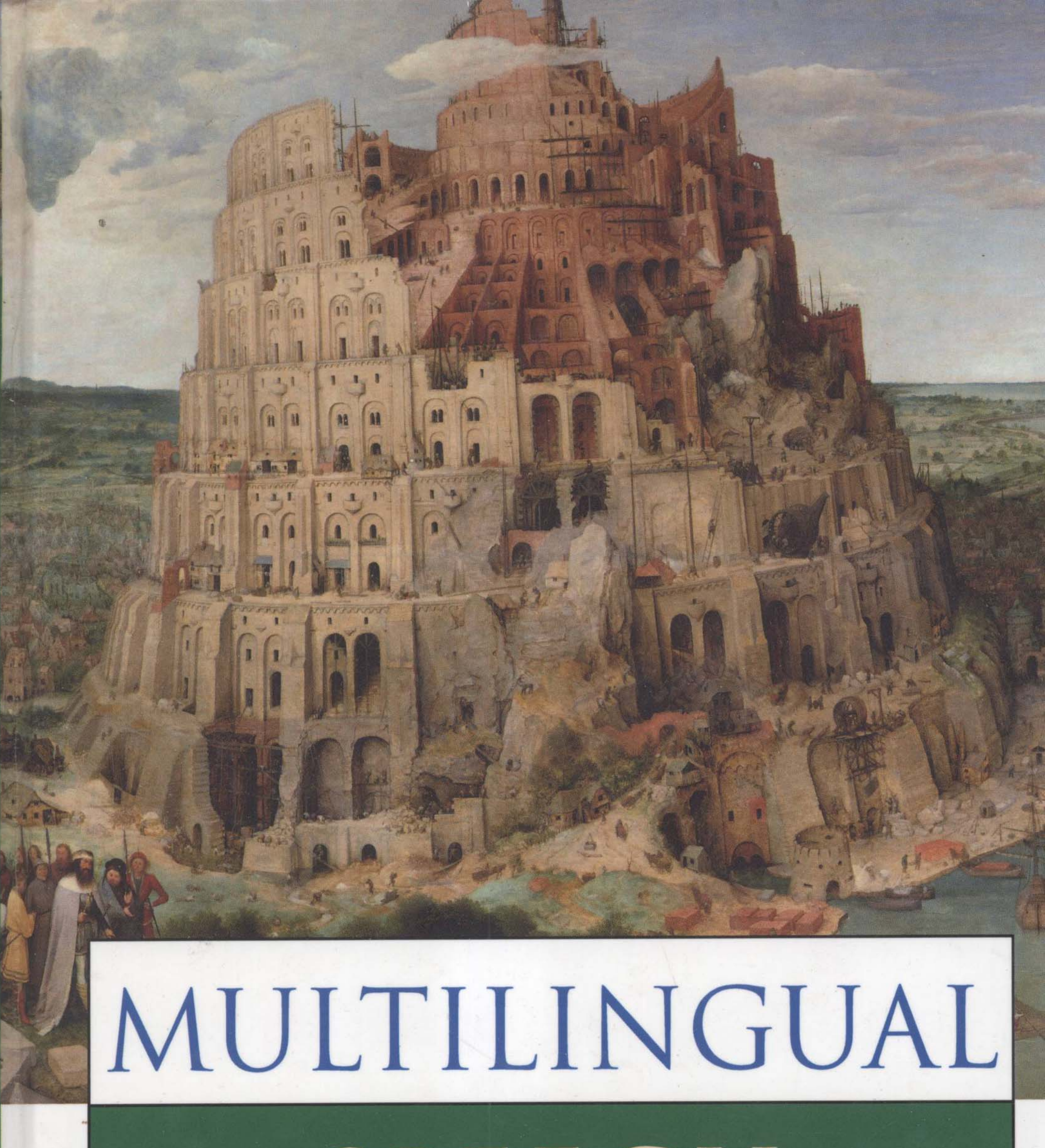

MULTILINGUAL

SPEECH PROCESSING

TANJA SCHULTZ

KATRIN KIRCHHOFF

Multilingual Speech Processing

Schultz & Kirchhoff

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. ∞

Library of Congress Cataloging-in-Publication Data

Application submitted.

British Library Cataloguing-in-Publication Data

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

ISBN 13: 978-0-12-088501-5
ISBN 10: 0-12-088501-8

For information on all Elsevier Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America
06 07 08 09 10 11 9 8 7 6 5 4 3 2 1

Multilingual Speech Processing

Contributor Biographies

Dr. Tanja Schultz received her Ph.D. and Masters in Computer Science from University Karlsruhe, Germany in 2000 and 1995, respectively, and earned a German Masters in Mathematics, Sports, and Education Science from the University of Heidelberg, Germany in 1990. She joined Carnegie Mellon University in 2000 and is a faculty member of the Language Technologies Institute as an Assistant Research Professor. Her research activities center around human-machine and human-human interaction. With a particular area of expertise in multilingual approaches, she directs research on portability of speech and language processing systems to many different languages. In 2001 Tanja Schultz was awarded with the FZI price for her outstanding Ph.D. thesis on language independent and language adaptive speech recognition. In 2002 she received the Allen Newell Medal for Research Excellence from Carnegie Mellon for her contribution to Speech-to-Speech Translation and the ISCA best paper award for her publication on language independent acoustic modeling. She is an author of more than 80 articles published in books, journals, and proceedings, and a member of the IEEE Computer Society, the European Language Resource Association, and the Society of Computer Science (GI) in Germany. She served as Associate Editor for IEEE Transactions and is currently on the Editorial Board of the Speech Communication journal.

Dr. Katrin Kirchhoff studied Linguistics and Computer Science at the Universities of Bielefeld, Germany, and Edinburgh, United Kingdom, and was a visiting researcher at the International Computer Science Institute, Berkeley, California. After obtaining her Ph.D. in Computer Science from the University of Bielefeld in 1999, she joined the University of

Washington, where she is currently a Research Assistant Professor in Electrical Engineering. Her research interests are in automatic speech recognition, language identification, statistical natural language processing, human-computer interfaces, and machine translation. Her work emphasizes novel approaches to acoustic-phonetic and language modeling and their application to multilingual contexts. She currently serves on the Editorial Board of the Speech Communication journal.

Dr. Christopher Cieri is the Executive Director of the Linguistic Data Consortium, where he has overseen dozens of data collection and annotation projects that have generated multilingual speech and text corpora. His Ph.D. is in Linguistics from the University of Pennsylvania. His research interests revolve around corpus based language description especially in phonetics, phonology, and morphology as they interact with nonlinguistic phenomena as in language contact and studies of linguistic variation.

Dr. Mark Liberman is Trustee Professor of Phonetics in Linguistics at the University of Pennsylvania, where he is also Director of the Linguistic Data Consortium, Co-Director of the Institute for Research in Cognitive Science, and Faculty Master of Ware College House. His Ph.D. is from the Massachusetts Institute of Technology in 1975, and he worked from 1975 to 1990 at AT&T Bell Laboratories, where he was head of the Linguistics Research Department.

Dr. Khalid Choukri obtained an Electrical Engineering degree (1983) from Ecole Nationale de l'aviation civile (ENAC), and Masters Degree (1984) and doctoral degrees (1987) in Computer Sciences and Signal Processing at the Ecole Nationale Supérieure des Télécommunications (ENST) in Paris. He was a research scientist at the Signal Department of ENST, involved in Man-Machine Interaction. He has also consulted for several French companies, such as Thomson, on various speech system projects and was involved in SAM, ARS, etc. In 1989, he joined CAP GEMINI INNOVATION, R&D center of CAP SOGETI to work as the team leader on speech processing, oral dialogs and neural networks. He managed several ESPRIT projects, such as SPRINT, and was involved in many others, such as SUNDIAL. He then moved to ACSYS in September 1992 to take on the position of Speech Technologies Manager. Since 1995, he has been the Executive Director of the European Language

Resources Association (ELRA) and the Managing Director of the Evaluations and Language Resources Distribution Agency (ELDA) for which the priority activities include the collection and distribution of Language Resources. In terms of experience with EC-funded projects, ELDA/ELRA has played a significant role in several European projects, such as C-Oral-Rom, ENABLER, NET-DC, OrienTel, CLEF, CHIL, and TC-STAR.

Dr. Victoria Arranz holds an M.Sc. in Machine Translation and a Ph.D. in Computational Linguistics (1998) from the Centre for Computational Linguistics, University of Manchester Institute of Science and Technology (UMIST), United Kingdom, where she participated in several international projects dealing with restricted domains and corpus study and processing. She has worked as a Research Scientist both at the Grup d'Investigació en Lingüística Computacional (gilcUB) and at the Centre de Llenguatge i Computació (CLIC), Universitat de Barcelona, Spain, working on the production of language resources and coordinating the development of terminological LRs for the medical domain. Then she joined the Universitat Politècnica de Catalunya, Barcelona, where she has been a Visiting Researcher, a Senior Researcher, and also a Lecturer of Computational Linguistics within the Natural Language Processing and Cognitive Science Ph.D. program. She has also participated in a number of national and international projects regarding Terminological LRs (SCRIPTUM), LRs for Speech-to-Speech Translation (LC-STAR), Dialogue Systems (BASURDE), Speech-to-Speech Translation (ALIADO, FAME, TC-STAR), and other NLP techniques. Currently, she is the Head of the Language Resources Identification Unit at ELDA, Paris, France, in charge of the BLARK and UNIVERSAL CATALOGUE projects, whose aims relate to the compiling of the existing LRs and the production of LRs in terms of language and technology needs.

Dr. Lori Lamel joined LIMSI as a permanent CNRS Researcher in October 1991 (http://www.limsi.fr/Individu/lamel/). She received her Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in May 1988. She obtained her 'Habilitation a diriger des Recherches' [Document title: Traitment de la parole (Spoken Language Processing)] in January 2004. Her research activities include multilingual studies in large vocabulary continuous speech recognition; acoustic-phonetics, lexical, and phonological modeling; spoken dialog systems; speaker and language identification; and the design, analysis, and realization of large speech corpora. She has been a prime contributor to the LIMSI participations in DARPA benchmark evaluations, being involved in acoustic model development and responsible for the pronunciation lexicon and has been involved in many European projects on speech recognition and spoken language dialog systems. She has over 150 publications and is a member of the Editorial Board of the Speech Communication journal, the Permanent Council of ICSLP and the coordination board of the L'Association Francophone de la Communication Parle.

Dr. Martine Adda-Decker has been a permanent CNRS Researcher at LIMSI since 1990 (http://www.limsi.fr/Individu/madda). She received an M.Sc. degree in Mathematics and Fundamental Applications in 1983 and her doctorate in Computer Science in 1988 from the University of Paris XI. Her main research interests are in multilingual, large vocabulary continuous speech recognition, acoustic and lexical modeling, and language identification. She has been a principal developer of the German ASR system. She is also interested in spontaneous speech phenomena, pronunciation variants, and ASR errors related to spontaneous speaking styles. More recently she has focused her research on using automatic speech recognizers as a tool to study phonetics and phonology in a multilingual context. In particular ASR systems can contribute to describe less studied languages, dialects, and regional varieties on the acoustic, phonetic, phonological, and lexical levels. She has been involved in many national CNRS and European projects.

Dr. Sanjeev P. Khudanpur is with the Department of Electrical & Computer Engineering and the Center for Language and Speech Processing at the Johns Hopkins University. He obtained a B. Tech. from the Indian Institute of Technology, Bombay, in 1988, and a Ph.D. from the University of Maryland, College Park, in 1997, both in Electrical Engineering. His research is concerned with the application of information theoretic and statistical methods to problems in human language technology, including automatic speech recognition, machine translation and information retrieval. He is particularly interested in maximum entropy and related techniques for model estimation from sparse data.

Dr. Alan W. Black is an Associate Research Professor on the faculty of the Language Technologies Institute at Carnegie Mellon University. He is a principal author of the Festival Speech Synthesis System, a standard free software system used by many research and commercial institutions throughout the world. Since joining CMU in 1999, with Kevin Lenzo, he has furthered the ease and robustness of building synthetic voices through the FestVox project using new techniques in unit selection, text analysis, and prosodic modeling. He graduated with a Ph.D. from the University of Edinburgh in 1993, and then worked in industry in Japan at ATR. He returned to academia as a Research Fellow at CSTR in Edinburgh and moved to CMU in 1999. In 2000, with Kevin Lenzo, he started the for-profit company Cepstral, LLC in which he continues to serve as Chief Scientist. He has a wide background in computational linguistics and has published in computational morphology, language modeling for speech recognition, computational semantics, and most recently in speech synthesis, dialog systems, prosody modeling and speech-to-speech translation. He is a strong proponent of building practical implementations of computational theories of speech and language.

Dr. Jiří Navrátil received M.Sc. and Ph.D. (summa cum laude) degrees from the Ilmenau Technical University, Germany in 1994 and 1999, respectively. From 1996 and 1998 he was Assistant Professor at the Institute of Communication and Measurement Technology at the ITU performing research on speech recognition and language identification. For his work in the field of language identification, Dr. Navrátil received the 1999 Johann-Philipp-Reis Prize awarded by the VDE (ITG), Deutsche Telekom, and the cities of Friedrichsdorf and Gelnhausen, Germany. In 1999, he joined IBM to work in the Human Language Technologies Department at the Thomas J. Watson Research Center, Yorktown Heights, New York. He has authored over 40 publications on language and speaker recognition, received several invention achievement awards and has a technical group award from IBM. His current interests include voice-based authentication, particularly conversational biometrics, language recognition, and user-interface technologies.

Dr. Etienne Barnard is a research scientist and coleader of the Human Language Technologies research group at the Meraka Institute in Pretoria, South Africa, and Professor in Electronic and Computer Engineering at the University of Pretoria. He obtained a Ph.D. in Electronic and Computer Engineering from Carnegie Mellon University in 1989, and is active in the development of statistical approaches to speech recognition and intonation modeling for the indigenous South African languages.

Dr. Marelie Davel is a research scientist and coleader of the Human Language Technologies research group at the Meraka Institute in Pretoria, South Africa. She obtained a Ph.D. in Computer Engineering at the University of Pretoria in 2005. Her research interests include pronunciation modeling, bootstrapping of resources for language technologies, and new approaches to instance-based learning.

Dr. Silke Goronzy received a diploma in Electrical Engineering from the Technical University of Braunschweig, Germany, in 1994. She joined the Man-Machine Interface group of the Sony Research Lab in Stuttgart, Germany, to work in the area of automatic speech recognition on speaker adaptation, confidence measures, and adaptation to non-native speakers. In 2002 she received a Ph.D. in Electrical Engineering from the University of Braunschweig. At Sony she also performed research in the area of multimodal dialog systems, and in 2002, she lead a team working on personalization and automatic emotion recognition. In 2004, she joined 3SOFT GmbH in Erlangen, Germany, where she is leading the Speech Dialog Systems team that is developing HMI solutions for embedded applications. She also gives lectures at the University of Ulm, Germany.

Dr. Laura Mayfield Tomokiyo holds a Ph.D. in Language Technologies and an M.S. in Computational Linguistics from Carnegie Mellon University. Her undergraduate degree is from the Massachusetts Institute of Technology in Computer Science and Electrical Engineering. She has held positions at Toshiba and the Electrotechnical Laboratories (ETL) in Japan. Currently, she is Director of Language Development at Cepstral, LLC, where she is responsible for expansion to new languages and enhancement of existing languages in text-to-speech synthesis. Her research interests include multilinguality in speech technology, application of speech technology to language learning, and the documentation and preservation of underrepresented languages.

Dr. Seichii Yamamoto graduated from Osaka University in 1972 and received his Masters and Ph.D. degrees from Osaka University in 1974 and 1983, respectively. He joined Kokusai Denshin Denwa Co. Ltd. in

April 1974, and ATR Interpreting Telecommunications Research Laboratories in May 1997. He was appointed president of ATR-ITL in 1997. He is currently a Professor of Doshisha University and invited researcher (ATR Fellow) at ATR Spoken Language Communication Research Laboratories. His research interests include digital signal processing, speech recognition, speech synthesis, natural language processing, and spoken language translation. He has received Technology Development Awards from the Acoustical Society of Japan in 1995 and 1997. He is also an IEEE Fellow and a Fellow of IEICE Japan.

Dr. Alex Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the University of Karlsruhe, Germany. He directs the Interactive Systems Laboratories at both Universities with research emphasizing speech recognition, handwriting recognition, language processing, speech translation, machine learning, and multimodal and multimedia interfaces. At Carnegie Mellon University, he also serves as Associate Director of the Language Technology Institute and as Director of the Language Technology Ph.D. program. He was one of the founding members of the CMU's Human Computer Interaction Institute (HCII) and continues on its core faculty. Dr. Waibel was one of the founders of C-STAR, the international consortium for speech translation research and served as its chairman from 1998–2000. His team has developed the JANUS speech translation system, the JANUS speech recognition toolkit, and a number of multimodal systems including the meeting room, the Genoa Meeting recognizer and meeting browser. Dr. Waibel received a B.S. in Electrical Engineering from the Massachusetts Institute of Technology in 1979, and his M.S. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 1980 and 1986. His work on the Time Delay Neural Networks was awarded the IEEE best paper award in 1990, and his work on speech translation systems the "Alcatel SEL Research Prize for Technical Communication" in 1994.

Dr. Stephan Vogel studied physics at Philips University in Marburg, Germany, and at Imperial College in London, England. He then studied History and Philosophy of Science at the University of Cambridge, England. After returning to Germany he worked from 1992 to 1994 as a Research Assistant in the Department of Linguistic Data Processing, University of Cologne, Germany. From 1994 to 1995 he worked as software developer at ICON Systems, developing educational software. In 1995 he joined the research team of Professor Ney at the Technical University of Aachen, Germany, where he started work on statistical machine translation. Since May 2001, he has worked at Carnegie Mellon University in Pittsburgh, Pennsylvania, where he leads a team of students working on statistical machine translation, translation of spontaneous speech, automatic lexicon generation, named entity detection and translation, and machine translation evaluation.

Dr. Helen Meng is a Professor in the Department of Systems Engineering & Engineering Management of The Chinese University of Hong Kong (CUHK). She received her S.B., S.M., and Ph.D. degrees in Electrical Engineering, all from the Massachusetts Institute of Technology. Upon joining CUHK in 1998, Helen established the Human-Computer Communications Laboratory, for which she serves as Director. She also facilitated the establishment of the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies in 2005 and currently serves as co-director. Her research interests include multilingual speech and language processing, multimodal human-computer interactions with spoken dialogs, multibiometric user authentication as well as multimedia data mining. Dr. Meng is a member of the IEEE Speech Technical Committee and the Editorial Boards of several journals, including *Computer Speech and Language, Speech Communication*, and the *International Journal of Computational Linguistics*, and *Chinese Language Processing*. In addition to speech and language research, Dr. Meng is also interested in the development of Information and Communication Technologies for our society. She is an appointed member of the Digital 21 Strategy Advisory Committee, which is the main advisory body to the Hong Kong SAR Government on information technology matters.

Dr. Devon Li is a chief engineer in the Human-Computer Communications Laboratory (HCCL), The Chinese University of Hong Kong (CUHK). He received his B.Eng. and M.Phil. degrees from the Department of Systems Engineering and Engineering Management, CUHK. His Masters thesis research focused on monolingual and English-Chinese cross-lingual spoken document retrieval systems. This work was selected to represent CUHK in the Challenge Cup Competition in 2001, a biennial competition where over two hundred universities across China compete in terms of their R&D projects. The project was awarded Second Level Prize in this competition. Devon extended his work to spoken query retrieval during his internship at Microsoft Research Asia, Beijing. In 2002, Devon began to work on the "Author Once, Present Anywhere (AOPA)" project in HCCL, which aimed to develop a software platform that enables multi-device access to Web content. The user interface is adaptable to a diversity of form factors including the desktop computer, mobile handhelds, and screenless voice browsers. Devon has developed the CU Voice Browser (a bilingual voice browser) and also worked on the migration of CUHK's Cantonese speech recognition (CU RSBB) and speech synthesis (CU VOCAL) engines toward SAPI-compliance. Devon is also among the first developers to be proficient with the emerging SALT (Speech Application Language Tags) standard for multimodal Web interface development. He has integrated the SAPI-compliant CUHK speech engines with the SALT framework.

Foreword

Speech recognition and speech synthesis technologies have enjoyed a period of rapid progress in recent years, with an increasing number of functional systems being developed for a diverse array of applications. At the same time, this technology is only being developed for fewer than twenty of the world's estimated four to eight thousand languages. This disparity suggests that in the near future, demand for speech recognition and speech synthesis technologies, and the automated dialog, dictation, and summarization systems that they support, will come to include an increasingly large number of new languages. Due to current globalization trends, multilingual speech recognition and speech synthesis technologies, which will enable things like speech-to-speech translation, and the ability to incorporate the voice of a speaker from one language into a synthesizer for a different language (known as polyglot synthesis), will become increasingly important.

Because current speech recognition and speech synthesis technologies rely heavily on statistical methods, when faced with the challenge of developing systems for new languages, it is often necessary to begin by constructing new speech corpora (databases). This in turn requires many hours of recording and large amounts of funding, and consequently, one of the most important problems facing minority language researchers will be how to significantly advance research on languages for which there are either limited data resources or scant funds. Moreover, because all languages have unique characteristics, technologies developed based on models of one language cannot simply be ported "as-is" to other languages; this process requires substantial modifications. These are a few of the major handicaps facing current efforts to develop speech technology and extend it into new areas and to new languages. If speech recognition and synthesis systems could be easily and effectively ported between different languages, a much greater number of people might share the benefits that this technology has to offer.

This book spans the state-of-the-art technologies of multilingual speech processing. Specifically, it focuses on current research efforts in Europe and America; it describes new speech corpora under development and discusses multilingual speech recognition techniques from acoustic and language modeling to multilingual dictionary construction issues. Regarding the issue of language modeling, it even touches on new morphological and lexical segmentation techniques. On the topic of multilingual text-to-speech conversion, it discusses possible methods for generating voices in new languages. Language identification techniques and approaches to recognition problems involving non-native speakers are also discussed. The last portion of the book is devoted to issues in speech-to-speech translation and automated multilingual dialog systems.

Presently enrolled at my laboratory here at Tokyo Tech are not only Japanese students but also students from eleven other countries, including the United States, England, France, Spain, Iceland, Finland, Poland, Switzerland, Thailand, Indonesia, and Brazil. Despite this large number of representative languages, the ones for which we are easily able to obtain comparatively large- scale spoken corpora are limited to widely researched languages like English, Japanese, and French. Developing new speech recognition and synthesis systems for any other languages is considerably expensive and time consuming, mainly because researchers must begin by first constructing new speech corpora. In our lab, for the purpose of developing an automated dialog system for Icelandic–a language spoken by approximately 300,000 people worldwide–we are currently conducting research in an effort to automatically translate a series of written English corpora into Icelandic. This is possible both because the English data is abundantly available and because the two languages' similarity in terms of grammatical structure makes the idea more feasible than with many other language pairings.

Constructing corpora is not, however, solely a problem for minority languages. In order to expand the application of current speech recognition systems, instead of just recognizing carefully read text from written manuscripts, the ability to recognize spontaneous speech with a high degree of precision will become essential. Currently available speech recognition technology is capable of achieving high levels of accuracy with speech read aloud carefully from a text; however, when presented with spontaneous speech, recognition rates drop off dramatically. This is because read speech and spontaneous speech are dramatically different, both from a linguistic standpoint and in terms of their respective acoustic models. This difference is actually greater than the difference between two closely related languages. In order to improve recognition accuracy for spontaneous speech, large scale spontaneous speech corpora are necessary, such as are currently available only for a limited number of widely spoken and recorded languages, like English or Japanese. But even the substantial size of spoken corpora in these languages is several orders of magnitude below what they have to offer in terms of written text. The reason for this is, again, the colossal amount of money required to construct a new corpus, which requires many hours of recording, followed by faithful transcription and detailed annotation. From this point on the most pressing difficulties in both spontaneous speech recognition and multilingual speech technology will go hand in hand.

Seen from these various perspectives, the publication of this new book appears not only very timely, but also promises to occupy a unique place among current speech-related textbooks and reference manuals. In addition to the respect and gratitude I have for the two young researchers who have made its publication a reality, in the years to come I expect this book to become an important milestone in the course of further speech-related research.

Sadaoki Furui
Tokyo Institute of Technology
August 26, 2005

Contents