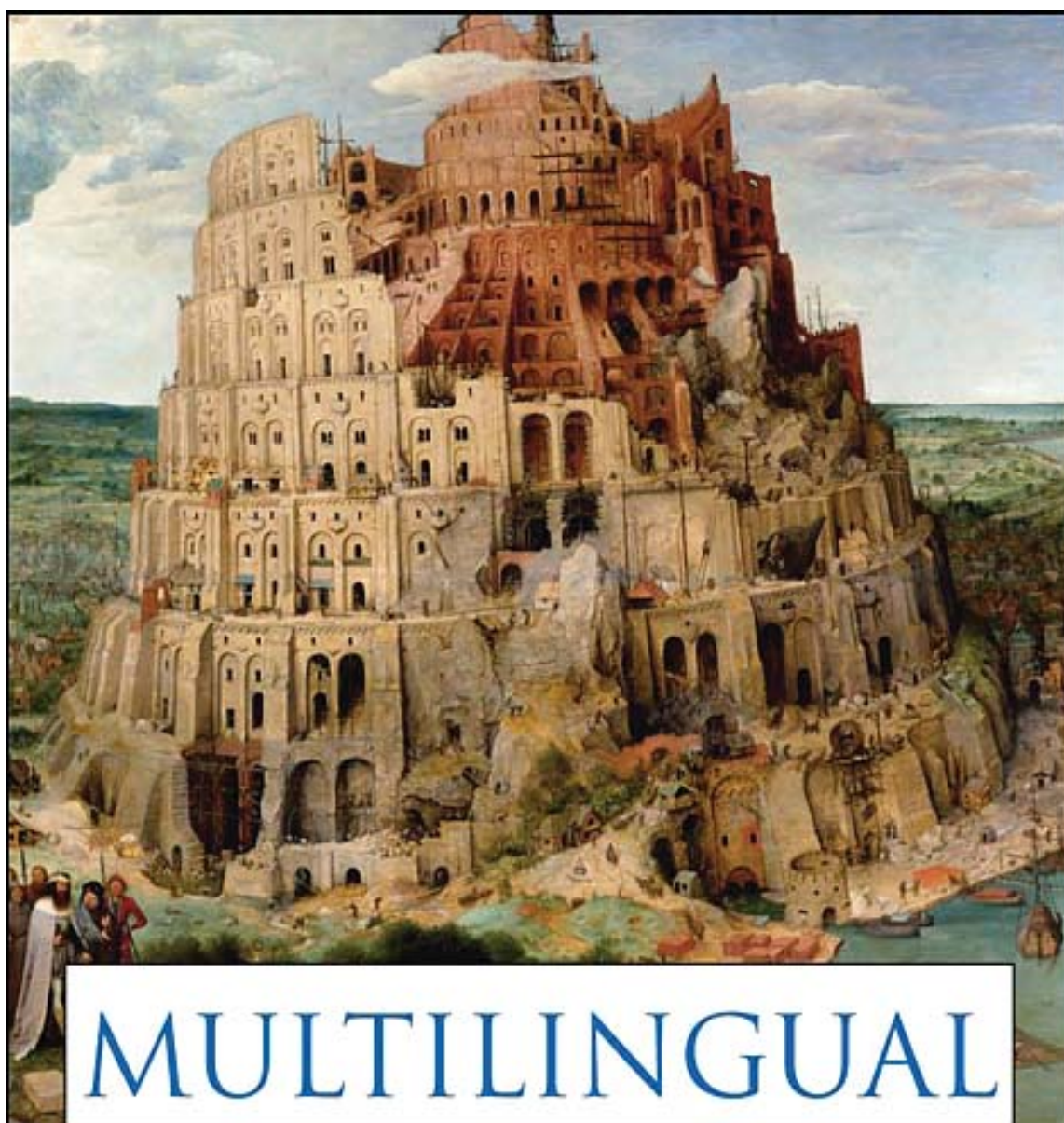




MULTILINGUAL — SPEECH — PROCESSING



TANJA SCHULTZ
KATRIN KIRCHHOFF

Multilingual Speech Processing

This Page Intentionally Left Blank

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. ∞

Copyright © 2006, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: permissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

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

Working together to grow libraries in developing countries		
www.elsevier.com www.bookaid.org www.sabre.org		
ELSEVIER	BOOK AID International	Sabre Foundation

Contents

Contributor Biographies	xvii
Foreword	xxvii
1 Introduction	1
2 Language Characteristics	5
<i>Katrin Kirchhoff</i>	
2.1 Languages and Dialects	5
2.2 Linguistic Description and Classification	8
2.3 Language in Context	20
2.4 Writing Systems	22
2.5 Languages and Speech Technology	30
3 Linguistic Data Resources	33
<i>Christopher Cieri and Mark Liberman, Victoria Arranz and Khalid Choukri</i>	
3.1 Demands and Challenges of Multilingual Data-Collection Efforts	33
3.2 International Efforts and Cooperation	40
3.3 Data Collection Efforts in the United States	44
3.4 Data Collection Efforts in Europe	55
3.5 Overview Existing Language Resources in Europe	64
4 Multilingual Acoustic Modeling	71
<i>Tanja Schultz</i>	
4.1 Introduction	71
4.2 Problems and Challenges	79
4.3 Language Independent Sound Inventories and Representations	91
4.4 Acoustic Model Combination	102
4.5 Insights and Open Problems	118

5 Multilingual Dictionaries	123
<i>Martine Adda-Decker and Lori Lamel</i>	
5.1 Introduction	123
5.2 Multilingual Dictionaries	125
5.3 What Is a Word?	129
5.4 Vocabulary Selection	141
5.5 How to Generate Pronunciations	149
5.6 Discussion	166
6 Multilingual Language Modeling	169
<i>Sanjeev P. Khudanpur</i>	
6.1 Statistical Language Modeling	169
6.2 Model Estimation for New Domains and Speaking Styles	174
6.3 Crosslingual Comparisons: A Language Modeling Perspective	177
6.4 Crosslinguistic Bootstrapping for Language Modeling	193
6.5 Language Models for Truly Multilingual Speech Recognition	199
6.6 Discussion and Concluding Remarks	202
7 Multilingual Speech Synthesis	207
<i>Alan W. Black</i>	
7.1 Background	208
7.2 Building Voices in New Languages	208
7.3 Database Design	213
7.4 Prosodic Modeling	216
7.5 Lexicon Building	219
7.6 Non-native Spoken Output	230
7.7 Summary	231
8 Automatic Language Identification	233
<i>Jiří Navrátil</i>	
8.1 Introduction	234
8.2 Human Language Identification	235
8.3 Databases and Evaluation Methods	240
8.4 The Probabilistic LID Framework	242
8.5 Acoustic Approaches	245
8.6 Phonotactic Modeling	251
8.7 Prosodic LID	262
8.8 LVCSR-Based LID	266
8.9 Trends and Open Problems in LID	268

9 Other Challenges: Non-native Speech, Dialects, Accents, and Local Interfaces	273
<i>Silke Goronzy, Laura Mayfield Tomokiyo, Etienne Barnard, Marelie Davel</i>	
9.1 Introduction	273
9.2 Characteristics of Non-native Speech	276
9.3 Corpus Analysis	278
9.4 Acoustic Modeling Approaches for Non-native Speech	287
9.5 Adapting to Non-native Accents in ASR	288
9.6 Combining Speaker and Pronunciation Adaptation.....	298
9.7 Cross-Dialect Recognition of Native Dialects	299
9.8 Applications	301
9.9 Other Factors in Localizing Speech-Based Interfaces	309
9.10 Summary.....	315
10 Speech-to-Speech Translation	317
<i>Stephan Vogel, Tanja Schultz, Alex Waibel, and Seichii Yamamoto</i>	
10.1 Introduction	317
10.2 Statistical and Interlingua-Based Speech Translation Approaches	320
10.3 Coupling Speech Recognition and Translation.....	341
10.4 Portable Speech-to-Speech Translation: The ATR System	347
10.5 Conclusion	394
11 Multilingual Spoken Dialog Systems	399
<i>Helen Meng and Devon Li</i>	
11.1 Introduction	399
11.2 Previous Work.....	403
11.3 Overview of the ISIS System.....	407
11.4 Adaptivity to Knowledge Scope Expansion	417
11.5 Delegation to Software Agents	425
11.6 Interruptions and Multithreaded Dialogs	427
11.7 Empirical Observations on User Interaction with ISIS	433
11.8 Implementation of Multilingual SDS in VXML	437
11.9 Summary and Conclusions	443
Bibliography	449
Index	491

This Page Intentionally Left Blank

List of Figures

Figure 2.1	The Indo-European language family (extinct languages are not shown).	10
Figure 2.2	The International Phonetic Alphabet (from the International Phonetic Association).	13
Figure 2.3	Example of within-utterance code-switching between French and Moroccan Arabic, from Naït M'Barek and Sankoff (1988).	21
Figure 2.4	Classification of writing systems.	23
Figure 2.5	Representation of the sentence <i>I come from Hunan</i> in Chinese Hanzi, with indication of word segmentation.	24
Figure 2.6	Arabic script representation of the sentence <i>I like to travel to Cairo</i> , with diacritics (top row) and without (bottom row).	25
Figure 2.7	Japanese Hiragana characters.	26
Figure 2.8	Basic Korean Hangul characters.	27
Figure 3.1	Contract model followed by ELRA.	63
Figure 4.1	Automatic speech recognition (ASR).	72
Figure 4.2	Generation of an observation sequence $O = o_1 o_2 \dots o_T$ with a three-state left-to-right hidden Markov model.	73
Figure 4.3	Context decision tree for the middle state of a quinphone HMM for phone /k/.	76
Figure 4.4	Multilingual ASR system.	77
Figure 4.5	Consonants (C) to Vowels (V) ratio (in %) and phone-based error rates for nine languages.	86
Figure 4.6	Number of polyphones over context width for nine languages.	87
Figure 4.7	Compactness for nine EU-languages: Word Type (vocabulary) vs. Word Tokens (top) and vs. Grapheme Tokens (bottom).	90
Figure 4.8	Average and range of the share factor for phoneme based and articulatory feature based units over the number of $\binom{12}{k}$ and $\binom{5}{l}$ involved languages, respectively, with $k = 1, \dots, 12$ and $l = 1, \dots, 5$.	98
Figure 4.9	Portuguese polyphone coverage by nine languages.	101

Figure 4.10	Acoustic model combination ML-sep, ML-mix, ML-tag (from left to right).	107
Figure 4.11	Entropy gain over number of subpolyphones for a five-lingual system.	110
Figure 4.12	Classification accuracy of articulatory feature detectors from five languages on English test data.	114
Figure 5.1	Language-dependent resources for transcription systems.	125
Figure 5.2	About 13% of French's entries are imported from other languages, mainly English, Italian, and Germanic (after Walter, 1997).	126
Figure 5.3	Language independent processing steps for pronunciation dictionary generation.	128
Figure 5.4	Sample word lists obtained using different text normalizations, with standard base-form pronunciations.	130
Figure 5.5	Number of distinct (left) and total number (right) of words in the training data for different normalization combination V_i .	134
Figure 5.6	OOV rates for different normalization versions V_i on the training data using 64,000 word lists.	135
Figure 5.7	Number of words as a function of length (in characters) for German, English, and French from 300 million words running texts in each language. Number of distinct entries in the full lexicon (top). Number of occurrences in the corpus (bottom).	136
Figure 5.8	Can a word be compounded after letter k .	138
Figure 5.9	Goëlette profile for decomposition: branching factor as a function of length k for a simple word (left) and a three-word based compound (right).	139
Figure 5.10	Hierarchical representation for a complex decomposition.	141
Figure 5.11	OOV rates on training and dev_{96} data for different normalization versions V_i and 64,000 most frequently words from 40 million training data highlighting the importance of training epoch.	144
Figure 5.12	OOV rates for normalization versions V_0 , V_5 , and V_7 on dev_{96} text data, using 64,000 word lists derived from different training text sets.	144
Figure 5.13	Word list comparisons between pairs of languages. The number of shared words is shown as a function of word list size (including for each language its N most frequent items).	148
Figure 5.14	Pronunciation dictionary development for ASR system.	152
Figure 5.15	Pronunciation generation tool.	154
Figure 5.16	Example of letter-to-sound rules standard French, German, and English, and related exception. Rule precedence corresponds to listed order; ctx specifies letter contexts.	156
Figure 5.17	Examples of alternate valid pronunciations for American English and French.	159

Figure 5.18	Two example spectrograms of the word <i>coupon</i> : (left) /kjupan/ and (right) /kupan/. The grid 100 ms by 1 kHz.	160
Figure 5.19	An acoustic phone like segment is temporally modeled as a sequence of three states, each state being acoustically modeled by a weighted sum of Gaussian densities.	163
Figure 5.20	Impact on acoustic/temporal modeling depending on the choice of one or two symbols for affricates or diphthongs.	163
Figure 6.1	The classical communication channel model of automatic speech recognition.	170
Figure 6.2	Illustration of multiple back-off paths implicitly followed in a factored language model to evaluate the probability $P(w_n r_{n-1}, e_{n-1}, r_{n-2}, e_{n-2})$ of (6.10).	185
Figure 6.3	Dynamic adaptation of language models using contemporaneous data from another language.	195
Figure 8.1	Levels of signal abstraction by acoustic analysis along with components of information.	235
Figure 8.2	Human language identification rates for individual languages in the first (dark) and last (gray) quarters of the experiment, averaged over all participants (Muthusamy et al., 1994b).	236
Figure 8.3	An illustrative example of an HMM structure.	248
Figure 8.4	Example of a binary decision tree.	257
Figure 8.5	Phonotactic architecture with a single mono- or multilingual phone decoder.	259
Figure 8.6	Phonotactic architecture with multiple phone decoders.	260
Figure 8.7	Illustration of modeling cross-stream dependencies for a specific token.	261
Figure 9.1	Phonetic confusions for native English and native Japanese speakers, full phone set (top), specific phones (bottom).	281
Figure 10.1	Stochastic source-channel speech translation system.	320
Figure 10.2	Dynamic grammar acquisition as by-product of clarification dialogs using the GSG system on an e-mail client application.	323
Figure 10.3	Phrase alignment as sentence splitting.	326
Figure 10.4	The concept of N -grams (a) in sequences (b) in trees.	332
Figure 10.5	An alignment between an English phrase and its corresponding IF representation.	332
Figure 10.6	Translation quality with and without ASR score (acoustic model and source language model scores).	346
Figure 10.7	Block diagram of the ATR S2ST system.	348
Figure 10.8	Contextual splitting and temporal splitting.	349
Figure 10.9	An overview of the machine translation system developed in the C-cube project.	351

Figure 10.10	Examples of Hierarchical Phrase Alignment.	355
Figure 10.11	Examples of transfer rules in which the constituent boundary is “at.”	356
Figure 10.12	Example of TDMT transfer process.	357
Figure 10.13	Example of transfer rule generation.	358
Figure 10.14	Example of generated rules from the sentence “The bus leaves Kyoto at 11 a.m.”	360
Figure 10.15	Example of word alignment.	361
Figure 10.16	Example-based decoding.	362
Figure 10.17	A block diagram of a TTS module.	367
Figure 10.18	Data-collection environment of MAD.	375
Figure 10.19	The result of an evaluation experiment for naturalness between several TTS products. The horizontal bars at the top, middle, and bottom of the boxes indicate 75%, 50%, and 25% quartiles. Mean values are indicated by “x” marks.	383
Figure 10.20	Diagram of translation paired comparison method.	385
Figure 10.21	Procedure of comparison by bilingual evaluator.	386
Figure 10.22	Evaluation result with the translation paired comparison method.	388
Figure 10.23	Procedure of the automatic evaluation method.	388
Figure 10.24	Evaluation results in real environments.	393
Figure 11.1	The ISIS architecture.	415
Figure 11.2	Screen shot of the client that provides options for language selection and input mode selection (text or voice). Should “Text Input” be selected, a text box will appear.	416
Figure 11.3	Example of an XML message produced by the NLU server object.	416
Figure 11.4	Screen shot of the ISIS client object presenting information in response to the user’s query “Do you have the real-time quotes of Artel?”	424
Figure 11.5	Multiagent communication in ISIS. The messages communicated between the agents and the dialog manager server object are in XML format and wrapped with indicator tags.	427
Figure 11.6	The “Notify” icon indicates the arrival of a notification message.	429
Figure 11.7	The presence of both the “Notify” and the “Buy/Sell” icons indicate that there are pending notification message(s) and buy/sell reminders in the queue.	430
Figure 11.8	Structures and mechanisms supporting interruptions in ISIS. Online interaction (OI) and offline delegation (OD) are managed as separate dialog threads.	431
Figure 11.9	The AOPA software platform supports universal accessibility.	437
Figure 11.10	Browsing Web content by voice (see www.vxml.org).	438
Figure 11.11	Architecture of the bilingual CU voice browser.	442

List of Tables

Table 2.1	Distribution of the world's languages by geographical origin, percentage of the world's languages, and percentage of native speakers. Data from Gordon (2005).	7
Table 2.2	The twenty most widely spoken languages in the world according to the number of first-language speakers. Data from Gordon (2005).	8
Table 2.3	Illustration of derived forms in Arabic for the roots ktb (write) and drs (study).	18
Table 2.4	Korean speech levels: forms of the infinitive verb <i>gada</i> (go).	22
Table 3.1	European research programs funding speech and language technology.	56
Table 4.1	Writing systems and number of graphemes for twelve languages.	84
Table 4.2	Out-of-vocabulary rates for ten languages.	88
Table 4.3	Global unit set for twelve languages.	94
Table 4.4	Global feature unit set for five languages.	96
Table 4.5	Triphone coverage matrix for ten GlobalPhone languages; two numbers are given for each matrix entry (i, j) , meaning that language i is covered by language j with triphone types (upper number) and triphone tokens (lower number).	100
Table 4.6	Comparison between monolingual and multilingual articulatory feature detectors.	115
Table 4.7	Word error rates for English when decoding with articulatory feature detectors as additional stream.	115
Table 4.8	Phone versus grapheme-based speech recognition [word error rates] for five languages.	117
Table 5.1	For each version V_i ($i = 0, \dots, 7$) of normalized text, the elementary normalization steps N_j ($j = 0, \dots, 6$) are indicated by 1 in the corresponding column.	133
Table 5.2	Example words with ambiguous decompositions.	138

Table 5.3	Given a word start <i>Wbeg(k)</i> of length <i>k</i> , the number of character successors <i>#Sc(k)</i> generally tends toward zero with <i>k</i> . A sudden increase of <i>#Sc(k)</i> indicates a boundary due to compounding. <i>#Wend(k)</i> indicates the number of words in the vocabulary sharing the same word start.	139
Table 5.4	Examples of decomposition rules, including composita with imported English and French items; the number of occurrences of the decomposed items is given in parentheses.	140
Table 5.5	Lexical coverage and complementary OOV rates measured for different-size vocabularies on a 300-million word German text corpus. Measures are given with and without decomposition. The last two columns indicate the absolute and relative gains in OOV reduction rates.	141
Table 5.6	Some example rules to strip and add affixes used by a pronunciation generation tool. Affix types are P (prefix) and S (suffix).	155
Table 5.7	Pronunciation counts for inflected forms of the word <i>interest</i> in 150 hours of broadcast news (BN) data and 300 hours of conversational telephone speech (CTS).	165
Table 6.1	Czech is a free-word-order language, as illustrated by this set of sentences from Kuboň and Plátek (1994).	179
Table 6.2	Illustration of inflected forms in Czech for the underlying noun <i>žena</i> (woman).	180
Table 6.3	Illustration of derived forms in Arabic for the roots <i>ktb</i> (write) and <i>drs</i> (study) from Kirchoff et al. (2002a).	181
Table 6.4	Illustration of an agglutinated form in Inuktitut for the root word <i>tusaa</i> (hear).	182
Table 6.5	Typical vocabulary growth and rate of out-of-vocabulary words for various languages.	183
Table 6.6	The TDT-4 corpus covers news in three languages (Strassel and Glenn, 2003).	194
Table 8.1	Language identification accuracy rates over all test persons for test sections A (full speech), B (syllables), and C (prosody).	238
Table 8.2	Some acoustic LID systems and their performance rates.	250
Table 8.3	Examples of phonotactic LID systems and their recognition rates.	263
Table 8.4	Some prosodic components and their recognition rates.	265
Table 8.5	Some LVCSR LID components and their error rates.	267
Table 8.6	Comparison of basic LID approaches from an application development aspect.	269
Table 9.1	Targets for annotation with examples.	280

Table 9.2	Speaking rate and pause distribution statistics for native and non-native speakers of English in spontaneous versus read speech. Average phone duration and pause duration are measured in seconds. The pause-word ratio is the number of pauses inserted per word in speech.	282
Table 9.3	Frequent trigrams in native and non-native speech.	284
Table 9.4	WERs for accented speech using adapted dictionaries and weighted MLLR; baseline results on native speech are 11% for German and 12.3% for French.	297
Table 10.1	Corpus statistics for the NESPOLE! training and test set.	338
Table 10.2	Scores of the translations generated by systems <i>IL</i> , <i>SIL</i> , and <i>SMT</i> ; the values are percentages and averages of four independent graders.	338
Table 10.3	Training (test in parentheses) corpora.	340
Table 10.4	NIST scores for translation from Chinese to Spanish.	340
Table 10.5	Results for automatic disfluency removal on the English Verbmobil (EVM) and the Chinese CallHome (CCH) corpora.	343
Table 10.6	Optimal density and acoustic score weight based on utterance length.	345
Table 10.7	Optimal density and acoustic score weight based on utterance length when using acoustic and source language model scores.	346
Table 10.8	Summary of translation results for tight coupling between recognition and translation.	347
Table 10.9	Size of BTEC and SLDB.	371
Table 10.10	Characteristics of bilingual and monolingual travel conversation databases.	373
Table 10.11	Statistics of MAD corpora.	376
Table 10.12	Phoneme units for Japanese ASR.	376
Table 10.13	Perplexity for Japanese BTEC test set 01.	377
Table 10.14	Recognition performance for Japanese BTEC test set 01.	377
Table 10.15	Word accuracy rates [%] for two different language model combinations (the MDL-SSS acoustic model).	378
Table 10.16	Acoustic model performance comparison.	380
Table 10.17	Language model performance comparison.	380
Table 10.18	Subword units for Chinese ASR system.	381
Table 10.19	Token coverage rates of different subword units.	381
Table 10.20	Chinese character based recognition performance.	382
Table 10.21	Translation quality of four systems for BTEC.	384
Table 10.22	Translation performance.	390
Table 10.23	Translation performance for BTEC with/without duplications.	390
Table 10.24	Translation performance of test set without duplications.	391

Table 11.1	An example dialog in the stocks domain illustrating the capabilities of a state-of-the-art spoken dialog system (source: www.speechworks.com).	400
Table 11.2	An Example dialog from the CU FOREX hotline (Meng et al., 2000).	405
Table 11.3	Example rejection dialog in the ISIS system.	417
Table 11.4	An example dialog from the ISIS system.	418
Table 11.5	Example illustrating the automatic acquisition of a new stock name in ISIS through a spoken dialog between the user and the system.	423
Table 11.6	Example dialog illustrating the interruption of an online interaction dialog by an offline delegation alert.	432
Table 11.7	Example task list prepared by a participating subject.	435
Table 11.8	A human-computer dialog in the weather domain (C: computer, H: human).	439
Table 11.9	The VXML document that specifies the English dialog in Table 11.8 (explanations are boldface).	440
Table 11.10	A bilingual human-computer dialog implemented with VXML in the CU weather system.	443
Table 11.11	VXML document that specifies the bilingual CU weather dialog in Table 11.10 (explanations are boldface).	444

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 prize 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 à diriger des Recherches' [Document title: Traitement 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. Marelle 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.

This Page Intentionally Left Blank