TEXT, SPEECH AND LANGUAGE TECHNOLOGY

SPOKEN MULTIMODAL HUMAN-COMPUTER DIALOGUE IN MOBILE ENVIRONMENTS

Edited by W. Minker, Dirk Bühler and Laila Dybkjær

Springer

Spoken Multimodal Human-Computer Dialogue in Mobile Environments

Edited by

W. Minker
University of Ulm, Germany

Dirk Bühler
University of Ulm, Germany

and

Laila Dybkjær
University of Southern Denmark, Odense, Denmark

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-3073-8 (HB)
ISBN 1-4020-3075-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Springer,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Springer,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands

Spoken Multimodal Human-Computer Dialogue in Mobile Environments

Text, Speech and Language Technology

VOLUME 28

The titles published in this series are listed at the end of this volume.

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 reseach 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 19988 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 (TIT), Japan. He received the ME degree in Information Science and Engineering from TIT 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.

Eric Woudenberg began work in speech recognition at ATR, Kyoto, Japan, in 1993. He joined Bell Laboratories, Murray Hill, New Jersey, USA, in 1998, and SpeechWorks International, New York, in 2000. Since 2002 he has been a senior developer at LOBBY7, Boston, working on commercially deployable multimodal systems.

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