving performance of spoken language systems. This topic is presented in Chapter 10; and (5) Chapter 11 addresses the critical issue of corpus design and annotation which is becoming a key concern for designing spoken language systems. The tonal and syllabic nature of Mandarin makes corpus design a challenging research problem.

Part II of the book is devoted to technology integration and spoken language system design. There are seven chapters: (1) Chapter 12 is about speech-to-speech translation which is one of the grand challenges for the speech community. A domain-specific system is presented and some current capabilities and limitations are illustrated; (2) Chapter 13 is concerned with spoken document retrieval. Data mining and information retrieval are two technical topics that are impacting our daily lives. Spoken document retrieval encompasses these two areas. It is a good illustration of speech technology integration; (3) Chapter 14 presents an in-depth study on speech act modeling and its application in spoken dialogue systems; (4) Chapter 15 deals with a unique problem of transliteration that translates out-of-vocabulary words from one letter-based language like English into another character-based language like Chinese for many emerging speech and language applications; (5) Chapters 16 and 17 are devoted to two major languages in spoken Chinese, namely Cantonese and Min-nan. Issues in modeling for speech recognition and synthesis are highlighted. Tone modeling with these two languages is of special interest because there are more tones to

deal with and some of the tone differences are subtle and pose challenging technical problems to researchers; (6) An increasingly important subject that has attracted attention among speech researchers is the use of speech technologies to assist in language learning and evaluation. For example, the Putonghua Proficiency Test is currently conducted almost entirely in a manual mode. Chapter 18 deals with some issues about automating such processes.

Part III of the book concerns with applications, tools and CSLP resources. There are five chapters: (1) Chapter 19 discusses an audio-based digital content management and retrieval system for data mining of audiovisual documents; (2) Chapter 20 presents a dialog system with multilingual speech recognition and synthesis for Cantonese, Putonghua and English; (3) Chapter 21 presents a large-scale directory inquiry field trial system. Due to the potential of having an unlimited vocabulary, many technical and practical considerations need to be addressed. Pronunciation variation is also a major problem here. It is a good example of illustrating technical concerns in designing real-life Chinese spoken language systems; (4) Chapter 22 describes a car navigation application in which robustness is a main concern because of the adverse speaking conditions in moving vehicles. Interactions between speech and acoustics are of utmost importance in this hand-free and eye-free application; (5) Finally Chapter 23 provides a valuable collection of language resources for supporting Chinese spoken language processing.

In summary, putting together such an extensive volume in such a short time is a daunting task. The editors would like to express their sincere gratitude to all the distinguished contributors. Without their timely endeavor, it would not be possible to have such a quality book. A special thank is extended to two industry sponsors in Singapore for their generous financial support. A dedicated team at World Scientific Publishing has also assisted the editors continuously from the time of inception to final production of the book. Finally the editors are greatly indebted to Ms. Mahani Aljunied for her painstaking effort to make all chapters consistent in style, uniform in quality, and conforming to a single standard in presentation. Her education training in linguistics and her interest in spoken language processing made a big difference in finishing this historic volume in time.

Chin-Hui Lee, Atlanta
Haizhou Li, Singapore
Lin-shan Lee, Taipei
Ren-Hua Wang, Hefei
Qiang Huo, Hong Kong

This page is intentionally left blank

LIST OF CONTRIBUTORS

Shuanhu Bai
Institute for Infocomm Research,
Heng Mui Keng Terrace, Singapore

Berlin Chen
National Taiwan Normal University,
Taipei

Jung-Kuei Chen
Chunghwa Telecommunication
Laboratories, Taoyuan

Sin-Horng Chen
Department of Communication
Engineering,
National Chiao Tung University,
Hsinchu

Yuan-chin Chiang
Institute of Statistics,
National Tsing-hua University,
Hsinchu

Jen-Tzung Chien
Department of Computer Science and
Information Engineering,
National Cheng Kung University,
Tainan

Pak-Chung Ching
Department of Electronic Engineering,
The Chinese University of Hong Kong,
Hong Kong

Min Chu
Speech Group,
Microsoft Research Asia, Beijing

Chuang-Hua Chueh
Department of Computer Science and
Information Engineering,
National Cheng Kung University,
Tainan

Jianwu Dang
School of Information Science,
Japan Advanced Institute of Science
and Technology, Ishikawa

Li Deng
Microsoft Research,
One Microsoft Way, Redmond

Pascale Fung
Human Language Technology Center,
Department of Electronic & Computer
Engineering, Hong Kong University of
Science & Technology, Hong Kong

Yuqing Gao
IBM T. J. Watson Research Center,
Yorktown Heights

Liang Gu
IBM T. J. Watson Research Center,
Yorktown Heights

Taiyi Huang
Institute of Automation,
Chinese Academy of Sciences, Beijing

Qiang Huo
Department of Computer Science,
The University of Hong Kong,
Hong Kong

Mei-Yuh Hwang
Department of Electrical Engineering,
University of Washington, Seattle

Chih-Chung Kuo
Industrial Technology Research
Institute, Hsinchu

Jin-Shea Kuo
Chunghwa Telecommunication
Laboratories, Taoyuan

Chin-Hui Lee
School of Electrical & Computer
Engineering, Georgia Institute of
Technology, Atlanta

Lin-shan Lee
Department of Electrical Engineering,
National Taiwan University, Taipei

Tan Lee
Department of Electronic Engineering,
The Chinese University of Hong Kong,
Hong Kong

Aijun Li
Institute of Linguistics, Chinese
Academy of Social Sciences, Beijing

Haizhou Li
Institute for Infocomm Research,
Heng Mui Keng Terrace, Singapore

Min-siong Liang
Department of Electrical Engineering,
Chang Gung University, Taoyuan

Qingfeng Liu
USTC iFlytek Speech Laboratory,
University of Science and Technology
of China, Hefei

Yi Liu
Human Language Technology Center,
Department of Electronic & Computer
Engineering, Hong Kong University of
Science & Technology, Hong Kong

Wai Kit Lo
Department of Systems Engineering
and Engineering Management,
The Chinese University of Hong Kong,
Hong Kong

Dau-cheng Lyu
Department of Electrical Engineering,
Chang Gung University, Taoyuan

Ren-yuan Lyu
Department of Computer Science &
Information Engineering,
Chang Gung University, Taoyuan

Helen M. Meng
Department of Systems Engineering &
Engineering Management,
The Chinese University of Hong Kong,
Hong Kong

Yao Qian
Microsoft Research Asia,
Haidian, Beijing

Jianhua Tao
NLPR, Institute of Automation,
Chinese Academy of Sciences, Beijing

Chiu-yu Tseng
Institute of Linguistics,
Academia Sinica, Taipei

Hsiao-Chuan Wang
National Tsing Hua University,
Hsinchu

Hsien-Chang Wang
Department of Information Management, Chang Jung University, Tainan County

Hsin-min Wang
Institute of Information Science, Academia Sinica, Taipei

Jhing-Fa Wang
Department of Electrical Engineering, National Cheng Kung University, Tainan City

Jia-Ching Wang
Department of Electrical Engineering, National Cheng Kung University, Tainan City

Ren-Hua Wang
USTC iFlytek Speech Lab, University of Science & Technology of China, Hefei

Si Wei
USTC iFlytek Speech Laboratory, University of Science and Technology of China, Hefei

Chung-Hsien Wu
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan

Meng-Sung Wu
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan

Bo Xu
Institute of Automation, Chinese Academy of Sciences, Beijing

Gwo-Lang Yan
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan

Chung-Chieh Yang
Chunghwa Telecommunication Laboratories, Taoyuan

Jui-Feng Yeh
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan

Shuwu Zhang
Institute of Automation, Chinese Academy of Sciences, Beijing

Thomas Fang Zheng
Department of Physics, Tsinghua University, Beijing

Bowen Zhou
IBM T. J. Watson Research Center, Yorktown Heights

Yiqing Zu
Motorola China Research Center, Shanghai

This page is intentionally left blank

CONTENTS

Preface v

List of Contributors xi

Part I: Principles of CSLP **1**

Chapter 1 Speech Analysis: The Production-Perception Perspective 3
L. Deng and J. Dang

Chapter 2 Phonetic and Phonological Background of Chinese Spoken Languages 33
C.-C. Kuo

Chapter 3 Prosody Analysis 57
C. Tseng

Chapter 4 Tone Modeling for Speech Synthesis 77
S.-H. Chen, C. Tseng and H.-m. Wang

Chapter 5 Mandarin Text-To-Speech Synthesis 99
R.-H. Wang, S.-H. Chen, J. Tao and M. Chu

Chapter 6 Large Vocabulary Continuous Speech Recognition for Mandarin Chinese: Principles, Application Tasks and Prototype Examples 125
L.-s. Lee

Chapter 7 Acoustic Modeling for Mandarin Large Vocabulary Continuous Speech Recognition 153
M.-Y. Hwang

Chapter 8 Tone Modeling for Speech Recognition 179
T. Lee and Y. Qian

Chapter 9 Some Advances in Language Modeling 201
C.-H. Chueh, M.-S. Wu and J.-T. Chien

Chapter 10 Spontaneous Mandarin Speech Pronunciation Modeling 227
P. Fung and Y. Liu

Chapter 11 Corpus Design and Annotation for Speech Synthesis and Recognition 243
A. Li and Y. Zu

Part II: CSLP Technology Integration **269**

Chapter 12 Speech-to-Speech Translation 271
Y. Gao, L. Gu and B. Zhou

Chapter 13 Spoken Document Retrieval and Summarization 301
B. Chen, H.-m. Wang and L.-s. Lee

Chapter 14 Speech Act Modeling and Verification in Spoken Dialogue Systems 321
C.-H. Wu, J.-F. Yeh and G.-L. Yan

Chapter 15 Transliteration 341
H. Li, S. Bai and J.-S. Kuo

Chapter 16 Cantonese Speech Recognition and Synthesis 365
P. C. Ching, T. Lee, W. K. Lo and H. M. Meng

Chapter 17 Taiwanese Min-nan Speech Recognition and Synthesis 387
R.-y. Lyu, M.-s. Liang, D.-c. Lyu and Y.-c. Chiang

Chapter 18 Putonghua Proficiency Test and Evaluation 407
R.-H. Wang, Q. Liu and S. Wei

Part III: Systems, Applications and Resources **431**

Chapter 19 Audio-Based Digital Content Management and Retrieval 433
B. Xu, S. Zhang and T. Huang

Chapter 20 Multilingual Dialog Systems 459
H. M. Meng

Chapter 21 Directory Assistance System 483
J.-K. Chen, C.-C. Yang and J.-S. Kuo

Chapter 22 Robust Car Navigation System 503
J.-F. Wang, H.-C. Wang and J.-C. Wang

Chapter 23 CSLP Corpora and Language Resources 523
H.-C. Wang, T. F. Zheng and J. Tao

Index 539

Advances in Chinese Spoken Language Processing **Part I**

Principles of CSLP

This page is intentionally left blank

CHAPTER 1

SPEECH ANALYSIS: THE PRODUCTION-PERCEPTION PERSPECTIVE

Li Deng[†] and Jianwu Dang[‡]

[†]*Microsoft Research*
One Microsoft Way, Redmond, WA 98052
[‡]*School of Information Science*
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292
Email: {deng@microsoft.com, jdang@jaist.ac.jp}

This chapter introduces the basic concepts and techniques of speech analysis from the perspectives of the underlying mechanisms of human speech production and perception. Spoken Chinese language has special characteristics in its signal properties that can be well understood in terms of both the production and perception mechanisms. In this chapter, we will first outline the general linguistic, phonetic, and signal properties of spoken Chinese. We then introduce human production and perception mechanisms, and in particular, those relevant to spoken Chinese. We also present some recent brain research on the relationship between human speech production and perception. From the perspectives of human speech production and perception, we then describe popular speech analysis techniques and classify them based on the underlying scientific principles either from the speech production or perception mechanism or from both.

1. Introduction

Chinese is the language of over one billion speakers. Several dialect families of Chinese exist, each in turn consisting of many dialects. Although different dialect families are often mutually unintelligible, systematic correspondences (e.g., in lexicon and syntax) exist among them, making it easy for speakers of one dialect to pick up another relatively quickly. The largest dialect family is the Northern family, which consists of over 70% of all Chinese speakers. Standard or Mandarin Chinese is a member of the Northern family and is based on the pronunciation of the Beijing dialect. Interestingly, most speakers of Standard