Spoken Multimodal Human-Computer Dialogue in Mobile Environments
Contents

Preface xi
Contributing Authors xiii
Introduction xxi

Part I Issues in Multimodal Spoken Dialogue Systems and Components

1 Multimodal Dialogue Systems 3
Alexander I. Rudnicky
1. Introduction 3
2. Varieties of Multimodal Dialogue 4
3. Detecting Intentional User Inputs 6
4. Modes and Modalities 7
5. History and Context 7
6. Domain Reasoning 8
7. Output Planning 9
8. Dialogue Management 9
9. Conclusion 10
References 11

2 Speech Recognition Technology in Multimodal/Ubiquitous Computing Environments 13
Sadaoki Furui
1. Ubiquitous/Wearable Computing Environment 13
2. State-of-the-Art Speech Recognition Technology 14
3. Ubiquitous Speech Recognition 16
4. Robust Speech Recognition 18
5. Conversational Systems for Information Access 21
6. Systems for Transcribing, Understanding and Summarising Ubiquitous Speech Documents 24
7. Conclusion 32
References 33
3
A Robust Multimodal Speech Recognition Method using Optical Flow Analysis
Satoshi Tamura, Koji Iwano, Sadaoki Furui
1. Introduction 38
2. Optical Flow Analysis 39
3. A Multimodal Speech Recognition System 40
4. Experiments for Noise-Added Data 43
5. Experiments for Real-World Data 48
6. Conclusion and Future Work 49
References 52

4
Feature Functions for Tree-Based Dialogue Course Management
Klaus Macherey, Hermann Ney
1. Introduction 55
2. Basic Dialogue Framework 56
3. Feature Functions 59
4. Computing Dialogue Costs 63
5. Selection of Dialogue State/Action Pairs 64
6. XML-based Data Structures 65
7. Usability in Mobile Environments 68
8. Results 69
9. Summary and Outlook 74
References 74

5
A Reasoning Component for Information-Seeking and Planning Dialogues
Dirk Bühler, Wolfgang Minker
1. Introduction 77
2. State-of-the-Art in Problem Solving Dialogues 80
3. Reasoning Architecture 81
4. Application to Calendar Planning 85
5. Conclusion 88
References 90

6
A Model for Multimodal Dialogue System Output Applied to an Animated Talking Head
Jonas Beskow, Jens Edlund, Magnus Nordstrand
1. Introduction 93
2. Specification 97
3. Interpretation 103
4. Realisation in an Animated Talking Head 105
5. Discussion and Future Work 109
References 111
Contents

Part II  System Architecture and Example Implementations

7  
Overview of System Architecture  
Andreas Kellner
1. Introduction  117
2. Towards Personal Multimodal Conversational User Interface  118
4. Standardisation of Application Representation  126
5. Conclusion  129
References  130

8  
XISL: A Modality-Independent MMI Description Language  
Kouichi Katayama, Hirobumi Yamada, Yusaku Nakamura, Satoshi Kobayashi, Tsuneo Nitta
1. Introduction  133
2. XISL Execution System  134
3. Extensible Interaction Scenario Language  136
4. Three Types of Front-Ends and XISL Descriptions  140
5. XISL and Other Languages  146
6. Discussion  147
References  148

9  
A Path to Multimodal Data Services for Telecommunications  
Georg Nikifeld, Michael Pucher, Robert Finan, Wolfgang Eckhart
1. Introduction  149
2. Application Considerations, Technologies and Mobile Terminals  150
3. Projects and Commercial Developments  154
4. Three Multimodal Demonstrators  156
5. Roadmap for Successful Versatile Interfaces in Telecommunications  161
6. Conclusion  163
References  164

10  
Multimodal Spoken Dialogue with Wireless Devices  
1. Introduction  169
2. Why Multimodal Wireless?  171
3. Walking Direction Application  172
4. Speech Technology for Multimodal Wireless  173
5. User Interface Issues  174
6. Multimodal Architecture Issues  179
7. Conclusion  182
References  184
Contents

15
Enhancing the Usability of Multimodal Virtual Co-drivers 269
Niels Ole Bernsen, Laila Dybkjær
1. Introduction 269
2. The VICO System 271
3. VICO Haptics - How and When to Make VICO Listen? 272
4. VICO Graphics - When might the Driver Look? 274
5. Who is Driving this Time? 278
6. Modelling the Driver 280
7. Conclusion and Future Work 284
References 285

16
Design, Implementation and Evaluation of the SENECA Spoken Language Dialogue System 287
Wolfgang Minker, Udo Haiber, Paul Heisterkamp, Sven Scheible
1. Introduction 288
2. The SENECA SLDS 290
3. Evaluation of the SENECA SLDS Demonstrator 301
4. Conclusion 308
References 309

17
Segmenting Route Descriptions for Mobile Devices 311
Sabine Geldof, Robert Dale
1. Introduction 311
2. Structured Information Delivery 315
3. Techniques 315
4. Evaluation 322
5. Conclusion 326
References 327

18
Effects of Prolonged Use on the Usability of a Multimodal Form-Filling Interface 329
Janienke Sturm, Bert Cranen, Jacques Terken, Ilse Bakx
1. Introduction 329
2. The Matis System 332
3. Methods 335
4. Results and Discussion 337
5. Conclusion 345
References 346

19
User Multitasking with Mobile Multimodal Systems 349
Anthony Jameson, Kerstin Klöckner
1. The Challenge of Multitasking 350
2. Example System 354
3. Analyses of Single Tasks 354
4. Analyses of Task Combinations 359
5. Studies with Users 364
Preface

This book is based on publications from the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments held at Kloster Irsee, Germany, in 2002. The workshop covered various aspects of development and evaluation of spoken multimodal dialogue systems and components with particular emphasis on mobile environments, and discussed the state-of-the-art within this area. On the development side the major aspects addressed include speech recognition, dialogue management, multimodal output generation, system architectures, full applications, and user interface issues. On the evaluation side primarily usability evaluation was addressed. A number of high quality papers from the workshop were selected to form the basis of this book.

The volume is divided into three major parts which group together the overall aspects covered by the workshop. The selected papers have all been extended, reviewed and improved after the workshop to form the backbone of the book. In addition, we have supplemented each of the three parts by an invited contribution intended to serve as an overview chapter.

Part one of the volume covers issues in multimodal spoken dialogue systems and components. The overview chapter surveys multimodal dialogue systems and links up to the other chapters in part one. These chapters discuss aspects of speech recognition, dialogue management and multimodal output generation. Part two covers system architecture and example implementations. The overview chapter provides a survey of architecture and standardisation issues while the remainder of this part discusses architectural issues mostly based on fully implemented, practical applications. Part three concerns evaluation and usability. The human factors aspect is a very important one both from a development point of view and when it comes to evaluation. The overview chapter presents the state-of-the-art in evaluation and usability and also outlines novel challenges in the area. The other chapters in this part illustrate and discuss various approaches to evaluation and usability in concrete applications or experiments that often require one or more novel challenges to be addressed.

We are convinced that computer scientists, engineers, and others who work in the area of spoken multimodal dialogue systems, no matter if in academia or in industry, may find the volume interesting and useful to their own work.
Graduate students and PhD students specialising in spoken multimodal dialogue systems more generally, or focusing on issues in such systems in mobile environments in particular, may also use this book to get a concrete idea of how far research is today in the area and of some of the major issues to consider when developing spoken multimodal dialogue systems in practice.

We would like to express our sincere gratitude to all those who helped us in preparing this book. Especially we would like to thank all reviewers who through their valuable comments and criticism helped improve the quality of the individual chapters as well as the entire book. A special thank is also due to people at the Department of Information and Technology in Ulm and at NISlab in Odense.

Wolfgang MINKER

Dirk BÜHLER

Laila DYBKJÆR
Contributing Authors

Ilse Bakx is a Researcher at the Department of Technology Management, Technical University Eindhoven, The Netherlands. She obtained her MSc degree in Psychology (cognitive ergonomics) in 2001 at University of Maastricht. Her current research is dealing with the user aspects and usability of multimodal interaction.

Niels Ole Bernsen is Professor at, and Director of, the Natural Interactive Systems Laboratory, the University of Southern Denmark. His research interests include spoken dialogue systems and natural interactive systems more generally, including embodied conversational agents, systems for learning, teaching, and entertainment, online user modelling, modality theory, systems and component evaluation, including usability evaluation, system simulation, corpus creation, coding schemes, and coding tools.

Jonas Beskow is a Researcher at the Centre for Speech Technology at KTH in Stockholm, where he received his PhD in 2003. During 1998/99 he was a Visiting Researcher at the Perceptual Science Lab at UC Santa Cruz, sponsored by a Fulbright Grant. He received his MSc in Electrical Engineering from KTH in 1995. His main research interests are in the areas of facial animation, speech synthesis and embodied conversational agents.

Dan Bohus is a PhD candidate in the Computer Science Department at Carnegie Mellon University, USA. He has graduated with a BS degree in Computer Science from Politechnica University of Timisoara, Romania. His research is focussed on increasing the robustness and reliability of spoken language systems faced with unreliable inputs.

Jonathan Bloom received his PhD in Experimental Psychology, specifically in the area of psycholinguistics, from the New School for Social Research, New York, USA, in 1999. Since then, he has spent time designing speech user interfaces for Dragon Systems and currently for SpeechWorks International. For both companies, his focus has been on the design of usable multimodal interfaces.
Dirk Bühler is a PhD student at the University of Ulm, Department of Information Technology, Germany. He holds an MSc in Computer Science with a specialisation in computational linguistics from the University of Tübingen. His research interests are the development and evaluation of user interfaces, including dialogue modelling and multimodality, domain modelling, knowledge representation, and automated reasoning. He worked at DaimlerChrysler, Research and Technology, Germany, from 2000 to 2002.

Bob Carpenter received a PhD in Cognitive Science from the University of Edinburgh, United Kingdom, in 1989. Since then, he has worked on computational linguistics, first as an Associate Professor of computational linguistics at Carnegie Mellon University, Pittsburgh, USA, then as a member of technical staff at Lucent Technologies Bell Labs, and more recently, as a programmer at SpeechWorks International, and Alias I.

Sasha Caskey is a Computer Scientist whose main research interests are in the area of human-computer interaction. In 1996 he joined The MITRE Corporation in the Intelligent Information Systems Department where he contributed to research in spoken language dialogue systems. Since 2000 he has been a Researcher in the Natural Dialog Group at SpeechWorks International, New York, USA. He has contributed to many open source initiatives including the GalaxyCommunicator software suite.

Rachel Coulston is a Researcher at the Center for Human-Computer Communication (CHCC) in the Department of Computer Science at the Oregon Health & Science University (OHSU). She holds her BA and MA in Linguistics, and does research on linguistic aspects of human interaction with interactive multimodal computer systems.

Bert Cranen is a Senior Lecturer at the Department of Language and Speech, University of Nijmegen, The Netherlands. He obtained his masters degree in Electrical Engineering in 1979. His PhD thesis in 1987 was on modelling the acoustic properties of the human voice source. His research is focussed on questions how automatic speech recognition systems can be adapted to be successfully deployed in noisy environments and in multimodal applications.

Robert Dale is Director of the Centre for Language Technology at Macquarie University, Australia, and a Professor in that University's Department of Computing. His current research interests include low-cost approaches to intelligent text processing tasks, practical natural language generation, the engineering of habitable spoken language dialogue systems, and computational, philosophical and linguistic issues in reference and anaphora.
Courtney Darves is a PhD student at the University of Oregon in the Department of Psychology. She holds an MSc in Psychology (cognitive neuroscience) and a BA in Linguistics. Her research focuses broadly on adaptive human behaviour, both in the context of human-computer interaction and more generally in terms of neural plasticity.

Laila Dybkjær is a Professor at NISLab, University of Southern Denmark. She holds a PhD degree in Computer Science from Copenhagen University. Her research interests are topics concerning design, development, and evaluation of user interfaces, including development and evaluation of interactive speech systems and multimodal systems, design and development of intelligent user interfaces, usability design, dialogue model development, dialogue theory, and corpus analysis.

Wolfgang Eckhart visited the HTBLuVA in St. Pölten, Austria, before he worked at the Alcatel Austria Voice Processing Centre. Since 2001 he is employed at Sonorys Technology GesmbH with main focus on host-based Speech Recognition. In 2001 he participated in the research of ftw. project “Speech&More”.

Jens Edlund started out in computational linguistics at Stockholm University. He has been in speech technology research since 1996, at Telia Research, Stockholm, Sweden and SRI, Cambridge, United Kingdom and, since 1999, at the Centre for speech technology at KTH in Stockholm, Sweden. His research interests centre around dialogue systems and conversational computers.

Robert Finan studied Electronic Engineering at the University of Dublin, Ireland, Biomedical Instrumentation Engineering at the University of Dundee, United Kingdom, and Speaker Recognition at the University of Abertay, Dundee. He currently works for Mobilkom Austria AG as a Voice Services Designer. Since 2001 he participates in the research of ftw. project “Speech&More”.

Sadaoki Furui is a Professor at Tokyo Institute of Technology, Department of Computer Science, Japan. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction.

Sabine Geldof has a background in linguistics and artificial intelligence. As part of her dissertation she investigated the influence of (extra-linguistic) context on language production, more specifically in applications for wearable and mobile devices. Her post-doctoral research focuses on the use of natural language generation techniques to improve efficiency of information delivery in a task-oriented context.
Paul Heisterkamp has obtained his MA in German Philology, Philosophy and General Linguistics from Münster University, Germany, in 1986. Starting out in 1987 with the AEG research at Ulm, Germany, that later became DaimlerChrysler corporate research, he has worked on numerous national and international research projects on spoken dialogue. The current focus of his work is shifting from dialogue management and contextual interpretation to dialogue system integration in mobile environments with special respect to the aspects of multimodality in vehicle human-computer interfaces, as well as cognitive workload assessment.

Koji Iwano is an Assistant Professor at Tokyo Institute of Technology, Department of Computer Science, Japan. He received the BE degree in Information and Communication Engineering in 1995, and the ME and PhD degrees in Information Engineering in 1997 and 2000 respectively from the University of Tokyo. His research interests include speech recognition, speaker recognition, and speech synthesis.

Anthony Jameson is a Principal Researcher at DFKI, the German Research Center for Artificial Intelligence, and an adjunct Professor of Computer Science at the International University in Germany. His central interests concern interdisciplinary research on intelligent user interfaces and user-adaptive systems.

Kouichi Katsurada received the BE degree in 1995 and the PhD degree in 2000 from Osaka University, Japan. He joined Toyohashi University of Technology as a Research Associate in 2000. His current interests are in multimodal interaction and knowledge-based systems.

Andreas Kellner received his Diploma degree in Electrical Engineering from the Technical University Munich, Germany, in 1994. He has been working in the “Man-Machine Interfaces” department at the Philips Research Laboratories in Aachen since 1995. There, he was responsible for the development of spoken language dialogue systems and conversational user interfaces for various applications. He has also been involved in standardization efforts such as the W3C Voice Browser Working group. His main research areas of interests are natural language processing, dialogue management, and systems architectures.

Kerstin Klöckner studied Computational Linguistics at the University of the Saarland, Germany, where she obtained her Diploma in 2001. Since then, she has been working as a Researcher at DFKI’s Evaluation Center for Language Technology Systems.

Satoshi Kobayashi received the BE degree in 1991, the ME degree in 1994 from Toyohashi University of Technology, Japan, and the PhD degree in
2000 from Shizuoka University, Japan. He joined Toyohashi University of Technology as a Research Associate in 1999. His current interests are in multimodal interaction and language communication.

**Klaus Macherey** received the Diploma degree in Computer Science from the Aachen University of Technology (RWTH), Germany, in 1999. Since then, he has been a Research Assistant with the Department of Computer Science of RWTH. In 2002, he was a summer student at IBM T. J. Watson Research Center, Yorktown Heights, New York, USA. His primary research interests cover speech recognition, confidence measures, natural language understanding, dialogue systems, and reinforcement learning.

**Wolfgang Minker** is a Professor at the University of Ulm, Department of Information Technology, Germany. He received his PhD in Engineering Science from the University of Karlsruhe, Germany, in 1997 and his PhD in Computer Science from the University of Paris-Sud, France, in 1998. He was a Researcher at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), France, from 1993 to 1999 and a member of the scientific staff at DaimlerChrysler, Research and Technology, Germany, from 2000 to 2002.

**Yusaku Nakamura** received the BE degree in 2001 from Toyohashi University of Technology, Japan. Since 2001, he has been pursuing his Masters degree at Toyohashi University of Technology. He is presently researching multimodal interaction.

**Hermann Ney** received the Diploma degree in Physics in 1977 from Göttingen University, Germany, and the Dr.-Ing. degree in Electrical Engineering in 1982 from Braunschweig University of Technology, Germany. He has been working in the field of speech recognition, natural language processing, and stochastic modelling for more than 20 years. In 1977, he joined Philips Research, Germany. In 1985, he was appointed Department Head. From 1988 to 1989, he was a Visiting Scientist at Bell Laboratories, Murray Hill, New Jersey. In 1993, he joined the Computer Science Department of Aachen University of Technology as a Professor.

**Georg Niklfeld** studied Computer Science at the TU Vienna, Linguistics/Philosophy at the University of Vienna, Austria, and Technology Management at UMIST, United Kingdom. He did research in natural language processing at ÖFAI and later was employed as development engineer at a telecom equipment manufacturer. Since 2001 he works at ftw. as Senior Researcher and Project Manager for speech processing for telecommunications applications.
Tsuneo Nitta received the BEE degree in 1969 and the Dr. Eng. degree in 1988, both from Tohoku University, Sendai, Japan. After engaging in research and development at R&D Center of Toshiba Corporation and Multimedia Engineering Laboratory, where he was a Chief Research Scientist, since 1998 he has been a Professor in Graduate School of Engineering, Toyohashi University of Technology, Japan. His current research interest includes speech recognition, multimodal interaction, and acquisition of language and concept.

Magnus Nordstrand has been a Researcher at Centre for Speech Technology at KTH in Stockholm since 2001, after MSc studies in Electrical Engineering at KTH. Basic research interests focus on facial animation and embodied conversational agents.

Sharon Oviatt is a Professor and Co-Director of the Center for Human-Computer Communication in the Department of Computer Science at the Oregon Health & Science University, USA. Her research focuses on human-computer interaction, spoken language and multimodal interfaces, and mobile and highly interactive systems. In the early 1990s, she was a pioneer in the area of pen/voice multimodal interfaces, which now are being developed widely to support map-based interactions on hand-held devices and next-generation smart phones.

Michael Phillips is the Chief Technology Officer and co-founder of SpeechWorks International. In the early 80s, he was a Researcher at Carnegie Mellon University, Pittsburgh, USA. In 1987, he joined the Spoken Language Systems group at MIT's Laboratory for Computer Science where he contributed to the development of one of the first systems to combine speech recognition and natural language processing technologies to allow users to carry on full conversations within limited domains. In 1994, he left MIT, and started SpeechWorks, licensing the technology from the group at MIT.

Roberto Pieraccini started his research on spoken language human-computer interaction in 1981 at CSELT (now Telecom Italia Lab), Torino, Italy. He then joined AT&T Bell Laboratories, Murray Hill, New Jersey, USA, in 1990 and AT&T Shannon Laboratories, Florham Park, New Jersey, in 1995. Since 1999 he is leading the Natural Dialog group at SpeechWorks International, New York.

Michael Pucher studied philosophy at the University of Vienna and computational logic at Vienna University of Technology, Austria. Since 2001 he has been working at ftw. as a Researcher. His current research interests are multimodal systems, speech synthesis and voice services for telecommunications.
Alexander Rudnicky is involved in research that spans many aspects of spoken language, including knowledge-based recognition systems, language modelling, architectures for spoken language systems, multimodal interaction, the design of speech interfaces and the rapid prototyping of speech-to-speech translation systems. His most recent work has been in spoken dialogue systems, with contributions to dialogue management, language generation and the computation of confidence metrics for recognition and understanding. He is a recipient of the Allen Newell Award for Research Excellence.

Sven Scheible studied communications engineering at the University of Applied Sciences in Ulm, Germany, where he obtained his Diploma in 1999. Since then, he has been working in the research department of Temic, Germany, for three years. During this time he joined the EU research project SENECA where he was responsible for the application development and system integration. Afterwards, he moved to the product development department and is currently responsible for tools supporting the grammar and dialogue implementation process.

Stephen Springer has over 19 years of experience in the design and implementation of intelligent language systems. He managed the Speech Services Technology Group at Bell Atlantic, where he worked with Victor Zue's Spoken Language System Group at MIT. At SpeechWorks International, he has designed enterprise systems that have handled over 10,000,000 calls, with transaction completion rates exceeding 95%. He leads the international User Interface Design team at SpeechWorks.

Janienke Sturm is a Researcher at the Department of Language and Speech of the University of Nijmegen. She graduated as computational linguist at the University of Utrecht, The Netherlands, in 1997. Since then her research focussed mainly on design and evaluation of spoken dialogue systems for information services.

Satoshi Tamura is a PhD candidate at Tokyo Institute of Technology (TTT), Japan. He received the ME degree in Information Science and Engineering from TTT in 2002. His research interests are speech information processing, especially multimodal audio-visual speech recognition.

Jacques Terken has a background in experimental psychology and received a PhD in 1985. He has conducted research on the production and perception of prosody and on the modelling of prosody for speech synthesis. Currently, his research interests include the application of speech for human-computer interaction, mainly in the context of multimodal interfaces.
Marilyn Walker is a Royal Society Wolfson Professor of Computer Science and Director of the Cognition and Interaction Lab at the University of Sheffield in England. Her research interests include the design and evaluation of dialogue systems and methods for automatically adapting such systems through experience with users. She received her PhD in Computer and Information Science from the University of Pennsylvania in 1993 and an MSc in Computer Science from Stanford University in 1988. Before coming to Sheffield, she was a Principal Research Scientist at AT&T Shannon Labs.

Matt Wesson is a Research Programmer at the Center for Human-Computer Communication (CHCC) in the Department of Computer Science at the Oregon Health & Science University (OHSU). He holds a BA in English and an MA in Computer Science.

Steve Whittaker is the Chair of the Information Retrieval Department at the University of Sheffield, United Kingdom. His main interests are in computer-mediated communication and human-computer interaction. He has designed and evaluated videoconferencing, email, voicemail, instant messaging, shared workspace and various other types of collaborative tools to support computer-mediated communication. He has also conducted extensive research into systems to support multimodal interaction, including speech browsing and multimodal mobile information access.


Hirobumi Yamada received the BE degree in 1993 and the PhD degree in 2002 from Shinshu University, Japan. He joined Toyohashi University of Technology as a Research Associate in 1996. His current interests are in multimodal interaction, E-learning systems and pattern recognition.
Introduction

Spoken multimodal human-computer interfaces constitute an emerging topic of interest not only to academia but also to industry. The ongoing migration of computing and information access from the desktop and telephone to mobile computing devices such as Personal Digital Assistants (PDAs), tablet PCs, and next generation mobile phones poses critical challenges for natural human-computer interaction. Spoken dialogue is a key factor in ensuring natural and user-friendly interaction with such devices which are meant for everybody. Speech is well-known to all of us and supports hands-free and eyes-free interaction, which is crucial, e.g. in cars where driver distraction by manually operated devices may be a significant problem. Being a key issue, non-intrusive and user-friendly human-computer interaction in mobile environments is discussed by several chapters in this book.

Many and increasingly sophisticated over-the-phone spoken dialogue systems providing various kinds of information are already commercially available. On the research side interest is progressively turning to the integration of spoken dialogue with other modalities such as gesture input and graphics output. This process is ongoing both regarding applications running on stationary computers and those meant for mobile devices. The latter is witnessed by many of the included chapters.

In mobile environments where the situation and context of use is likely to vary, speech-only interaction may sometimes be the optimal solution while in other situations the possibility of using other modalities possibly in combination with speech, such as graphics output and gesture input, may be preferable.

Users who interact with multimodal devices may benefit from the availability of different modalities in several ways. For instance, modalities may supplement each other and compensate for each others’ weaknesses, a certain modality may be inappropriate in some situations but the device and its applications can then still be used via another modality, and users’ different preferences as to which modalities they use can be accommodated by offering
different modalities for interaction. Issues like these are also discussed in several of the included chapters in particular in those dealing with usability and evaluation issues.

We have found it appropriate to divide the book into three parts each being introduced by an overview chapter. Each chapter in a part has a main emphasis on issues within the area covered by that part. Part one covers issues in multimodal spoken dialogue systems and components, part two concerns system architecture and example implementations, and part three addresses evaluation and usability. The division is not a sharp one, however. Several chapters include a discussion of issues that would make them fit almost equally well under another part. In the remainder of this introduction, we provide an overview of the three parts of the book and their respective chapters.

Issues in Multimodal Dialogue Systems and Components.
The first part of the book provides an overview of multimodal dialogue systems and discusses aspects of speech recognition, dialogue management including domain reasoning and inference, and multimodal output generation. By a multimodal dialogue system we understand a system where the user may use more than one modality for input representation and/or the system may use more than one modality for output representation, e.g. input speech and gesture or output speech and graphics.

In his overview chapter Rudnicky discusses multimodal dialogue systems and gives a bird’s-eye view of the other chapters in this part. He discerns a number of issues that represent challenges across individual systems and thus are important points on the agenda of today’s research in multimodal dialogue systems. These issues include the detection of intentional user input, the appropriate use of interaction modalities, the management of dialogue history and context, the incorporation of intelligence into the system in the form of domain reasoning, and finally, the problem of appropriate output planning.

On the input side speech recognition represents a key technique for interaction, not least in ubiquitous and wearable computing environments. For the use of speech recognition to be successful in such environments, interaction must be smooth, unobtrusive, and effortless to the user. Among other things this requires robust recognition also when the user is in a noisy environment.

Two chapters in this part deal with the robustness issue of speech recognition systems. Furui provides an overview of the state-of-the-art in speech recognition. Moreover, he addresses two major application areas of speech recognition technology. One application area is that of dialogue systems. The user speaks to a system e.g. to access information. A second major area using speech technology is that of systems for transcription, understanding, and summarisation of speech documents, e.g. meeting minute transcription systems. Furui discusses the very important issue of how to enhance the robustness of speech
recognisers facing acoustic and linguistic variation in spontaneous speech. To this end he proposes a paradigm shift from speech recognition to speech understanding so that the recognition process rather delivers the meaning of the user's input than a word to word transcription.

Tamura et al. discuss audio-visual speech recognition, a method that draws not only on the speech signal but also takes visual information, such as lip movements, into account. This approach seems promising in improving speech recognition accuracy not least in noisy environments. The authors propose a multimodal speech recognition method using optical flow analysis to extract visual information. The robustness of the method to acoustic and visual noises has been evaluated in two experiments. In the first experiment white noise was added to the speech wave form. In the second experiment data from a car was used. The data was distorted acoustically as well as visually. In both experiments significantly better results were achieved using audio-visual speech recognition compared to using only audio recognition.

The next two chapters in this part by Macherey and Ney and Bühler and Minker focus on aspects of dialogue management. Ideally dialogue managers should be application-independent. To achieve this, one must, according to Macherey and Ney, distill the steps which many domains have in common leading to parameterisable data structures. Macherey and Ney propose trees as an appropriate way in which to represent such data structures. Using a tree-based structure the focus of the chapter is on dialogue course management, dialogue cost features, and the selection of dialogue actions. Based on proposed cost functions the dialogue manager should in each state be able to choose those actions which are likely to lead as directly as possible through the tree to the user's goal. Encouraging results are presented from evaluating an implemented tree-based dialogue manager in a telephone directory assistance setting.

Bühler and Minker present a logic-based problem assistant, i.e. a reasoning component which interacts with and supports dialogue management. The problem assistant constantly draws on various contextual information and constraints. In case of conflicts between new constraints provided by the user and information already in the system, the problem assistant informs the dialogue manager about its inferences. Thereby it enables the dialogue manager to explain the problem to the user and possibly propose solutions to it. The functionality of the problem assistant is illustrated on calendar planning using a scenario in which a user plans a series of meeting appointments in various locations.

The last chapter in part one concerns multimodal output generation. Beskow et al. present a formalism for GEneric System Output Markup (GESOM) of verbal and non-verbal output. The idea is that the dialogue manager will use the markup formalism to annotate the output to be produced to the user next.
The markup provides information about the communicative functions of output but it does not contain details on how these are to be realised. The details of rendering are encapsulated in a generation component, allowing the dialogue manager to generate fairly abstract requests for output which are then rendered using the relevant and available output devices. This is valuable both for applications that operate over a variety of output devices and also for developers who no longer need to spend time coding details of generation. The markup formalism has been used in an animated talking head in the Swedish real-estate AdApt system. The use of GESOM in this system is also discussed in the chapter.

System Architecture and Example Implementations. The second part of the book discusses architectural issues in system design and example implementations. Most existing implementations of multimodal and unimodal dialogue systems are based on architectural infrastructures that allow for a distribution of computation between host computers, operating systems, and programming languages. With multimodal dialogue systems evolving from speech-only server-based systems, such as call centre automation systems, to personal multimodal and mobile interaction partners, such as PDAs and mobile phones, a new dimension of requirements is being placed on the underlying architectures.

The overview chapter by Kellner presents an analysis of these requirements and applies it to two particular system architectures, the Galaxy Communicator infrastructure used by the US American DARPA (Defence Research Project Agency) Community and the SmartKom testbed that served as a basis for a German national research project where partners from university and industry were involved. The common goal of both of these frameworks is to integrate several independently developed modules into one coherent system. Multi-modality and cooperative behaviour are key factors, requiring the base architectures to provide a more sophisticated handling of streams of information than those used for implementing traditional telephone-based systems. Related to this comparison, an overview of existing and emerging standards for speech-enabled applications, such as VoiceXML and SALT, is given. Kellner also pays attention to a second emerging requirement, namely, the need to enable the user to access and combine different applications through a coherent user interface (e.g., using an assistant metaphor impersonated as an avatar.)

As claimed by Katsurada et al. in their chapter multimodal dialogue systems, in particular those being used in different environments or on hardware as distinct as PCs, PDAs, and mobile phones, require abstractions from traditional presentation management techniques, such as hypertext (HTML). These abstractions should enable a developer to describe modalities more flexibly, so that it becomes possible to add or modify modalities as needed when port-
Introduction

ing an application to new devices supporting new types of modalities. Ideally, application logic that is not modality-dependent should be reusable on all devices. To this end, a modality-independent Man-Machine Interface (MMI) description language called XISL (eXtensible Interaction Scenario Language) is proposed. The authors describe the implementation of three execution environments for XISL: a PC terminal, a mobile phone terminal, and a PDA terminal. The PC and the PDA feature multimodal interaction via touch screen, keyboard, and audio device, while the phone uses speech and Dual Tone Multi-Frequency (DTMF).

Mobile terminal capabilities seem especially relevant since appealing applications and services are necessary to convince industrial developers and device manufacturers of the possibilities for the commercial exploitation of multimodal interfaces. As described in the chapter by Niklfield et al., these interfaces must be tailored to the specific capabilities and limitations of the end device, which is particularly important for mobile phones that may be based on different standards such as GPRS, UMTS, or WLAN. It is shown that multimodality can indeed bring about usability advantages for specific applications, such as a map service.

Pieraccini et al. also discuss the various issues related particularly to the design and implementation of multimodal dialogue systems with wireless handheld devices. Focus is on the design of a usable interface that exploits the complementary features of the audio and visual channels to enhance usability. One aspect arising from the mobility of the user is the fact that the handheld devices could potentially be used in a variety of different situations in which certain channels are or are not preferred. Pieraccini et al. present two implementations of client-server architectures demonstrated by map and navigation applications.

Also dealing with map interaction and navigation as an application of multimodal interaction, the chapter by Bühlcr and Minker presents the mobile scenario of the SmartKom project. Like Pieraccini et al. the authors focus on a specific issue in mobile multimodal dialogue systems, namely the required ability of the system to dynamically adapt itself or be adaptable by the user to the current environment of use in terms of the modalities used for interaction. The authors present situations in SmartKom's integrated driver and pedestrian scenario in which the user might want to change the modalities used by the system, or in which the system might decide to enable or disable certain channels. From a developer's point of view this is also related to the modality-independent MMI description language presented in the chapter by Katsurada et al., but the focus is on different mobile situations of use of one device rather than on the use of a single application on different devices.

Finally, the required adaptations of a dialogue system when porting it to a new application domain and environment of use is also investigated in the
chapter by Bohus and Rudnicky. The intended use of a dialogue system as an aircraft maintenance and repair assistant requires major adaptations and adjustments of existing dialogue technologies, originally developed for telephone-based problem solving. Multimodality constitutes an important factor in this development process. Results from two field evaluations of the maintenance and repair system are reported.

Evaluation and Usability. The chapters in the third and last part of the book have a main focus on evaluation and usability issues. Evaluation and usability of dialogue systems, unimodal as well as multimodal, are as crucial as ever, and the importance is likely to increase along with the technical advances and market growth in the field. The chapters included here give an impression of some of the research challenges being addressed today and some which will soon need our attention.

Dybkjaer et al. provide an overview of the state-of-the-art in evaluation and usability and review a number of initiatives with a main focus on evaluation and/or usability. In addition they point to a number of challenges ahead. Clearly there has been significant progress on unimodal dialogue systems evaluation and usability. However, the emergence of, among others, multimodal, mobile, and non-task-oriented systems pose entirely new challenges. For instance we need to address issues such as online user modelling, emotions, non-task-oriented dialogue, mobile environments, and user preferences and priorities in interaction. These issues are important to the usability of many future and emerging systems and must of course also be evaluated. A major challenge is in realising these issues in appropriate ways in systems and in finding adequate ways in which to evaluate them.

Some of the challenges mentioned by Dybkjaer et al. are addressed in more detail by other chapters in this part of the book. User models are addressed both by Whittaker and Walker and by Bernsen and Dybkjaer. Mobile environments are mentioned by all chapters but are central to the applications in Bernsen and Dybkjaer, Minker et al., Geldof and Dale, Sturm et al., and Jameson and Klöckner. The ability to take into account user preferences and priorities is of course central to user models but aspects of the issue are also discussed in the chapters by Sturm et al. and Jameson and Klöckner and to some extent by Geldof and Dale. An important issue is that of adaptability. To which extent will users naturally adapt to the system and in which ways must the system be able to adapt to users in order to be accepted? This question is touched upon by most chapters in this part but most clearly in Sturm et al., Jameson and Klöckner, and Oviatt et al.

The chapter by Whittaker and Walker addresses the problem of how to select the most relevant options to mention and what to say about them when the task domain is complex as in restaurant information. The hypothesis is
that a user model may be an adequate solution to the problem. Results from 
a Wizard-of-Oz experiment showed that interaction strategies tailored to user 
requirements and which reduce dialogue length are preferable. In a subse-
quent experiment subjects overheard dialogues about restaurant selection and 
provided information quality judgements for each dialogue. Dialogues with 
individual user models scored higher than those with a default user model.

Bernsen and Dybkjær discuss four user-oriented design analysis problems in 
the context of a research prototype of an in-car system. The prototype is mul-
timodal and accepts input via speech and via a push-to-activate button which 
will activate speech recognition. Output is provided in terms of speech and 
on the in-car display. The prototype enables navigation assistance to addresses 
and to points of interest and hotel reservation. The problems addressed con-
cern when the system should (not) listen, how to optimise use of the display, 
driver identification, and online adaptive user modelling. Possible solutions to 
these four problems are proposed and discussed regarding their pros and cons.

Like Bernsen and Dybkjær, Minker et al. address in-vehicle applications. 
They describe a spoken command-based user interface for telephone and navi-
gation applications. The system can be operated via speech or manually either 
via input from a remote control (for navigation) or by using the keypad on 
the phone. Among the topics addressed in the chapter is the evaluation of the 
system with users who were driving while carrying out various tasks given to 
them. The evaluation showed that speech input was faster, driving skills were 
in general less affected by speech input, and users preferred speech input due 
to e.g., better safety and comfort and less distraction. However, task comple-
tion rate was considerably lower with speech input compared to manual input 
the main reason being forgotten command words and use of out-of-vocabulary 
words.

Geldof and Dale also address navigation though not as an in-car application. 
They focus on route descriptions via mobile devices and how to present the 
descriptions in an adequate way. The descriptions must be recognisable and 
rememberable and the small screen size of mobile devices must be taken into 
account. The approach proposed is to summarise the route description. This 
description can then be expanded in a tree-like way to provide more detail if 
desired by the user. The approach has been evaluated and compared to a route 
description representation in terms of a flat numbered list of instructions. The 
evaluation was performed in a car. Three teams on half of the route used the 
tree representation while for the other half of the route the list representation 
was used. Two teams preferred the tree structure while one team preferred the 
list structure.

Sturm et al. present experience from a study on users' development of in-
teraction patterns over time. The system used in the study is a train timetable 
information system designed for small portable devices. The system accepts
spoken and pointing input and provides spoken and graphics output. After an introduction subjects completed six scenarios as a pre-test. Then they practised for a couple of hours in total divided on different days. In a post-test the first six scenarios were carried out again. Evaluation showed that while effectiveness (measured as dialogue success rate) was not really affected, efficiency (measured as task completion time) clearly improved over time and interaction patterns changed. For instance, the users learned how to speak to the system, how to use pointing for more reliable data entry and error correction, and how to speed up interaction in various ways. 

Jameson and Klöckner discuss the implications of user multitasking for the design of mobile multimodal systems. They address the tasks of using a mobile phone while walking. Eye-based dialling and ear-based dialling are analysed theoretically, and possible resource conflicts are pointed out when adding the task of walking. In an experiment subjects were asked to walk around in a room with several obstacles while calling different phone numbers and talking to those who answered the call. Half of them used eye-based dialling, one used ear-based dialling, while two used a hybrid approach. Several different subjective factors seemed to influence their preferences, such as habits from other tasks, dislikes related to design, and ideas of what is socially acceptable. A conclusion of the chapter is that the observed factors may be hard to be aware of and design for but nevertheless need attention since they seem crucial to user acceptance.

In the final chapter of this part, Oviatt et al. investigate to which extent the basic acoustic-prosodic features of users' speech are influenced by text-to-speech (TTS). Young children interacted for an hour each with animated marine animals. In addition to spoken input and output the system enabled pen-based input and graphics output. The output voices were tailored to represent opposite ends of the introvert-extrovert personality spectrum. Amplitude, duration, and dialogue response latencies were measured. Good convergence with the TTS voice was found: When exposed to an extrovert voice, the children increased their amplitude and decreased utterance duration and dialogue response latencies, and contrarily, when interacting with an introvert voice. The hope is that future systems may be able to guide users' speech to fall within a range that is easily processed by the recogniser without any explicit instructions to users.
ISSUES IN MULTIMODAL SPOKEN DIALOGUE SYSTEMS AND COMPONENTS