

Spoken Dialogue Systems Technology and Design

Wolfgang Minker • Gary Geunbae Lee
Satoshi Nakamura • Joseph Mariani
Editors

Spoken Dialogue Systems Technology and Design

 Springer

Editors

Wolfgang Minker
Ulm University
Institute of Information Technology
Albert-Einstein-Allee 43
89081 Ulm
Germany
wolfgang.minker@uni-ulm.de

Satoshi Nakamura
Keihanna Research Laboratories
National Institute of Information
and Communications Technology
Kyoto, Japan
satoshi.nakamura@nict.go.jp

Gary Geunbae Lee
Department of Computer Science
and Engineering
Pohang University of Science
and Technology (POSTECH)
San 31, Hyoja-dong
790-784 Pohang, Kyungbuk
Nam-Gu
Korea, Republic of (South Korea)
gblee@postech.ac.kr

Joseph Mariani
Centre National de la Recherche Scientifique
Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur (LIMSI)
and
Institute for Multilingual and Multimedia
Information (IMMI)
B.P. 133
91403 Orsay cedex, France
joseph.mariani@limsi.fr

ISBN 978-1-4419-7933-9 e-ISBN 978-1-4419-7934-6
DOI 10.1007/978-1-4419-7934-6
Springer New York Dordrecht Heidelberg London

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Spoken dialogue systems have become an increasingly important interface between humans and computers as they constitute the most natural way of communication.

This book covers all key topics in spoken language dialogue interaction from a variety of leading researchers. It brings together several perspectives in the areas of spoken dialogue analysis (Chapters 1-3), processing emotions in dialogue (Chapters 4-5), multimodality (Chapters 6-9) as well as resources and evaluation (Chapters 10-11).

The book is based on a selected subset of papers from the International Workshop on Spoken Dialogue Systems Technology (IWSDS) held at Kloster Irsee, Germany in 2009. There, the latest formal, corpus-based, implementational and analytical work on spoken dialogue systems technology have been discussed among researchers and developers in the area, from both industry and academia. The IWSDS workshop series (former Tutorial and Research Workshop Series at Kloster Irsee) provides since 1999 a regular forum for the presentation of research in the discourse and dialogue area to both the larger Spoken Dialogue Systems community as well as to researchers outside this community.

The 2009 workshop programme featured 36 papers from more than 14 different countries representing the 3 continents. Of these nine were invited for publication in this book along with two papers by invited speakers, i.e. a total of eleven papers. All workshop papers were extended and revised before they were submitted as book chapters. Each chapter has subsequently been reviewed by at least two reviewers and further improved on the basis of their comments.

We would like to thank all those who contributed to and helped us in preparing this book. In particular we would like to express our gratitude to the following reviewers for their valuable comments and criticism on the submitted drafts of the book chapters: Jan Alexandersson, Hélène Bonneau-Maynard, Rainer Gruhn, Joakim Gustafson, Li Haizhou, Gerhard Hanrieder, Paul Heisterkamp, Harald Hüning, Tatsuya Kawahara, HongKook Kim, Lin-Shan Lee, López-Cózar, Mike McTear, Sebastian Möller, Mikio Nakano, Elmar Nöth,

Tim Paek, Patrick Paroubek, Norbert Reithinger, Laurent Romary, Gabriel Skantze and Harald Traue. We are also grateful to Kseniya Zablostkaya and Sergey Zablotkiy at the Institute of Information Technology at the University of Ulm for their support in editing the book.

In the following we give an overview of the book contents by providing excerpts of the chapter abstracts.

Spoken Dialogue Analysis In Chapter 1 **Raab et al.** propose a solution for automatically recognizing several languages simultaneously (Raab et al., 2010). Keeping the recognition feasible on embedded systems and handling the native speech of the user with highest priority constitute the main goals. The authors describe experiments that address the impact of non-native accent. One potential application of the approach is to extend existing in-car infotainment systems to international navigation input and music selection via voice.

Misu et al. propose in Chapter 2 an efficient online learning method for dialogue management using speech (Misu et al., 2010). The method is based on the Bayes risk criterion for document retrieval. The framework has been extended to be trainable via online learning. The performance is evaluated and compared with a reinforcement learning method in terms of convergence speed.

According to **Lopez-Cozar et al.** (Chapter 3) continuous advances in the field of spoken dialogue systems make the processes of design, implementation and evaluation increasingly complex (López-Cózar et al., 2010). A technique which has attracted interest during the last decades is based on the automatic generation of dialogues between the system and a user simulator. The authors describe the main methodologies and techniques developed to create such user simulators. Additionally, a user simulation technique is proposed. It is based on a novel approach to simulating different levels of user cooperativeness, which allows carrying out a more detailed system assessment.

Processing Emotions in Dialogue In Chapter 4 **Polzehl et al.** report on performance optimization techniques for anger classification using acoustic cues (Polzehl et al., 2010). The authors evaluate the performance of a broad variety of features on voice portal databases which contain non-acted, continuous speech of narrow-band quality. After determining optimal sets of feature combinations, classification scores have been calculated using discriminative training and Support-Vector Machine classification.

Psychomimes have become increasingly important as they reflect the emotional status of the speaker and frequently appear in communication. **Kurosawa et al.** experimentally classified psychomimes (Chapter 5) by using a self-organizing map algorithm as the basis of implementing a spoken dialogue application with emotional agents (Kurosawa et al., 2010). The results are

represented as maps with meaningful three-vector dimensions, i.e., three verb classes are assigned using a thesaurus dictionary.

Multimodality Spoken dialogue technology has developed considerably over the past thirty years both in terms of research activity and commercially deployed applications. In Chapter 6 **McTear** presents an overview of trends in dialogue research based on an analysis of papers from past Eurospeech-Interspeech conferences (McTear, 2010). Some challenges are identified, in particular, the issues faced when building viable end-to-end dialogue systems. Difficulties are often encountered when trying to access the many resources and toolkits that are required for dialogue research. The chapter concludes with a discussion of some opportunities for future research, including the integration of dialogue technology into voice search applications and the application of spoken dialogue technology in ambient intelligence environments.

Partially Observable Markov Decision Processes (POMDPs) are applied in action control to manage and support natural dialogue communication with conversational agents. In Chapter 7, **Minami et al.** present an approach to resolve robustness and flexibility issues by applying rewards, from statistics of agent action predictive probabilities, to POMDP value iterations (Minami et al., 2010). These rewards can generate an action sequence whose predictive distribution is maximized under POMDP conditions.

In Chapter 8 **Araki and Hattori** present a novel rapid prototyping method for a spoken multimodal dialogue system (Araki and Hattori, 2010). In contrast to previous approaches that define individual dialogue states the proposed method automatically generates these states from a data model definition. The authors show rapid development examples of various types of spoken multimodal dialogue systems.

In Chapter 9, **Weiss et al.** discuss the contributions of different modalities to the overall quality of multimodal interaction (Weiss et al., 2010). The authors present experimental results from several multimodal scenarios, involving different interaction paradigms, degrees of interactivity, and modalities. The results show that the impact of each modality on overall quality in interaction significantly depends on the scenario and degree of interactivity.

Resources and Evaluation Chapter 10 by **Ohtake et al.** introduce a new corpus of consulting dialogues designed for training a dialogue manager (Ohtake et al., 2010). The authors have collected more than 150 hours of dialogues in a tourist guidance domain. The chapter outlines the employed taxonomy of dialogue act annotation.

Statistical modeling for dialogue systems development requires large amounts of dialogues annotated with function labels (usually dialogue acts). This annotation is a time-consuming process. In Chapter 11 **Tamarit et al.**

compare two statistical models for dialogue annotation, a classical Hidden Markov Model and a novel N-gram Transducers model (Tamarit et al., 2010). This latter one should allow for faster data annotation. A comparison between both models is performed on two standard corpora.

Wolfgang Minker
Ulm University
Germany

References

- Araki, M. and Hattori, T. (2010). Proposal for a Practical Spoken Dialogue System Development Method. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Kurosawa, Y., Mera, K., and Takezawa, T. (2010). Psychomime Classification and Visualization Using a Self-Organizing Map. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- López-Cózar, R., Griol, D., Espejo, G., Callejas, Z., and Ábalos, N. (2010). Towards Fine-Grain User-Simulation For Spoken Dialogue Systems. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- McTear, M. (2010). Trends, Challenges and Opportunities in Spoken Dialogue Research. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Minami, Y., Mori, A., Meguro, T., Higashinaka, R., Dohsaka, K., and Maeda, E. (2010). Dialogue Control by POMDP using Dialogue Data Statistics. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Misu, T., Sugiura, K., Kawahara, T., Ohtake, K., Hori, C., Kashioka, H., and Nakamura, S. (2010). Online Learning of Bayes Risk-Based Optimization of Dialogue Management. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Ohtake, K., Misu, T., Hori, C., Kashioka, H., and Nakamura, S. (2010). Dialogue Acts Annotation to Construct Dialogue Systems for Consulting. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Polzehl, T., Schmitt, A., and Metze, F. (2010). Approaching Multi-Lingual Emotion Recognition from Speech - On Language Dependency of Acoustic/Prosodic Features for Anger Detection. In *SpeechProsody*, Chicago, U.S.A.
- Raab, M., Gruhn, R., and Nöth, E. (2010). Multilingual Speech Interfaces for Resource-constrained Dialogue Systems. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.

- Tamarit, V., Martínez-Hinarejos, C.-D., and Benedí, J.-M. (2010). On the Use of N-gram Transducers for Dialogue Annotation. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.
- Weiss, B., Möller, S., Wechsung, I., and Kühnel, C. (2010). Quality of Experiencing Multi-Modal Interaction. In *"Spoken Dialogue Systems Technology and Design"*. Springer, This Edition.

Contents

Preface	v
References	viii
Contributing Authors	xv
1	
Multilingual Speech Interfaces for Resource-Constrained Dialogue Systems	1
<i>Martin Raab, Rainer Gruhn and Elmar Nöth</i>	
1. Introduction	2
2. Literature Review	3
3. Approach	5
4. Experimental Setup	5
5. Accent Adaptation	7
6. Scalable Architecture	15
7. Summary	24
References	25
2	
Online Learning of Bayes Risk-Based Optimization	29
<i>Teruhisa Misu, Komei Sugiura, Tatsuya Kawahara and Kiyonori Ohtake, Chiori Hori, Hideki Kashioka, Satoshi Nakamura</i>	
1. Introduction	30
2. Dialogue Management and Response Generation in Document Retrieval System	32
3. Optimization of Dialogue Management in Document Retrieval System	37
4. Online Learning of Bayes Risk-based Dialogue Management	42
5. Evaluation of Online Learning Methods	46
6. Conclusions	48
References	49
3	
Towards Fine-Grain User-Simulation For Spoken Dialogue Systems	53
<i>Ramón López-Cózar, David Griol and Gonzalo Espejo, Zoraida Callejas, Nieves Ábalos</i>	
1. Introduction	54

2.	Related Work	56
3.	Our User Simulators	66
4.	Experiments	69
5.	Results	71
6.	Conclusions	76
	Acknowledgments	77
	References	77
4		
	Salient Features for Anger Recognition	83
	<i>Tim Polzehl, Alexander Schmitt and Florian Metze</i>	
1.	Introduction	84
2.	Related Work	85
3.	Overview of Database Conditions	85
4.	Selected Corpora	87
5.	Prosodic and Acoustic Modeling	88
6.	Feature Ranking	92
7.	Normalization	94
8.	Classification	95
9.	Experiments and Results	97
10.	Discussion	101
11.	Conclusions	102
	Acknowledgments	103
	References	103
5		
	Psychomime Classification and Visualization	107
	<i>Yoshiaki Kurosawa, Kazuya Mera, Toshiyuki Takezawa</i>	
1.	Introduction	108
2.	Psychomimes and Emotional Spoken Dialogue Systems	109
3.	Self-Organizing Map	111
4.	Experiment	115
5.	Conclusions and Future Work	132
	References	132
6		
	Trends, Challenges and Opportunities in Spoken Dialogue Research	135
	<i>Michael McTear</i>	
1.	Introduction	135
2.	Research in Spoken Dialogue Technology	136
3.	Challenges for Researchers in Spoken Dialogue Systems	143
4.	Opportunities for Future Research in Dialogue	148
5.	Concluding Remarks	156
	References	158
7		
	Dialogue Control by POMDP using Dialogue Data Statistics	163
	<i>Yasuhiro Minami, Akira Mori, Toyomi Meguro and Ryuichiro Higashinaka and Kohji Dohsaka, Eisaku Maeda</i>	

1.	Introduction	164
2.	Partially Observable Markov Decision Process	166
3.	Dialogue Control using POMDP from Large Amounts of Data	169
4.	Evaluation and Results	177
5.	Discussion	179
6.	Future Work	181
7.	Conclusions	183
	Acknowledgments	184
	References	184
8		
	Proposal for a Practical Spoken Dialogue System Development Method	187
	<i>Masahiro Araki and Takashi Hattori</i>	
1.	Introduction	187
2.	Overview of the Data-Management Centered Prototyping Method	189
3.	Prototyping of a Slot-Filling Dialogue System	191
4.	Prototyping of a DB-Search Dialogue System	198
5.	Prototyping of a Multi-Modal Interactive Presentation System	200
6.	Incorporation of the User Model	207
7.	Conclusions	209
	Acknowledgments	210
	References	210
9		
	Quality of Experiencing Multi-Modal Interaction	213
	<i>Benjamin Weiss, Sebastian Möller, Ina Wechsung and Christine Kühnel</i>	
1.	Introduction	213
2.	Advantages of Systems Providing Multi-Modal Interaction	214
3.	Quality of Experience	216
4.	Audio-Video Quality Integration in AV-Transmission Services	217
5.	Quality of Embodied Conversational Agents	221
6.	Quality of Systems with Multiple Input Modalities	223
7.	Conclusions	226
	Acknowledgments	228
	References	228
10		
	Dialogue Acts Annotation to Construct Dialogue Systems for Consulting	231
	<i>Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, and Satoshi Nakamura</i>	
1.	Introduction	231
2.	Kyoto Tour Guide Dialogue Corpus	233
3.	Annotation of Communicative Function and Semantic Content in DA	236
4.	SA Tags	236
5.	Semantic Content Tags	245
6.	Usage of the Kyoto Tour Guide Corpus	249

7.	Conclusions	252
	References	252
11		
	On the Use of N-gram Transducers for Dialogue Annotation	255
	<i>Vicent Tamarit, Carlos-D. Martínez-Hinarejos, and José-Miguel Benedí</i>	
1.	Introduction	255
2.	The HMM-based Annotation Model	257
3.	The NGT Annotation Model	260
4.	Corpora	264
5.	Experimental Results	268
6.	Conclusions and Future Work	273
	Acknowledgments	274
	References	274
	Index	277

Contributing Authors

Nieves Ábalos is a M. S. and PhD student in Computer Science at University of Granada, Spain. She has also a B. S. in Computer Science from this University. Her research interests and activities have been related to speech technologies and include dialogue systems, multimodal systems and ambient intelligence among others. She is currently participating in several research projects related to these areas.

Masahiro Araki received B.E., M.E. and Ph. D. degrees in information science from Kyoto University, Kyoto, Japan, in 1988 and 1990, 1998 respectively. He is now an associate professor at department of information science, graduate school of science and technology, Kyoto Institute of Technology. His current interests are methodologies of spoken dialogue system development and multimodal interaction. He is a member of ISCA, ACL, IPSJ, IEICE, JSAI and ANLP.

José-Miguel Benedí received the Licenciado degree in physics from the University of Valencia in 1980 and the Ph.D. degree in computer science from the Polytechnic University of Valencia in 1989. Since 1987, he has been with the Department of Information Systems and Computation of the Polytechnic University of Valencia first as an Associate Professor and from 2002 as a Full Professor.

His current research interest lies in the areas of syntactic pattern recognition and their application to speech recognition, language modeling, language translation and understanding, and dialogue systems.

Dr. Benedí is a member of the Spanish Association for Artificial Intelligence (AEPIA) and the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), which is an affiliate society of IAPR, where he served as a member of the governing board, during the period 2001 to 2008.

Zoraida Callejas is Assistant Professor in the Department of Languages and Computer Systems at the Technical School of Computer Science and Telecommunications of the University of Granada (Spain). She completed a

PhD in Computer Science at University of Granada in 2008 and has been a visiting researcher in University of Ulster (Belfast, UK), Technical University of Liberec (Liberec, Czech Republic) and University of Trento (Trento, Italy). Her research activities have been mostly related to speech technologies and in particular to the investigation of dialogue systems. Her results have been published in several international journals and conferences. She has participated in numerous research projects, and is a member of several research associations focused on speech processing and human-computer interaction.

Ramon López-Cózar Delgado is Professor at the Faculty of Computer Science and Telecommunications of the University of Granada (Spain). His main research interests in the last 15 years include spoken and multimodal dialogue systems, focusing on speech processing and dialogue management. He has coordinated several research projects, has published a number of journal and conference papers, and has been invited speaker at several scientific events addressing these topics. In 2005 he published the book "Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment" (Wiley). Recently he has co-edited the book "Human-Centric Interfaces for Ambient Intelligence" (Elsevier Academic Press, 2010), in which he has coordinated the section concerning speech processing and dialogue management. He is a member of ISCA (International Speech Communication Association), FoLLI (Association for Logic, Language and Information), AIPO (Spanish Society on Human-Computer Interaction) and SEPLN (Spanish Society on Natural Language Processing).

Kohji Dohsaka is a Senior Research Scientist at NTT Communication Science Laboratories, NTT Corporation, Japan. He is also an invited Professor at Osaka University, Japan. He received his M.E. of Information and Computer Science degree from Osaka University, Japan, in 1986 and his Doctor of Information Science degree from Japan Advanced Institute of Science and Technology, Japan, in 2004. His research interests mainly focus on natural language communication, dialogue systems and human-computer interaction.

Gonzalo Espejo obtained the Degree in Computer Science in 2009 from the University of Granada. He is currently a PhD student in this University, where he has worked in several projects concerned with spoken and multimodal dialogue systems. His main research interests include spoken and multimodal systems as well as ambient intelligence. He has attended several workshops related to natural language processing and dialogue systems.

David Griol obtained his Ph.D. degree in Computer Science from the Technical University of València (Spain) in 2007. He has also a B.S. in Telecommunication Science from this University. He is currently professor at

the Department of Computer Science in the Carlos III University of Madrid (Spain). He has participated in several European and Spanish projects related to natural language processing and dialogue systems. His research activities are mostly related to the development of statistical methodologies for the design of spoken dialogue systems. His research interests include dialogue management/optimization/simulation, corpus-based methodologies, user modelling, adaptation and evaluation of spoken dialogue systems and machine learning approaches. Before starting his Ph.D. study, he worked as a network engineer in Motorola.

Rainer Gruhn studied computer science at the University of Erlangen-Nuremberg, Germany and graduated with Master degree in 1998. Afterwards he joined Advanced Telecommunication Research (ATR) Institute in Kyoto, Japan and received his Ph.D. degree from the University of Ulm, Germany in 2008. From 2005 to 2009 he has been with the speech division of Harman/Becker Automotive Systems and is currently in the ASR group of SVOX Deutschland GmbH. His main research interests are multilingual and non-native speech recognition.

Takashi Hattori received B.E. and M.E. in information science from Kyoto Institute of Technology, Kyoto Japan, in 2007 and 2009 respectively. He is now a member of Software Service, Inc.

Ryuichiro Higashinaka is a Research Scientist at NTT Cyber Space Laboratories, NTT Corporation. He received the B.A. degree in environmental information, the Master of Media and Governance degree, and the Ph.D degree from Keio University, Tokyo, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. From 2004 to 2006, he was a visiting researcher at the University of Sheffield, UK. His research interests are in utterance understanding and generation in spoken dialogue systems and question answering systems. He is a member of the Association for Natural Language Processing and IPSJ.

Chiori Hori received the B.E. and the M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan in 1994 and 1997, respectively. From April 1997 to March 1999, she was a Research Associate in the Faculty of Literature and Social Sciences, Yamagata University. In April 1999, she started the doctoral course in the Graduate School of Information Science and Engineering at Tokyo Institute of Technology (TITECH) and received her Ph.D. degree in March 2002. She was a Researcher in NTT Communication Science Laboratories (CS Labs) at Nippon Telegraph and Telephone Corporation (NTT), Kyoto, Japan from April 2002 to March 2004. She was a visiting researcher at Carnegie Mellon

University in Pittsburgh from April 2004 to March 2006. She was a senior researcher at ATR from January 2007 to March 2009. She is currently an expert researcher at NICT. She has received the Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2001 and Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2003. She is a member of the IEEE, the ASJ, and the IEICE.

Hideki Kashioka received his Ph.D. in Information Science from Osaka University in 1993. From 1993 to 2009, he worked for ATR Spoken Language Translation Research Laboratories. From 2006, he works for NICT. He is currently the research manager of Spoken Language Communication Group at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology. He is also the visiting associate professor of the graduate school of Information Science at the Nara Institute of Science and Technology from 1999.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Adjunct Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR, currently National Institute of Information and Communications Technology (NICT).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE SPS Speech Technical Committee. He was a general chair of the IEEE Automatic Speech Recognition & Understanding workshop (ASRU-2007). He is a senior member of IEEE.

Yoshiaki Kurosawa is a Research Associate at Hiroshima City University, Graduate School of Information Sciences, Japan. He received his Master's thesis in Human and Environmental Studies (psycholinguistics) from Kyoto University, Japan in 1994.

Christine Kühnel is working as a “Wissenschaftlicher Mitarbeiter” at the Quality and Usability Lab of Deutsche Telekom Laboratories, TU-Berlin. She studied Electrical Engineering and Business Administration in Kiel (Germany) and Madrid (Spain) and received her diploma degree in 2007 from the Christian-Albrechts University of Kiel. During her studies she completed an internship in Melbourne (Australia). At T-Labs, she is working towards her

PhD thesis in the domain of evaluation of multimodal systems.

Eisaku Maeda is an executive research scientist of Research Planning Section, NTT Communication Science Laboratories, NTT Corporation. He received the B.E. and M.E. degrees in biological science and the Ph.D. degree in mathematical engineering from the University of Tokyo, Tokyo, in 1984, 1986, and 1993, respectively. He joined NTT in 1986, and a guest researcher at the University of Cambridge, UK in 1996-1997. From 2002 He is also a invited professor of Osaka University, Osaka.

Carlos D. Martínez-Hinarejos is a Senior Lecturer (Profesor Contratado Doctor) at the Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación, Spain. He got his degree in Computer Science (1998) and his Ph.D. in Pattern Recognition and Artificial Intelligence (2003) from the same university. From 1999 to 2000 he got a grant from the Spanish Ministry of Education, and since 2000 he has been a member of the University staff in several positions. He did a postdoctoral visit to the University of Sheffield, Natural Language Processing group during 2005. His main research interests cover dialogue systems, speech recognition, dialogue annotation and bioinformatics.

Michael McTear is Professor of Knowledge Engineering at the University of Ulster with a special research interest in spoken language technologies. He graduated in German Language and Literature from Queens University Belfast in 1965, was awarded MA in Linguistics at University of Essex in 1975, and a PhD at the University of Ulster in 1981. He has been Visiting Professor at the University of Hawaii (1986-87), the University of Koblenz, Germany (1994-95), and University of Granada, Spain (2006, 2007, 2008). He has been researching in the field of spoken dialogue systems for more than fifteen years and is the author of the widely used text book "Spoken dialogue technology: toward the conversational user interface" (Springer Verlag, 2004).

Toyomi Meguro is a Research Scientist at NTT Communication Science Laboratories, NTT Corporation. She received the B.E. and M.E. degrees in electric engineering from Tohoku University, Japan, in 2006 and 2008, respectively. She joined NTT in 2008. Her research interests are in building listening agents.

Kazuya Mera is a research associate at Hiroshima City University, Graduate School of Information Sciences, Japan. He received his PhD from Tokyo Metropolitan Institute of Technology, Japan in 2003.

Florian Metze has worked on a wide range of problems in speech processing and applications. He is one of the developers of the dynamic "Ibis" decoder for the JRTk speech recognition toolkit, has worked on acoustic modeling for

distant microphones and conversational speech, particularly using discriminatively trained combinations of articulatory features. Extracting information such as age, gender, or emotion of a speaker and using it in adaptive speech applications was a recent focus of his work. He has also worked on the evaluation of multi-modal speech dialog systems from a user perspective, and the corresponding usage of modalities for mobile devices. The use of speech in interactive applications for accessing knowledge was also part of his work.

Yasuhiro Minami is a Senior Research Scientist at NTT Communication Science Laboratories, NTT Corporation. He received the M. Eng. Degree in Electrical Engineering and the Ph. D. in Electrical Engineering from the Keio University, in 1988 and 1991, respectively. He joined NTT in 1991. He had worked in robust speech recognition. He was a visiting researcher at MIT from 1999 to 2000. Since February, 2000, he has been with NTT Communication Science Laboratories. He is interested in modeling for speech recognition and speech dialogue systems.

Teruhisa Misu received the B.E. degree in 2003, the M.E. degree in 2005, and the Ph.D. degree in 2008, all in information science, from Kyoto University, Kyoto, Japan. From 2005 to 2008, he was a Research Fellow (DC1) of the Japan Society for the Promotion of Science (JSPS). In 2008, he joined NICT Spoken Language Communication Group.

Akira Mori is a Senior Research Scientist at NTT Communication Science Laboratories NTT Corporation. He received the B.E. and M.E. degrees in electronic engineering from Tohoku University in 1986 and 1988 respectively. He joined NTT in 1988. He worked on LSI designing, and broadband business development. He has been studying robotics and sensor fusion for interaction mechanism between human and robots at NTT Communication Science Laboratories.

Sebastian Möller studied electrical engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). He received a Doctor-of-Engineering degree at Ruhr-University Bochum in 1999 for his work on the assessment and prediction of speech quality in telecommunications. In 2000, he was a guest scientist at the Institut dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) in Martigny (Switzerland) where he worked on the quality of speech recognition systems. He gained the qualification needed to be a professor (*venia legendi*) at the Faculty of Electrical Engineering and Information Technology at Ruhr-University Bochum in 2004, with a book on the quality of telephone-based spoken dialogue systems. Since June 2005, he works at Deutsche Telekom Laboratories, TU Berlin. He was appointed Professor at TU Berlin for the subject "Usability" in April 2007, and heads the

“Quality and Usability Lab” at Deutsche Telekom Laboratories.

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was an associate professor of the graduate school of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008. He is ATR Fellow. He launched the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007. He is currently Vice Executive Director of Knowledge Creating Communication Research Center, and Director of MASTAR project, National Institute of Information and Communications Technology, Japan. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, and Docomo Mobile Science Award in 2007, Telecom System Award, and ASJ Award for Distinguished Achievements in Acoustics. He served as a general chair of International Workshop of Spoken Language Translation (IWSLT2006) and Oriental Cocosda 2008. He will be serving as Program Chair of INTERSPEECH 2010.

Elmar Nöth is a professor for Applied Computer Science at the University of Erlangen-Nuremberg. He studied in Erlangen and at M.I.T. and received the Dipl.-Inf. (univ.) degree and the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 1985 and 1990, respectively. Since 1990 he was an assistant professor at the Institute for Pattern Recognition in Erlangen. Since 2008 he is a full professor at the same institute and head of the speech group. He is one of the founders of the Sympalog Company, which markets conversational dialogue systems. He is on the editorial board of Speech Communication and EURASIP Journal on Audio, Speech, and Music Processing and member of the IEEE Speech and Language Technical Committee (SLTC). His current interests are prosody, analysis of pathologic speech, computer aided language learning and emotion analysis.

Kiyonori Ohtake is an Expert Researcher at National Institute of Information and Communications Technology, Spoken Language Communication Group of MASTAR Project, where he has worked since 2006. He received his Dr. of Eng. in 2001 from Toyohashi University of Technology. In 2001, he joined Spoken Language Communication research laboratories at Advanced Telecommunication Research institute international (ATR). His research focuses on dependency parsing of spoken Japanese, paraphrasing for language understanding, and proactive dialogue system.

Tim Polzehl studied Science of Communication at Berlin’s Technical University and American Science at Humboldt University of Berlin. Combining

linguistic knowledge with signal processing skills he focussed on speech interpretation and automatic data- and metadata extraction. He gathered experience within the field of machine learning as exercised when recognizing human speech utterances and classifying emotional expression subliminal in speech, the latter of which became his M.A. thesis. Tim Polzehl is a research scientist at Deutsche Telekom Laboratories (T-Labs) at present. Among other works and publications, he recently developed an emotion recognition system that was successfully contributed to the international Emotion Challenge Benchmark, held at Interspeech 2009.

Martin Raab has studied computer sciences at Carnegie Mellon University, USA, Trinity College Dublin, Ireland and the University of Karlsruhe, Germany where he received his Master degree in 2006. From 2007 to 2010 he was a speech recognition researcher at Harman Becker Automotive Systems and Nuance Communications in Ulm, Germany. The academic supervision during that time was done by Prof. Dr. Elmar Nöth from the Friedrich-Alexander-University of Erlangen-Nuremberg. Currently he is a software engineer at CAS Software AG in Karlsruhe, Germany.

Alexander Schmitt studied Computer Science in Ulm (Germany) and Marseille (France) with focus on media psychology and spoken language dialogue systems. He received his Masters degree in 2006 when graduating on Distributed Speech Recognition on mobile phones at Ulm University. Schmitt works as research assistant at Ulm University and is currently carrying out his PhD research project under the supervision of Prof. Dr. Wolfgang Minker. His research is centered around the statistical detection of problematic phone calls in Interactive Voice Response Systems and is carried out in cooperation with SpeechCycle, NYC, USA. Schmitt worked for Daimler Research Ulm on Statistical Language Modeling and regularly gives lectures on the development of Voice User Interfaces.

Komei Sugiura is an expert researcher at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology (NICT). He received his B.E. degree in electrical and electronic engineering, M.S. and Ph. D. degrees in informatics from Kyoto University in 2002, 2004, and 2007, respectively. From 2006 to 2008, he was a research fellow, Japan Society for the Promotion of Science, and he has been with NICT since 2008. His research interests include robot language acquisition, spoken dialogue systems, machine learning, and sensor evolution. He is a member of the Society of Instrument and Control Engineers, and the Robotics Society of Japan.

Toshiyuki Takezawa is a Professor at Hiroshima City University, Graduate School of Information Sciences, Japan. He received his PhD in Engineering from Waseda University, Japan in 1989. He was a Researcher at Advanced Telecommunications Research Institute International (ATR), Japan, from 1989 to 2007.

Vicent Tamarit got his degree in Computer Science in 2006 at the Universidad Politénica de Valencia. In 2008 got his master degree in Artificial Intelligence, Pattern Recognition and Digital Image by the Departamento de Sistemas Informáticos y Computación. Currently, he has a scholarship by the Ministry of Science of the Spanish Government. He develops his scholarship at the Instituto Tecnológico de Informática, in Valencia. His research interests are speech recognition and dialogue systems.

Ina Wechsung is working as a research assistant (Wissenschaftlicher Mitarbeiter) at the Quality and Usability Lab of Deutsche Telekom Laboratories, TU-Berlin. She studied Psychology and received her diploma degree in 2006 from the Chemnitz University of Technology. At T-Labs, she is working towards her PhD thesis.

Benjamin Weiss joined the Quality and Usability Lab in January 2007. He studied communication science & phonetics, educational studies and Scandinavian studies at the universities of Bonn, Trondheim and Berlin. After his graduation in 2002, he was with a “Graduiertenkolleg” at the Linguistics department at Humboldt University Berlin, doing his dissertation about speech tempo and pronunciation. He received his Ph.D. in Linguistics in 2008. Currently, he is working on evaluating quality and usability of multimodal interfaces.

Chapter 1

MULTILINGUAL SPEECH INTERFACES FOR RESOURCE-CONSTRAINED DIALOGUE SYSTEMS

Martin Raab

*Nuance Communications, Speech Technologies
Ulm, Germany*

martin.raab@informatik.uni-erlangen.de

Rainer Gruhn

*SVOX Deutschland GmbH
Ulm, Germany*

rainer.gruhn@alumni.uni-ulm.de

Elmar Nöth

*University of Erlangen-Nuremberg, Chair of Pattern Recognition
Erlangen, Germany*

noeth@informatik.uni-erlangen.de

Abstract A key component for the successful provision of spoken dialogue systems is speech recognition. The capabilities of the applied speech recognizer influence many other design criteria for spoken dialogue systems. In this chapter, the problem of multilingual speech for speech recognizers is analyzed and a solution for recognizing many languages simultaneously is proposed. Two main design goals for the described system were to keep the recognition feasible on embedded systems and to handle the native speech of the user with highest priority. In addition, experiments are described that address the effect of non-native accent. With the help of the added multilinguality, existing in-car infotainment systems can be extended to international navigation input and music selection via voice.

Keywords: Speech recognition; Non-native; Embedded.

1. Introduction

It is a frequent problem that spoken dialogue systems have to handle speech from multiple languages. One example are speech operated navigation systems that should allow destination input for many countries. In the near future companies are also preparing to launch music players that are operated via speech. A major problem in these applications is that the dialogue systems operate with constrained resources, and current speech recognition technology typically requires more resources for each language that has to be covered.

As this chapter focuses mainly on speech recognition, it is necessary to understand what components are involved in a speech recognition system, and how it is linked to the spoken dialogue system. Therefore [Figure 1](#) depicts a semi-continuous speech recognition system as it is applied in the experiments.

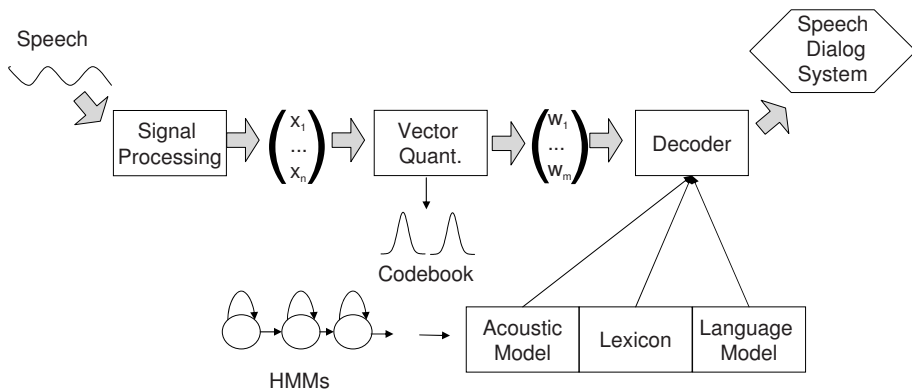


Figure 1. Overview of a semi-continuous speech recognition system. The incoming speech signal is first reduced to a feature vector that contains the most relevant information. This vector is matched against Gaussians in the codebook, and the corresponding probabilities are passed to the decoder that considers three knowledge sources to find the most likely speech utterance that it passes to the speech dialogue system.

The backbone of the whole process is the pattern matching with Hidden Markov Models (HMMs) and Gaussians in the acoustic model. In our special case, we work with one codebook, i.e. with one set of Gaussians that is shared across all HMMs. This allows to evaluate the Gaussians separately in the vector quantization step, and keeps the total number of Gaussians limited, which is important for resource constrained systems. In this process the Gaussians are responsible for the judgment of individual feature vectors, and the HMM structure models the change of the feature vectors over time. The other components like the lexicon and the language model represent additional knowledge sources that can help to retrieve correct decisions in case of almost equal hy-

potheses. For further details of speech recognition systems, see (Huang et al., 2001).

The rest of this chapter is organized as follows. As we discuss quite a lot of literature from the field of speech recognition we decided to have a separate Section 2 for our literature review. Based on the findings from the literature, Section 3 gives a brief overview of the experiments and approaches in this work. Section 4 then presents the experimental setup. Section 5 and 6 report experiments. The first one focuses on accent adaptation and the second one tests the combination of the most promising adaptation technique with a new method for the generation of multilingual acoustic models. Finally, Section 7 summarizes the findings from this chapter.

2. Literature Review

2.1 Review of Multilingual Speech Recognition

A common method for multilingual speech recognition is to model similar sounds in different languages with the same models. In this approach, phonemes from different languages can share one acoustic model when they have the same IPA (International Phonetic Alphabet, (Ladefoged, 1990)) symbol. Examples are (Koehler, 2001; Schultz and Waibel, 2001; Niesler, 2006). The sharing can also be based on acoustic model similarity determined by a distance measure. For example, Koehler (2001) and P. Dalsgaard et al. (1998) measure the log-likelihood difference on development data to determine the similarity of phonemes, as motivated by Juang and Rabiner (1985).

The advantages of all these approaches are that they cover many languages with much less parameters than a combination of all the monolingual recognizers. Thus they are very appropriate in all cases where one really needs all languages equally. However, in the examples mentioned at the beginning, the languages are not of equal importance. There is one device, which is typically owned by one user with one native language and that language is more important for the system than the other languages as the user usually utters commands, spellings and digit sequences in that language. Hence it is vital for a commercial system to recognize this main language with maximum performance.

Motivated by the finding that parameter sharing achieves better performance than sharing of full Hidden Markov Models (HMM) (Koehler, 2001), we developed the Multilingual Weighted Codebooks (MWC) algorithm for semi-continuous HMMs (Raab et al., 2008a; Raab et al., 2008b). The advantage of semi-continuous HMMs is that all HMMs use one codebook with a small number of Gaussians compared to the number of Gaussians that are used in continuous HMMs. However, the performance of a semi-continuous HMM with a suboptimal codebook from a foreign language is significantly reduced (Raab

et al., 2008b). The MWCs solve this problem by adding the most different Gaussians from the foreign language codebooks to the main language codebook.

2.2 Review of Non-Native Speech Recognition

An additional problem in this scenario is that the users are not perfect in the pronunciation of the foreign city names or the foreign song titles. They utter the utterances with non-native accent. This accent is amongst others influenced by the native language of the user (the main language of the system).

In the second part of our literature review we consider the work that has been done on non-native speech recognition. Tomokiyo and Waibel (2001) present several results with different adaptation techniques and achieve up to 30% rel. Word Error Rate (WER) improvement. Bouselmi et al. (2007) introduce confusion based acoustic model integration that allows additional HMM structures for frequently confused phoneme models. They report improvements of up to 70% rel. WER and an absolute Word Accuracy (WA) of up to 98.0% without speaker adaptation on the Hiwire test data (Segura et al., 2007) that we also use. However, using the Hiwire data for adaptation and testing is likely to give good results as the lexicon size is very limited and the same speakers are in the adaptation and test set (Lang, 2009).

There are also many works that seem to indicate that solely modifying the lexicon cannot achieve comparable improvements as acoustic model adaptation. For example, Gruhn et al. (2004) achieve about 11% relative WER improvement with rescoring with statistical pronunciation HMMs for many language pairs. Amdal et al. (2000) get only 3% relative improvement with rules extracted from non-native phoneme recognition results. However, Kim et al. (2007) obtain up to 19% relative WER improvement for Korean by Chinese speakers with rules derived via a decision tree.

The discussed adaptation methods have the drawback that they need development data from the corresponding accents. An overview of existing databases can be found in (Raab et al., 2007). The biggest database known to the authors covers almost 30 different accents (Schaden, 2006), but there are a lot more accents that are not covered. Therefore other techniques try to circumvent the need for special training or adaptation data. For example, Bartkova and Juvet (2006) and Goronzy et al. (2001) use manually derived pronunciation rules for the modification of lexicons. However, their approaches require expensive human work and achieve only moderate improvements in the range of 15% to 30% rel. WER.

Special training data can also be avoided if the required information about a non-native accent is derived from a comparison of the native language of the speaker and the spoken language. Witt (1999) proposes three different

algorithms for this, amongst other model merging. Improvements of up to 27% rel. WER are reported for the methods without online adaptation. However, the work of Witt was performed on continuous HMMs and cannot directly be applied to a semi-continuous HMM. Witt's algorithms do also benefit from adding Gaussians from other languages, so there is the question to what extent for example model merging can add on top of MWCs. The same question arises with work from (Tan and Besacier, 2007a).

3. Approach

Based on the findings in the first part of our literature review, we decided to use MWCs as this data driven model sharing technique has the advantage that it does not degrade the performance on the additional languages. The second part of the literature review showed that many works present very good improvements with non-native data. However, often the results yield the advantage that data from the same speakers is used for adaptation and for testing. In one of our experiments we therefore analyze how the effects change if we remove this additional adaptation benefit. In the the same set of experiments, we analyze the effect of different amounts of training data for context dependent and context independent modeling for non-native speech. We also verify that model merging helps for semi-continuous systems. All these experiments will show that MWCs are an appropriate technique for the efficient modeling of non-native speech.

Finally, we treat one of the major drawbacks of the MWC approach, the increased training effort. While from a purely theoretical point of view such a problem can be neglected, an exponential dependency on the number of languages supported will probably prevent every algorithm to be applied in real world commercial products for more than a couple of languages. We therefore propose projections between Gaussian spaces as an alternative to time consuming retraining, and together with the MWC algorithm this forms a scalable architecture that can be tailored according to the current user needs. Regarding this last aspect, different projections were proposed by (Raab et al., 2009b) and (Raab et al., 2009a). In this chapter, we show for the first time the effects of the combination of the MWC approach and the idea of the projections between Gaussian spaces.

4. Experimental Setup

4.1 Training and Test Data

Our semi-continuous speech recognizer uses 11 MFCCs with their first and second derivatives per frame. Monolingual recognizers for English, French, German, Spanish and Italian are each trained on 200 hours of Speecon

data (Iskra et al., 2002) with 1024 Gaussians in the codebook. The HMMs are context dependent and the codebook for each language is different.

Table 1 describes the native test sets and Table 2 the non-native test sets. All results will be reported in Word Accuracy (WA). The native tests are city names from a Nuance internal database.

Table 1. Descriptions of the native test set for each language.

Testset	Language	Words	Vocabulary
GE_City	German	2005	2498
US_City	English	852	500
IT_City	Italian	2000	2000
FR_City	French	3308	2000
SP_City	Spanish	5143	3672

Table 2. Description of the non-native test sets.

Testset	Accent	Words	Vocabulary
Hiwire_IT	Italian	3482	140
Hiwire_FR	French	5192	140
Hiwire_SP	Spanish	1759	140
Hiwire_GR	Greek	3526	140
IFS_MP3	German	831	63

The first four non-native test sets contain command and control utterances in accented English from the Hiwire database (Segura et al., 2007). The Hiwire database was identified as the most appropriate existing and available database after a review of existing non-native databases (Raab et al., 2007). To indicate that the language information in the test name only specifies the accent, the language information is given after the test name. The last non-native accent test was collected specifically for this work and contains Italian, French and Spanish song titles names spoken by Germans. A key difference for this last test is that the speakers sometimes had to pronounce utterances from a language that they cannot speak.

For some experiments development data was applied. Utterances from two different databases are used for adaptation. The development sets for Hiwire are very similar to the Hiwire test sets and as specified in the Hiwire database (Segura et al., 2007). The Hiwire development sets are non-overlapping with the Hiwire test sets, but they have the same speakers and cover consequently the same accents as the Hiwire tests.

However, as most of the experiments are tested on data from the Hiwire corpus, the question was if the observed improvements are only due to adaptation to non-native speech, or if other issues like adaptation to the recording conditions did also improve the performance. Therefore 80% of the speech from the ISLE Corpus (Interactive Spoken Language Education, (Menzel et al., 2000)) constitutes a second development set. Altogether, the ISLE development set contained 4446 utterances of German speakers and 4207 utterances of Italian speakers. The utterances style (natural sentences) is different to Hiwire, but this is less relevant, as the algorithms focus only on the acoustic realization of phonemes.

4.2 Benchmark System

The starting point for our comparison systems are trained monolingual semi-continuous HMM speech recognizers. This means that we have trained triphone models for all languages.

The benchmark system for the recognition of multiple languages combines all triphone models in one large model set. This is nothing else than evaluating all monolingual recognizers in parallel. Thus this system can achieve monolingual performance in all languages. However, in this approach all Gaussians from all languages that are currently set active for recognition have to be evaluated. This violates the motivation for the use of a semi-continuous system as no longer only one fixed number of Gaussians has to be evaluated for all models. To summarize, this approach can be considered as an upper bound in performance, but requires a linear increase of resources on the embedded system with the number of considered languages.

5. Accent Adaptation

5.1 Monophones vs. Triphones

In our first set of experiments we analyze the effect of more native training data for non-native speech recognition. On the one hand, it could be that the sharp modeling of context dependencies will not help as non-native speakers are not able to adhere to these subtle rules, as indicated in experiments in (Tan and Besacier, 2007b). On the other hand, experiments in (He and Zhao, 2001) showed that significant benefits can be obtained with context dependent models. Of course, such observations depend on the fluency of the speakers tested.

[Figure 2](#) evaluates these issues for one native English city test and four non-native tests. The lines show a couple of interesting aspects. First, as expected, for native speech context dependent modeling is much better than context independent modeling in all cases. Second, the non-native tests also benefit from more native training data, and third, when a large amount of data is applied

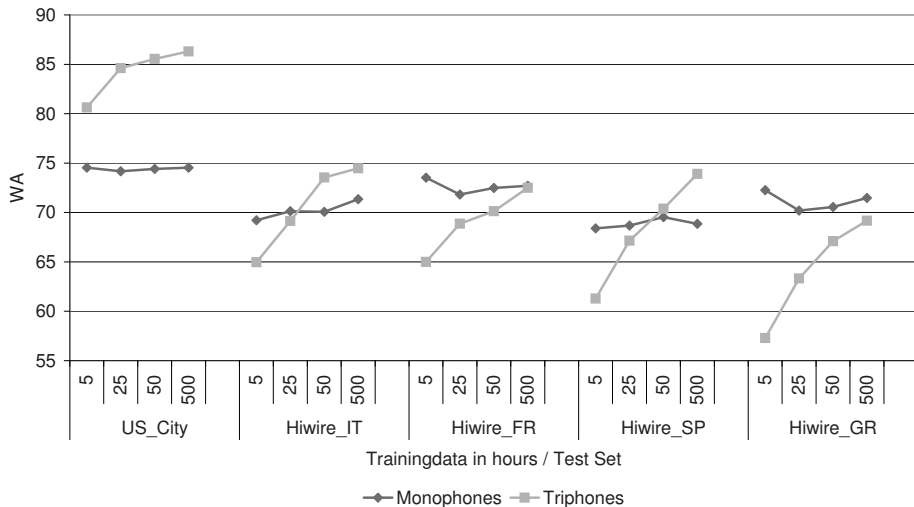


Figure 2. Comparison of monophones vs. triphones with different amounts of training data for non-native speech. In all tests more data helps more for the context dependent triphones. The results also show that while the difference is not as high as for the native city test, the performance on non-native speech with context dependent models is in average better than the context independent models if a large amount of training data is used.

(500 hours¹), the performance of context dependent modeling is clearly better for Spanish and Italian accented speech, and similar to context independent modeling for the other two non-native tests.

Due to these results, we applied context dependent models in the following experiments to have the potential to benefit from more native training, which is today accumulating at large companies as never before.

5.2 Multilingual MWC System

In the Multilingual Weighted Codebook (MWC) system the goals are to recognize many languages with limited decoding resources and to have monolingual performance for the native language of the user.

Figure 3 shows that our MWC systems are basically build like monolingual systems. From left to right, the following happens. First, one large HMM model set is generated that contains all HMM models of both languages. All these HMMs are untrained at this stage. Then, the HMMs that belong to one language are trained with speech from that language. This happens iteratively till all HMMs are trained. The only difference to training monolingual recognizers is the MWC itself.

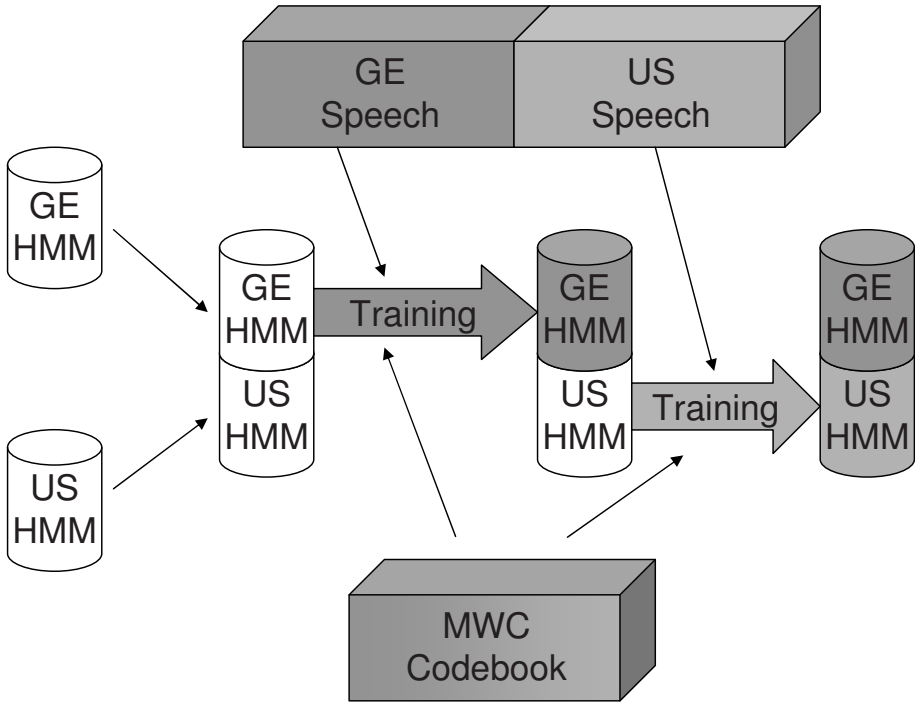


Figure 3. Generation of the multilingual system at the example of a bilingual German/English system. The HMMs for both languages are trained as usual, the German ones with German speech, and the English ones with English speech. The only difference to a monolingual system is that a multilingual codebook is applied.

The basic idea of an MWC is that it contains all Gaussians from the main language of the speaker and the most different Gaussians from the other modeled languages. The complete inclusion of the main language Gaussians guarantees optimal performance for the main language of the speaker. The additional Gaussians ensure that the errors made by the vector quantization step are kept at a small magnitude, and due to the fact that many Gaussians are not added, that the overall size of an MWC is much smaller than a simple pool of all Gaussians from all languages. More details about the MWC generation can be found in (Raab et al., 2008a) and (Raab et al., 2008b).

There are two aspects of interest regarding the MWC system. How well can it recognize very fluent speakers, and how does it perform for stronger accented speakers.

To analyze the first aspect an MWC system based on a German codebook is tested on native city names from several languages. [Figure 4](#) shows the strong benefits for the modeling of foreign sounds if additional Gaussians are allowed.

In fact, the baseline system is in this case an acoustic model that only applies Gaussians from the German codebook. The benchmark system is as described in Section 2 the monolingual system for each language and thus needs 5120 Gaussians for the recognition of all languages.

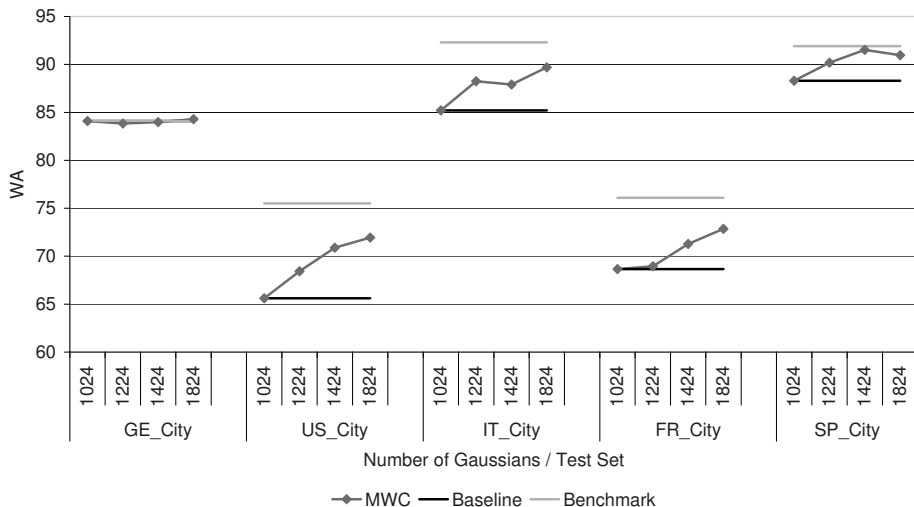


Figure 4. MWC performance starting from a German codebook with 1024 Gaussians for five languages. The x-axis indicates the total number of Gaussians, 1424 means for example that 400 Gaussians were added from the 4096 Gaussians of the other four languages.

Two more comments about the results in Figure 4 might be helpful. First, the different behavior of the German test is reasonable, as all system always have all Gaussians that stem from the German codebook. Second, the WA differences between the languages are due to different noise conditions in the different tests and are less relevant here.

The second question is how the performance difference between the MWC system and the benchmark system changes for average non-native speakers. This question is analyzed in Figure 5. The results demonstrate that the replacement of the codebook of the spoken language by the native codebook is actually an efficient method for modeling the accent of the speakers as long as the speakers are familiar with the correct pronunciation of the utterances. This is the case for the three Hiwire tests. For them, the improvement can be up to 25% rel. WER.

However, Figure 5 also clearly shows that in the case when speakers do not know the origin of the song title and/or cannot speak the corresponding language no improvements are observed. One reason for this is probably that in this work only the phoneme sequence of the correct language is allowed. At this point, further work is required, to evaluate how misjudged language origins

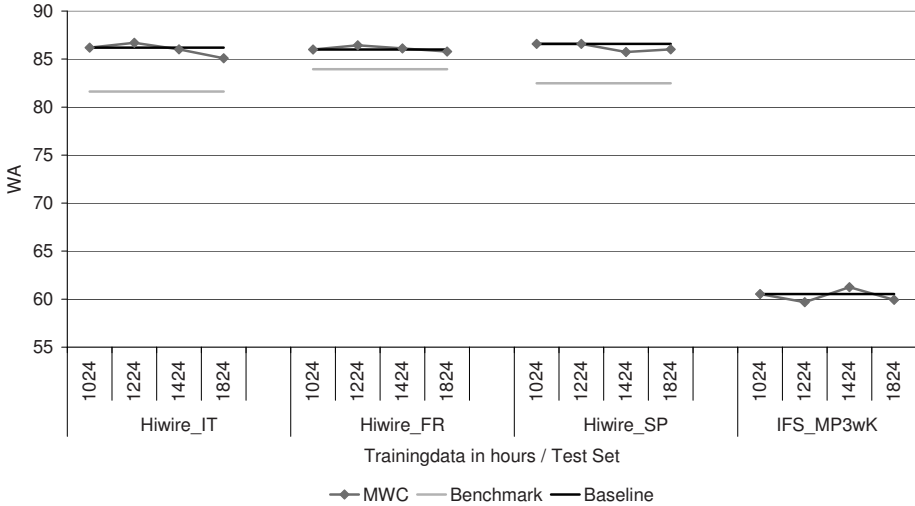


Figure 5. MWCs on non-native speech. The main language is always the native language of the current speaker group. The x-axis indicates how many Gaussians have been added from the four other languages.

can be handled. For example, acoustic language detection could help here to add the phoneme sequences of the language that the system has detected. The rest of this work concentrates on the behaviour in the case when the speakers know the correct pronunciation of the utterances they have to speak.

Another observation for average fluent speakers is that the addition of Gaussians does not help. However, it also does not reduce the performance, and together with the results from Figure 4 this means that an MWC system with 400 or 800 additional Gaussians has good performance for almost fluent and less fluent non-native speakers.

5.3 Model Merging

Model merging is a technique that was first proposed by (Witt and Young, 1999). In their work, improvements of up to 27% WER were achieved without additional adaptation data. The basic idea is to find similar sounds in the spoken language and the native language of the speaker and to create a new HMM that contains Gaussians from both languages. Thus it is to some extent similar to the MWC approach, however, there is also the additional factor that the mixture weights of the Gaussians are modified to keep the constraint that all mixture weights sum up to one.

In this work, the original approach had to be modified in several ways. First, adding all additional Gaussians is not an option, as this is basically just our

benchmark system and not desirable in resource constrained systems. Second, their method for determining if a sound is similar was also very expensive and included manual checks by humans. In our commercial scenario with hundreds of targeted accents, this should be avoided. We therefore compared the expected value of the feature vector for each HMM state and determined similarity by a Mahalanobis distance between these expected values. This procedure is described in more detail in the projection section in this chapter. Finally, to avoid the addition of Gaussians we worked only with acoustic models that are trained on the same codebook, thus there were no additional Gaussians, and model merging reduced to a weighted combination of mixture weights for each state, meaning

$$c_i = \alpha a_i + (1 - \alpha) b_i,$$

where c_i is the mixture weight for Gaussian i of the generated HMM and a_i and b_i are the mixture weights for the same Gaussian in a similar sound in the spoken and in the native language. α determines the influence of the native language of the speaker, and can vary between 0 and 1.

In a first experiment, we started from the native English system. In this case, as in the mentioned paper, significant improvements were achieved, even without additional Gaussians. The results are shown in Table 3 for three successful weight settings of alpha. However, the same table also shows that replacing the codebook yields better performance. We also tried to combine the two effects, and tested model merging on the native language codebooks, however no additional improvements were observed.

Table 3. Performance of model merging. Line 1 shows the performance of our monolingual English system. Line 2-4 show the performance of systems with an English codebook that are merged with models from the native language of the speakers with the specified value of alpha. Finally, Line 5-7 show in comparison the results when the codebook is replaced by the native language codebook.

Codebook	alpha	Hiwire_IT	Hiwire_FR	Hiwire_SP
English	0	81.6	83.9	82.5
English	0.1	82.2	84.8	84.1
English	0.2	83.8	85.0	84.4
English	0.3	83.4	84.9	84.9
English	0.4	83.1	84.7	84.8
Italian	0	86.2	-	-
French	0	-	86.0	-
Spanish	0	-	-	86.6

Our conclusions from these results are that model merging can also be applied to semi-continuous HMM systems to achieve improvements of up to 14%

rel. WER. The improvements are less pronounced as in the original work, as model merging in our case does not include the addition of relevant Gaussians. In fact, our results show that a replacement of the Gaussians alone is sufficient, and a combination of both techniques gives no additional improvements. However, although this is a reasonable decision for the scenario described in this work, in a scenario that only has to recognize English speech, model merging remains an appropriate technique to model different accents of English efficiently.

5.4 Adaptation with Non-Native Speech

A common method to improve the modeling of non-native speech is also to use some non-native adaptation data. In our scenario this is, as stated in the introduction, not very desirable, as hundreds of accents should be supported and therefore many data collections would be needed. Nevertheless, it is an interesting question how much can be gained with appropriate training data, and how the performance changes if data from other speakers and other databases is applied.

In the first experiment, our monolingual English system is retrained with non-native speech from the Hiwire database. [Figure 6](#) shows as expected that retraining is in all cases significantly better than the native system alone. These tests are again performed both for context dependent and context independent models. Among others the results show again that well trained triphones tend to outperform monophones. However, the major observation in this experiment is that improvements of more than 80% rel. WER are achieved.

Yet, these improvements might be hard to replicate in real scenarios as in many situations users will not like to have to train their system for one or two hours before they can use it, and the training data may also stem from another recording scenario as the actual test data. To evaluate these aspects, a second experiment was performed. This time development data from the ISLE corpus is applied. However, the tests remain from the Hiwire corpus.

[Figure 7](#) shows the significant impact this change yields for the recognition performance. The different adaptation data, although it is actually 2-3 times more adaptation data than the Hiwire data, does not give comparable improvements as the adaptation data from the same speakers. In fact, the context dependent models become worse. Nevertheless, the monophone system still outperforms the native benchmark system by 33% rel. WER.

However, due to the high development costs as well as the corresponding follow up costs (systems have to be built for all accents, have to be provided to customers and maintained) the accent adaptation described in this section is probably out of scope for current spoken dialogue systems. Some of these arguments are actually also preventing large scale application of the previously

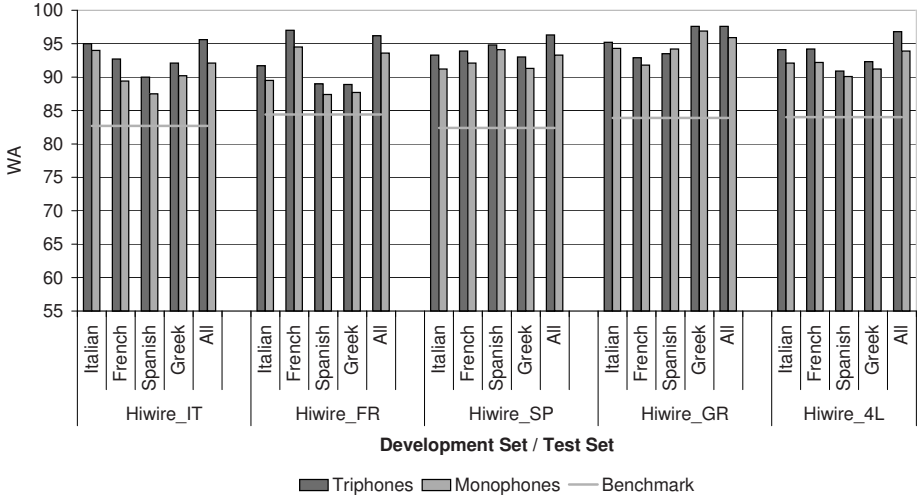


Figure 6. Retraining and testing on Hiwire. The vertical languages at the x-axis indicate which of the Hiwire development sets was applied. The results are reported for all four accents of the Hiwire database individually and for all accents together.

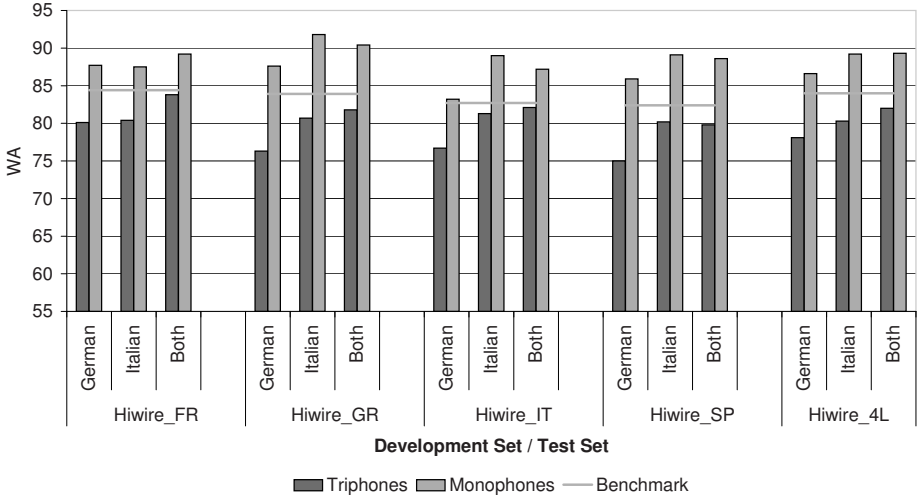


Figure 7. Retraining on ISLE and testing on Hiwire.

mentioned approaches. However, the following section will show how these problems can be at least partially circumvented for MWC systems, for example.

6. Scalable Architecture

In the previous section it was shown that there are a couple of ways to improve the modeling of non-native speech. For our scenario of one user of the system who is mainly interacting in his native language, but sometimes has to use words from foreign languages, the MWC approach is promising, as it does not require additional data collections and performs very well for the native language of the speaker.

Nevertheless, the MWC has to some extent the same problems as the other approaches, for example, specific systems have to be built and maintained for every accent. However, the actual creation of the codebook is less the problem, it is more the fact that only once the codebook is determined the HMMs that model the sounds can be estimated. In this section we will show that this issue can be circumvented with projections between Gaussian Mixture Models (GMMs).

6.1 Projections between GMMs

As discussed, the main problem with the provision of MWC systems for many accents is that the number of acoustic models that need to be trained grows exponentially with the number of languages supported. A training of a speech recognizer still takes a couple of hours on a current desktop computer, and thus prevents the application of this technique for many languages. However, the process is also very redundant, as HMMs for the same sound have to be estimated over and over again. This is depicted in Figure 8. Figure 8 also indicates that this problem can be avoided if there is a way to project the estimated distribution of one sound to another set of Gaussians. In this case, no retraining would need to be performed. Our first approach to this task was to minimize a distance between two Gaussian mixture models.

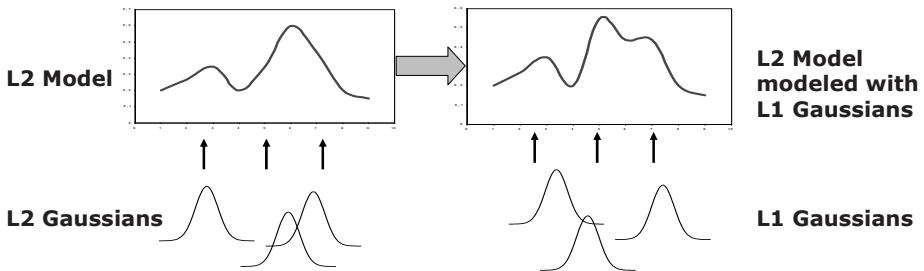


Figure 8. Motivation for the projection algorithm. The left L2 model and L2 Gaussians are from a monolingual acoustic model. In a multilingual system, different Gaussians (L1 Gaussians) have to be used. The idea is to project the L2 model to the L1 Gaussians to have a representation of the model that does not need additional Gaussians.

In the literature, many distances or divergences between GMMs are proposed, like the Kullback Leibler divergence, the likelihood on a development set (Juang and Rabiner, 1985) or the L2 distance (Jian and Vemuri, 2005). We chose the L2 distance, as it is the only one of them that can be solved analytically. The L2 distance between two Gaussian mixture models A and B is defined as follows:

$$D_{L2}(A, B) = \int (\alpha^T \mathbf{a}(\mathbf{x}) - \beta^T \mathbf{b}(\mathbf{x}))^2 d\mathbf{x}, \quad (1.1)$$

α and β are the weight vectors of the Gaussian vectors $\mathbf{a}$ and $\mathbf{b}$.

$$\alpha = \begin{pmatrix} w_1^a \\ w_2^a \\ \vdots \\ w_n^a \end{pmatrix}, \quad \mathbf{a}(\mathbf{x}) = \begin{pmatrix} \mathcal{N}(\mathbf{x}; \mu_1^a, \Sigma_1^a) \\ \mathcal{N}(\mathbf{x}; \mu_2^a, \Sigma_2^a) \\ \vdots \\ \mathcal{N}(\mathbf{x}; \mu_n^a, \Sigma_n^a) \end{pmatrix}, \quad (1.2)$$

$$\beta = \begin{pmatrix} w_1^b \\ w_2^b \\ \vdots \\ w_m^b \end{pmatrix}, \quad \mathbf{b}(\mathbf{x}) = \begin{pmatrix} \mathcal{N}(\mathbf{x}; \mu_1^b, \Sigma_1^b) \\ \mathcal{N}(\mathbf{x}; \mu_2^b, \Sigma_2^b) \\ \vdots \\ \mathcal{N}(\mathbf{x}; \mu_m^b, \Sigma_m^b) \end{pmatrix}, \quad (1.3)$$

where μ and Σ are the mean and the covariances of the Gaussians. The distance D_{L2} can be calculated as follows

$$\begin{aligned} D_{L2}(A, B) &= \int (\alpha^T \mathbf{a}(\mathbf{x}) - \beta^T \mathbf{b}(\mathbf{x}))^2 d\mathbf{x} \\ &= \sum_i \sum_j \alpha_i \alpha_j \int a_i(\mathbf{x}) a_j(\mathbf{x}) d\mathbf{x} \\ &\quad - 2 \sum_i \sum_j \alpha_i \beta_j \int a_i(\mathbf{x}) b_j(\mathbf{x}) d\mathbf{x} \\ &\quad + \sum_i \sum_j \beta_i \beta_j \int b_i(\mathbf{x}) b_j(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (1.4)$$

with $a_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_i^a, \Sigma_i^a)$ and $b_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_i^b, \Sigma_i^b)$. In order to solve this problem, the correlation $\int \mathcal{N}(\mathbf{x}; \mu_1, \Sigma_1) \mathcal{N}(\mathbf{x}; \mu_2, \Sigma_2) d\mathbf{x}$ between the Gaussians needs to be calculated. Petersen and Pedersen (2008) state that

$$\mathcal{N}(\mathbf{x}; \mu_1, \Sigma_1) \mathcal{N}(\mathbf{x}; \mu_2, \Sigma_2) = c_c \mathcal{N}(\mathbf{x}; \mu_c, \Sigma_c). \quad (1.5)$$

The elements of the resulting Gaussian $c_c \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ are

$$\begin{aligned} c_c &= \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)) \\ &= \frac{1}{\sqrt{\det(2\pi(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2))}} e^{[-1/2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]}, \end{aligned} \quad (1.6)$$

$$\boldsymbol{\mu}_c = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1} (\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2), \quad (1.7)$$

$$\boldsymbol{\Sigma}_c = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}. \quad (1.8)$$

With this, all correlations between all Gaussians can be calculated and written in three matrices $\mathbf{M}^{AA}$, $\mathbf{M}^{AB}$ and $\mathbf{M}^{BB}$:

$$M_{ij}^{AA} = \int a_i(\mathbf{x}) a_j(\mathbf{x}) d\mathbf{x}, \quad (1.9)$$

$$M_{ij}^{AB} = \int a_i(\mathbf{x}) b_j(\mathbf{x}) d\mathbf{x}, \quad (1.10)$$

$$M_{ij}^{BB} = \int b_i(\mathbf{x}) b_j(\mathbf{x}) d\mathbf{x}. \quad (1.11)$$

Hence Equation (1.4) can be written as

$$D_{L2}(A, B) = \boldsymbol{\alpha}^T \mathbf{M}^{AA} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{M}^{AB} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{M}^{BB} \boldsymbol{\beta}. \quad (1.12)$$

Setting the derivative with respect to $\boldsymbol{\alpha}$ of Equation 1.12 to zero yields the optimal distribution of the L2 model with L1 Gaussians. As we actually achieved a more desirable runtime/performance trade off with less sophisticated projections, we refer to (Raab et al., 2009b) for details of the minimum search and focus here on the more relevant approximated projections.

The goal of the approximated projections is to map all HMMs of all L languages to one fixed set of N Gaussians (= Recognition Codebook, RC). When we have chosen any codebook all M^l Gaussians of each Monolingual Codebook (MC^l) can be mapped individually to the RC . Each Gaussian $\mathcal{N}$ is represented by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We map based on the smallest Mahalanobis distance (Gaussian Distance D_G)

$$\begin{aligned} \text{map}_G(\mathcal{N}_{MC^l}^i) &= \mathcal{N}_{RC}^j, \quad 0 \leq i < M^l, 0 \leq j < N, 0 \leq l < L \\ j &= \arg \min_k D_G(\boldsymbol{\mu}_{MC^l}^i, \boldsymbol{\mu}_{RC}^k, \boldsymbol{\Sigma}_{MC^l}^i). \end{aligned} \quad (1.13)$$

In the introduction we have motivated that it is advisable to use the monolingual codebook from the main language as RC . This case offers further possibilities how HMMs from other languages can be linked to the RC . All states from the main language map only to Gaussians from the RC . Thus when all S states are mapped to RS main language states only Gaussians from the RC are

used. The same applies when all HMMs are mapped to main language HMMs. Both of these additional mappings have the advantage that they consider the combination of Gaussians in their distance.

We map states based on the minimum Mahalanobis distance (D_S) between the expected values of their probability distributions. In our system the probability distribution p_s of every state s is a Gaussian mixture distribution with M^l Gaussians:

$$p_{s_l}(\mathbf{x}) = \sum_{i=0}^{M^l} w^i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (1.14)$$

The expected value of $\mathbf{x}$ for each state s is then

$$\begin{aligned} E(p_{s_l}(\mathbf{x})) &= E\left(\sum_{i=1}^{M^l} w_{s_l}^i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right) \\ &= \sum_{i=1}^{M^l} w_{s_l}^i \boldsymbol{\mu}_i. \end{aligned} \quad (1.15)$$

The covariance which is needed for the Mahalanobis distance is a global diagonal covariance $\boldsymbol{\Sigma}_{All}$ estimated on all training samples. With D_S we define our state based mapping as

$$\begin{aligned} \mathbf{maps}(s_l^i) &= s_{RS}^j, 0 \leq i < S_l, 0 \leq j < RS, 0 \leq l < L \\ j &= \arg \min_k D_S(E(s_l^i), E(s_{RS}^k), \boldsymbol{\Sigma}_{All}). \end{aligned} \quad (1.16)$$

Based on D_S we can also define a distance between HMMs (D_H). In our system each context dependent phoneme is represented through a three state HMM model. In this case the distance between two phonemes $\mathbf{q}_1$ and $\mathbf{q}_2$ is

$$D_H(\mathbf{q}_1, \mathbf{q}_2) = \sum_{i=1}^3 D_S(s_{q_1}^i, s_{q_2}^i). \quad (1.17)$$

Similar as for D_S , map_H can be defined with D_H . D_G and D_S provide consistently good performance for different tests, while they use rather different information for their calculation. Therefore we also wanted to test a combined

map_{G+S} . This map is defined as

$$\mathbf{map}_{G+S}(\mathbf{s}_l^i) = \gamma_{G+S} \mathbf{maps}(\mathbf{s}_l^i) + (1 - \gamma_{G+S}) \begin{pmatrix} w_{s_l^i}^1 map_G(\mathcal{N}_{MC^l}^1) \\ w_{s_l^i}^2 map_G(\mathcal{N}_{MC^l}^2) \\ \vdots \\ w_{s_l^i}^{M^l} map_G(\mathcal{N}_{MC^l}^{M^l}) \end{pmatrix}$$

$$0 \leq l < L, 0 \leq i < S_l^+ \quad (1.18)$$

with the combination weight γ_{G+S} . γ_{G+S} has to be determined in experiments.

Figure 9 compares the L2 based projection and the four approximated projections for the native city tests. The results demonstrate that the L2 projection and the combined $\mathbf{map}_{G+S}$ perform roughly equal, and that there is still a gap to the monolingual benchmark systems. However, the last approximated projection needs only 300ms on an Intel Xeon Dual Core system with 3.6 GHz and two gigabyte of RAM for the mapping of one additional language. This is also the reason why in the following only this last projection will be considered, as it makes the realization of the generation of the multilingual system on the embedded system itself possible. This removes almost all additional training effort, and allows the creation of the system a user currently needs.

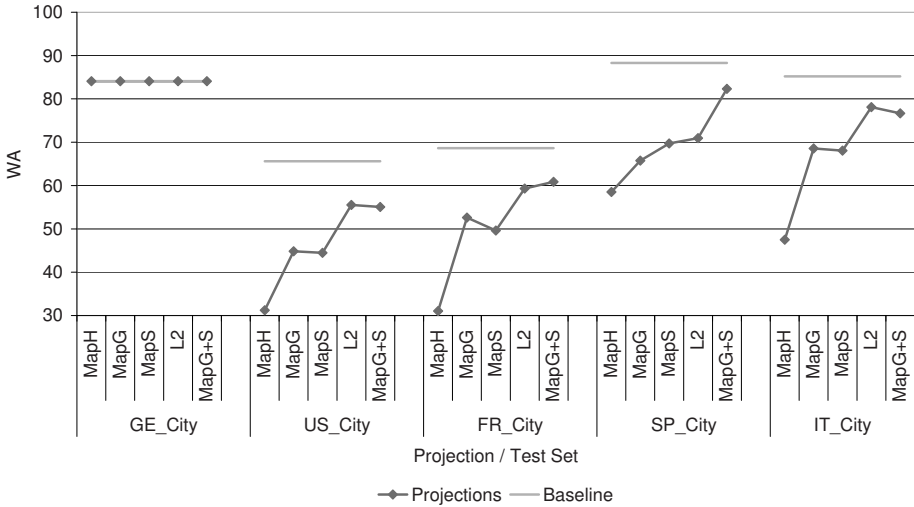


Figure 9. Performance of online generated HMMs on native speech of different languages. The recognition codebook contains all German Gaussians.

6.2 Scalable Architecture

The projections between GMMs have only been proposed to circumvent some of the problems that are inherent with other approaches, therefore the experiments of real interest are to combine the MWC algorithm and the projection algorithm. This is depicted in Figure 10 for two tasks of major interest for future car infotainment system, music selection and world wide destination input via voice.

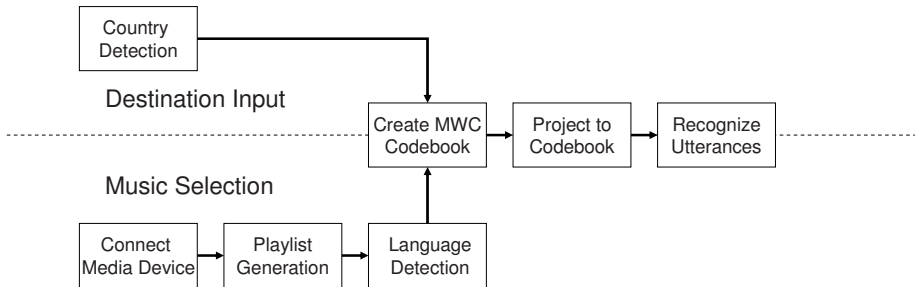


Figure 10. This diagram shows how the two application examples mentioned in the introduction can be processed with the proposed algorithms. Due to the MWC algorithm the performance can be scaled, and due to the projections algorithm systems for any language combination can be created in seconds.

This section evaluates again both native and non-native speech. In the first part, native speech of five languages is evaluated, first with German as the main language, then in a second set of experiments with English as the main language. The codebook of the baseline system only contains Gaussians from the main language, and the MWC systems always contain all Gaussians from the main language. However, the MWC systems also add some additional Gaussians for the language that is tested.

Figure 11 shows the Word Accuracies for city name tests. The x-axis indicates both which test is performed as well as the size of the MWC that is applied. The curves show that in the German test all three systems perform equal, the MWC based on the German codebook generated through the projections, the baseline and the benchmark system. This is actually a strength of the MWC approach, as commonly parameter sharing methods for multilingual speech recognition have slightly negative impacts for all languages. This is not the case for the main language in the MWC system.

The picture is different for the additional languages. Here it becomes evident that benchmark systems are better than other systems with parameter tying across languages. In the case that no additional Gaussians were added to the codebook, the baseline system also outperforms the projection system.

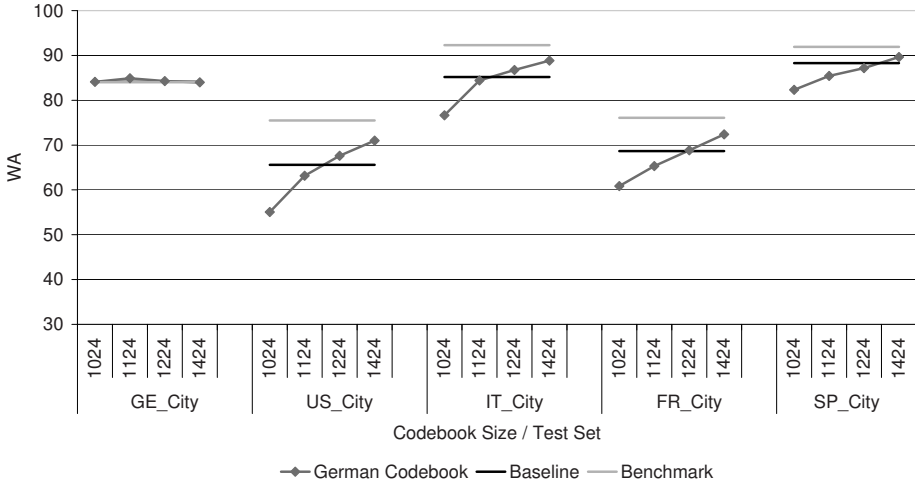


Figure 11. Performance of the scalable architecture on native speech of different languages with German as main language.

However, when the two algorithms are combined, and Gaussians are added before the projection is executed, the combined system outperforms the baseline system. This requires some additional resources at runtime for the additional Gaussians. However, a provision of the combined system requires less effort than a provision of the baseline system.

The set of results that is missing in this figure are the results of the MWC with a standard retraining. It is quite certain that this would give better performance than the MWC performance shown, but only for the experiments in this figure the same effort as building 20 monolingual recognizers would need to be done. In general it is unrealistic to provide Baum-Welch trained MWC systems for many languages and all their combinations today.

To verify that these results are not just due to some peculiarities of the German codebook, the same experiments are repeated with English as main language in Figure 12. The main difference is that now the English system remains unchanged. For the other tests, the trends are the same as before, the baseline system is better than projections alone and the combined system can outperform the baseline when more Gaussians are added. However, in general the performance is a little worse for the additional languages when compared to the results with German as the main language. We attribute this to the fact that German has more phonemes than English (59 compared to 46 phonemes), and that there are thus more sounds that are not well covered with an English codebook.

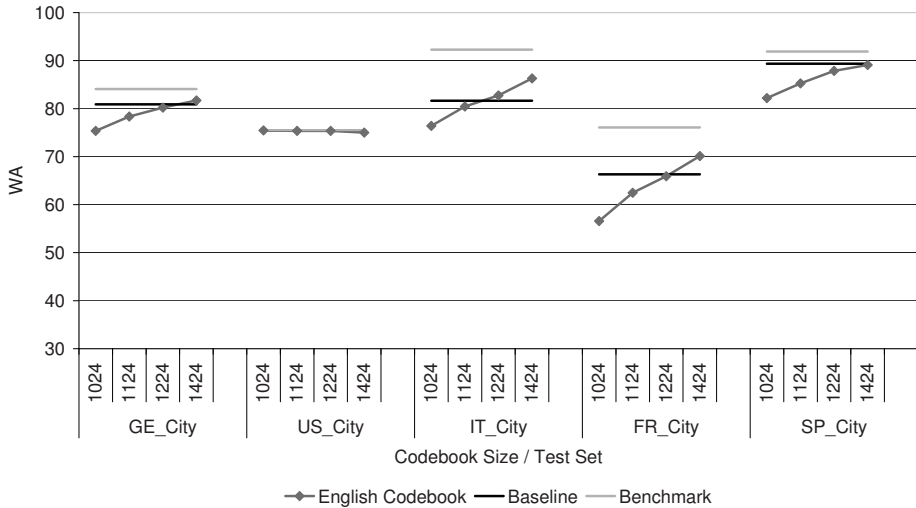


Figure 12. Performance of the scalable architecture on native speech of different languages with English as main language.

The second part of this section deals with the performance of the scalable architecture on non-native speech. The first three tests are non-native English by Spanish, French and Italian speakers. The last test contains song titles uttered by Germans. The last test contains stronger accented speech, as not all speakers were familiar with the language from which the name originated. Figure 13 shows the performance on these four tests. A major difference to the previous charts is that this time the main language was different for each test set, as it was always set to the native language of the speakers. It is also the case that no benchmark performance for the last test is given, as no monolingual system can recognize speech from three languages.

In the case of the speakers which were familiar with the spoken language, the proposed system benefits from the addition of the additional Gaussians, and comes close to the performance of the benchmark system when Gaussians are added. In the case of the less familiar speakers in the IFS MP3 test, however, the addition of Gaussians did not help. The performance on this test is also worse than for the other tests, which might indicate that their speech is just too different from the speech of native speakers of Spanish, French and Italian. This might again be due to misjudged language origins, as discussed in Section 5.2.

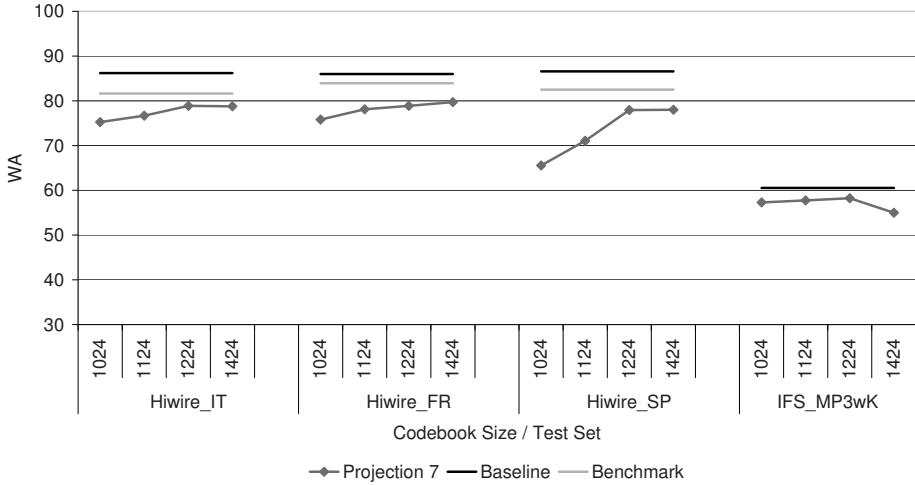


Figure 13. Performance of the scalable architecture on non-native speech, the main language is always the mother tongue of the speakers.

6.3 Footprint

In the previous section it was shown that the systems created in seconds by the scalable architecture can outperform traditionally trained systems with the help of more Gaussians. However, the main advantage of the scalable architecture over the benchmark systems is that it uses less memory. This is depicted in Figure 14.

The figure shows the memory usage for the complete speech recognizer application, not only for the acoustic model. This is also the reason why the rightmost system needs $\approx 400\%$ of memory compared to the one lingual system, and not $\approx 500\%$, which is true for the increase of the acoustic model. As discussed in previous experiments, context dependent modeling is not as important as for native speech, therefore it might be a viable solution to apply only monophones on devices with very constrained resources. In this case, the recognition of four more languages can be realized with 10% more resources. In other words, this is only 3% of the increase that would occur if the four monolingual recognizers are added.

Even in the case of improved modeling with additional Gaussians and with triphones in the 5LT 1424 system, the resource increase is only 29% of the increase of the 5LT 5120 system. However, one could also imagine of modeling only the most frequent triphones, to reduce the size of this system further, and still have sharper models as with monophones.

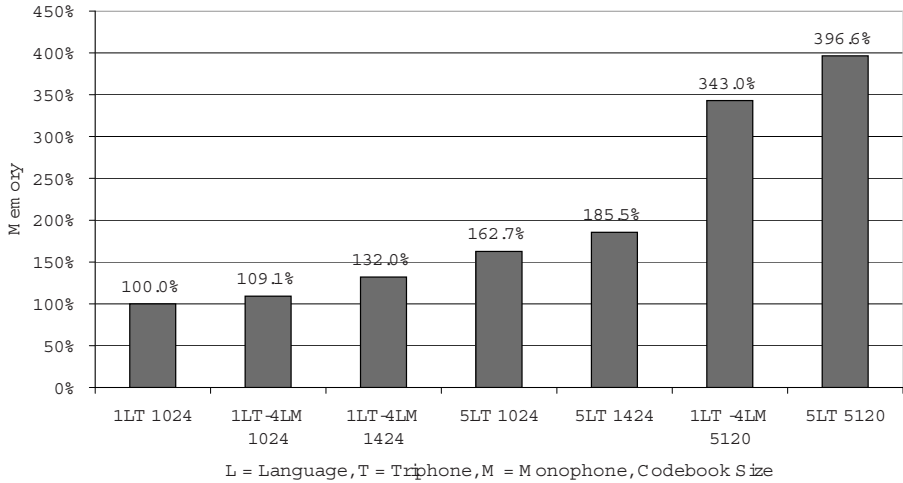


Figure 14. Increase of the memory footprint of the recognizer for five languages. The x-axis indicates the different system names. For example, the leftmost system covers 1 language with 1024 Gaussians, it is the monolingual comparison number. 1LT-4LM 1024 means that the main language is modeled with triphones, but the four other languages are modeled with monophones. At the other extreme, the 5LT 5120 is the memory demand of five monolingual acoustic models in parallel.

Thus, as always in multilingual speech recognition, it becomes again apparent that there are a multitude of choices that can be made. The advantage with the scalable architecture is that these decisions have not to be made once and then kept for a long time. Due to the fast provision, the models can be adapted to each new situation. For some devices, the smallest acoustic models can be created, and for more powerful devices, more efficient models are possible. That is the real advantage of the scalable architecture: It allows to provide just the system that is currently most appropriate.

7. Summary

This chapter has dealt with the problem of multilingual speech for resource-constrained dialogue systems. The application we had in mind during the construction of the algorithms was a future car infotainment system that allows music selection via voice and international destination input. However, there are many other scenarios in which such systems can be of use, for example in music players or cell phones. The two main goals for our work were to allow the recognition of many languages with limited resources and to keep monolingual performance on the native language of the user.

The described work has first concentrated on the improved modeling of non-native speech. In our experiments with monophones and triphones, we could show that triphones that are trained with large amounts of data tend to outperform context independent models. We therefore performed the other experiments mostly with context dependent models. We then evaluated three different methods for accent adaptation. Our MWC system allowed a resource efficient modeling of multiple languages, and improved the recognition by up to 25% rel. WER for non-native accented speakers. The model merging approach as introduced by Witt and Young (1999) did in our different scenario only yield 14% rel. WER improvement, and a combination with the MWC algorithm did not show any additional improvements. Finally, we also tested retraining with non-native development data and achieved up to 33% rel. WER improvements in the realistic case where we did not have adaptation data from the test speakers. Due to the costs involved with the collection of accent specific data for hundreds of accents, we came to the conclusion that the MWC approach is the most appropriate technique in our scenario.

In the second part of the work some drawbacks of the MWC approach have been tackled. With a projection between GMMs a scalable architecture on embedded systems can be realized. This architecture allows the creation of user or task adapted systems within less than a second on a dual core computer and will be feasible on many embedded systems. Furthermore it has removed almost all additional training effort, allows an efficient decoding of many languages due to the MWC approach and can provide a system for every language combination a user needs, as long as there is a monolingual speech recognizer for each of the requested languages available.

Notes

1. For the experiments with varying amount of training data, another training database was applied, to have more training data available. This is also the reason why the following sections report different numbers for the same tests.

References

- Amdal, I., Korkmazskiy, F., and Surendran, A. C. (2000). Joint Pronunciation Modelling of Non-Native Speakers Using Data-Driven Methods. In *Proceedings of Interspeech*, pages 622–625, Beijing, China.
- Bartkova, K. and Jouvet, D. (2006). Using Multilingual Units for Improved Modeling of Pronunciation Variants. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1037–1040, Toulouse, France.
- Bouselmi, G., Fohr, D., and Illina, I. (2007). Combined Acoustic and Pronunciation Modeling for Non-Native Speech Recognition. In *Proceedings of Interspeech*, pages 1449–1552, Antwerp, Belgium.

- Dalsgaard, P., Andersen, O., and Barry, W. (1998). Cross-Language Merged Speech Units and their Descriptive Phonetic Correlates. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, page no pagination, Sydney, Australia.
- Goronzy, S., Sahakyan, M., and Wokurek, W. (2001). Is Non-Native Pronunciation Modeling Necessary? In *Proceedings of Interspeech*, pages 309–312, Aalborg, Denmark.
- Gruhn, R., Markov, K., and Nakamura, S. (2004). A Statistical Lexicon for Non-Native Speech Recognition. In *Proceedings of Interspeech*, pages 1497–1500, Jeju Island, Korea.
- He, X. and Zhao, Y. (2001). Model Complexity Optimization for Nonnative English Speakers. In *Proceedings of Interspeech*, pages 1461–1464, Aalborg, Denmark.
- Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing: a Guide to Theory, Algorithm and System Development*. Prentice Hall.
- Iskra, D., Grosskopf, B., Marasek, K., van den Huevel, H., Diehl, F., and Kiessling, A. (2002). Speecon - Speech Databases for Consumer Devices: Database Specification and Validation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 329–333, Las Palmas de Gran Canaria, Spain.
- Jian, B. and Vemuri, B. C. (2005). A Robust Algorithm for Point Set Registration using Mixture of Gaussians. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1246–1251, Beijing, China.
- Juang, B. H. and Rabiner, L. R. (1985). A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal*, 64(2):391–408.
- Kim, M., Oh, Y. R., and Kim, H. K. (2007). Non-Native Pronunciation Variation Modeling using an Indirect Data Driven Method. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 231–236, Kyoto, Japan.
- Koehler, J. (2001). Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. *Speech Communication Journal*, 35(1-2):21–30.
- Ladefoged, P. (1990). The Revised International Phonetic Alphabet. *Language*, 66(3):550–552.
- Lang, H. (2009). Methods for the Adaptation of Acoustic Models to Non-Native Speakers. Diplomarbeit, Institute of Information Technology, University Ulm, Ulm, Germany.
- Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., and Souter, C. (2000). The ISLE Corpus of Non-Native Spoken English. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 957–963, Athens, Greece.

- Niesler, T. (2006). Language-Dependent State Clustering for Multilingual Speech Recognition in Afrikaans, South African English, Xhosa and Zulu. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing (MULTILING)*, Stellenbosch, South Africa.
- Petersen, K. and Pedersen, M. (2008). The Matrix Cookbook. <http://matrixcookbook.com>.
- Raab, M., Aradilla, G., Gruhn, R., and Nöth, E. (2009a). Online Generation of Acoustic Models for Multilingual Speech Recognition. In *Proceedings of Interspeech*, pages 2999–3002, Brighton, UK.
- Raab, M., Gruhn, R., and Nöth, E. (2007). Non-Native Speech Databases. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 413–418, Kyoto, Japan.
- Raab, M., Gruhn, R., and Nöth, E. (2008a). Multilingual Weighted Codebooks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4257–4260, Las Vegas, USA.
- Raab, M., Gruhn, R., and Nöth, E. (2008b). Multilingual Weighted Codebooks for Non-Native Speech Recognition. In *Proceedings of International Conference on Text, Speech and Dialogue (TSD)*, pages 485–492, Brno, Czech Republic.
- Raab, M., Schreiner, O., Herbig, T., Gruhn, R., and Nöth, E. (2009b). Optimal Projections between Gaussian Mixture Feature Spaces for Multilingual Speech Recognition. In *Proceedings of International Conference on Acoustics (NAG-DAGA)*, pages 411–414, Rotterdam, Netherlands.
- Schaden, S. (2006). *Regelbasierte Modellierung fremdsprachlich akzent-behafteter Aussprachevarianten*. PhD thesis, University Duisburg-Essen, Duisburg, Germany.
- Schultz, T. and Waibel, A. (2001). Language-Independent and Language-Adaptive Acoustic Modeling for Speech Recognition. *Speech Communication*, 35:31–51.
- Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.-A., Clot, V., Gemello, R., Matassoni, M., and Maragos, P. (2007). The HIWIRE Database, a Noisy and Non-Native English Speech Corpus for Cockpit Communication. <http://www.hiwire.org/>.
- Tan, T. P. and Besacier, L. (2007a). Acoustic Model Interpolation for Non-Native Speech Recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1009–1013, Honolulu, Hawaii.
- Tan, T. P. and Besacier, L. (2007b). Modeling Context and Language Variation for Non-Native Speech Variation. In *Proceedings of Interspeech*, pages 1429–1432, Antwerp, Belgium.

- Tomokiyo, L. M. and Waibel, A. (2001). Adaptation Methods for Non-Native Speech. In *Proceedings of Multilinguality in Spoken Language Processing (MSLP)*, pages 39–44, Aalborg, Denmark.
- Witt, S. (1999). *Use of Speech Recognition in Computer-Assisted Language Learning*. PhD thesis, Cambridge University Engineering Department, Cambridge, UK.
- Witt, S. and Young, S. (1999). Off-Line Acoustic Modeling of Non-Native Accents. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, Budapest, Hungary.

Chapter 2

ONLINE LEARNING OF BAYES RISK-BASED OPTIMIZATION OF DIALOGUE MANAGEMENT FOR DOCUMENT RETRIEVAL SYSTEMS WITH SPEECH INTERFACE

Teruhisa Misu, Komei Sugiura

National Institute of Information and Communications Technology (NICT)
Kyoto, Japan

{teruhisa.misu, komei.sugiura}@nict.go.jp

Tatsuya Kawahara

Kyoto University
Kyoto, Japan

kawahara@i.kyoto-u.ac.jp

Kiyonori Ohtake, Chiori Hori, Hideki Kashioka, Satoshi Nakamura

National Institute of Information and Communications Technology (NICT)
Kyoto, Japan

{kiyonori.ohtake, chiori.hori, hideki.kashioka, satoshi.nakamura}@nict.go.jp

Abstract

We propose an efficient online learning method of dialogue management based on Bayes risk criterion for document retrieval systems with a speech interface. The system has several choices in generating responses. So far, we have optimized the selection as minimization of Bayes risk based on reward for correct information presentation and penalty for redundant turns. In this chapter, this framework is extended to be trainable via online learning by maximum likelihood estimation of success probability of a response generation. Effectiveness

of the proposed framework was demonstrated through an experiment with a large amount of utterances of real users. The online learning method was then compared with the method using reinforcement learning and discussed in terms of convergence speed.

Keywords: Dialogue management.

1. Introduction

Numerous spoken dialogue systems have been developed in the past years. Their typical task domains have included airline information (ATIS & DARPA Communicator) (Levin et al., 2000; Potamianos et al., 2000; Rudnicky et al., 2000; Seneff and Polifroni, 2000), train information (ARISE and MASK) (Lamel et al., 1999; Sturm et al., 1999; Lamel et al., 2002), weather information (Zue et al., 2000), and bus location tasks (Komatani et al., 2005; Raux et al., 2005). Although these systems can handle simple database retrieval or transactions with constrained dialogue flows, there has been increasing demand for spoken dialogue systems that can handle more complex tasks. Meanwhile, increasingly more electronic text resources have recently been accumulated, attracting attention as the wisdom of crowds, which potentially can provide almost any desired kind of knowledge. In addition, since most documents are indexed (e.g., via Web search engines), we can potentially access these. The recent progress in natural language processing techniques has also enabled us to manage complex search queries. In such situations, the demand for document retrieval by using speech input has been increasing. In fact, in recent years, the target of spoken dialogue systems has been extended to the retrieval of documents (Chen et al., 2005; Reithinger and Sonntag, 2005; Pan and Lee, 2007), including manuals (Misu and Kawahara, 2006b) and Web pages (Brøndsted et al., 2006). We have also developed an information navigation system based on document retrieval and presentation called the “Dialogue Navigator for Kyoto City” (Misu and Kawahara, 2007).

There are quite a few choices in spoken dialogue systems for handling user utterances and generating responses that involve parameter tuning. Since a subtle change in these choices may affect the behavior of the entire system, they are usually manually tuned by experts. In addition, every time the system is updated, such as when knowledge bases are added or changes are made to the ASR modules, the parameters must be re-tuned. Due to the high cost of such tuning, there have been many studies that have addressed the automatic optimization of dialogue management (Bohus et al., 2006; Lemon and Pietquin, 2007; Kim et al., 2008b), and most of these have dealt with database retrieval tasks (Young et al., 2007; Kim et al., 2008a). The dialogue process in these studies has been designed using the formulation of Markov decision processes

(MDPs) and trained by reinforcement learning (RL) (Roy et al., 2000; Levin et al., 2000; Singh et al., 2002). The dialogue process in these frameworks needs to be mapped into a finite number of states. Since a list of database slots and a definite set of keywords are prepared a priori and manually in relational database (RDB) query tasks, the dialogue process is easily managed based on these. Such mapping is straightforward. For example, in a train information task, the combination of statuses of database slots, (such as “blank”, “filled” and “confirmed”) can be used as one dialogue state.

However, they cannot be directly applied to document retrieval tasks with a speech interface, where there is no relational structure in the document and every word is used in matching. The system has several choices for generating responses. Confirmation is needed to eliminate any misunderstandings caused by ASR errors, but users easily become irritated with too many redundant confirmations. Although there have been several studies dealing with dialogue management in call routing systems (Levin and Pieraccini, 2006; Horvitz and Paek, 2006), these methods cannot be applied to complex decision making processes in information guidance tasks. For example, our navigation system classifies user utterances into two types of information queries and factoid wh-questions, and generates appropriate responses to respective inputs. Unlike conventional question-answering (QA) tasks, in which all user inputs are assumed to be wh-questions, such as the TREC QA Track ¹, it is often difficult to tell whether an utterance is an information query or a wh-question (Rosset et al., 2006). In addition, there is not necessarily an exact answer to a wh-question in a document set. Therefore, it is not always optimal to respond to a wh-question with only its answer.

In order to manage the choices in response generation efficiently, we have proposed Bayes risk-based dialogue management based on reward for correct information presentation and penalty for redundant turns as well as the score of document retrieval and answer extraction (Misu and Kawahara, 2009). However, this method requires a large amount of training data for parameter optimization. The strategy must be retrained when the reward/penalty parameters were updated. In this chapter, we extend this framework to be trainable online. The proposed method is intended for efficient learning with a small number of samples. It is based on maximum likelihood estimation using Fisher’s scoring algorithm, and optimality of the dialogue strategy is guaranteed even when the reward or penalty parameters were re-tuned after the learning. We also compare the proposed method with a method using reinforcement learning and discussed in terms of convergence speed.

2. Dialogue Management and Response Generation in Document Retrieval System

2.1 System Overview

We first describe the speech based document retrieval task. For this information navigation task, the user produces a spoken query and the system provides a response retrieved from a set of documents. When generating responses, the user utterances are classified into two categories. The first consists of information queries, such as “Please tell me about the Golden Pavilion”. For such information queries, the system retrieves the appropriate document from the knowledge base (KB) and presents it to the user. The second category consists of wh-questions, such as “When was it built?”. The system extracts the sentence from the KB that includes the answer to the wh-question and presents it to the user. This six-step procedure is summarized below and is also outlined in [Figure 1](#). The system:

- 1 Recognizes the user utterance.
- 2 Classifies it as an information query or a wh-question.
- 3 Concatenates contextual information (on previous user utterances).
- 4 Makes a retrieval from the KB.
- 5
 - (a) Extracts one sentence that includes the answer to the wh-question from the search result if the utterance is a wh-question.
 - (b) Summarizes the document with the maximum matching score if the utterance is an information query.
- 6 Outputs the response through synthesized speech.

When the utterance is recognized/identified as an information query, the system presents (or confirms to present) a document. Even though it is possible to present the whole document, it is preferable to suppress the amount of information in the speech output. The system makes a summary by extracting important sentences, taking into consideration the positions of sentences and the co-occurrences count of nouns. However, summarizing the retrieved document may cause important parts of the information to be lost that the user wanted to know about or may have been interested in. Therefore, we incorporate a factoid QA technique to follow up the initial query, enabling random access to any part of the document using named entities (NEs) as a clue to access (Misu and Kawahara, 2009). Identification of wh-question types is done by detecting cue phrases in individual utterances and is backed off to the information query mode if unreliable.

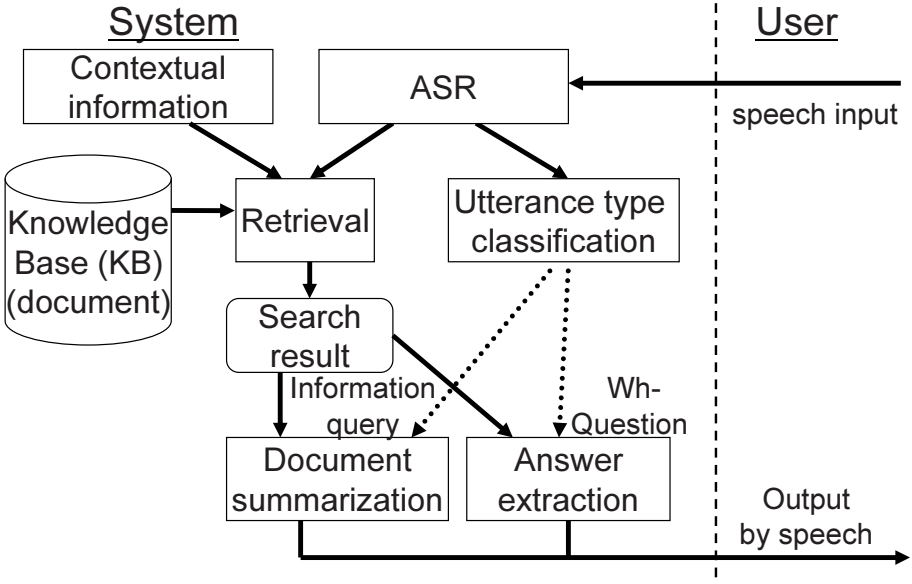


Figure 1. Overview of document retrieval and QA with speech interface.

2.2 Knowledge Base (KB)

We adopted a sightseeing guide for Kyoto city as the target domain. We use two document sets as the KBs for this domain. The first set consisted of Wikipedia documents concerning Kyoto. We selected the entries that contained the word “Kyoto”. Many of these documents concern sightseeing spots, such as Kinkaku-ji temple, Kiyomizu-dera temple, or Nijo-jo castle. As shown in Figure 2, these documents have a structure of section units. The second document set is the official information database on Kyoto city, containing brief explanations of the spots, together with access information, business hours, and entrance fees. Table 1 lists the sizes of these KBs.

Table 1. Specifications of sightseeing guidance knowledge base (KB).

	No. of documents	No. of sections	No. of words	Vocabulary size
Wikipedia	269	678	150 K	14,285
Tourist information	541	541	70 K	8,639
Total	810	1,219	220 K	17,933

Kinkaku-ji**(Overview)**

Kinkaku-ji is the informal name of Rokuon-ji in Kyoto, Japan. The Golden Pavilion, or Kinkaku, is a three-story building on the grounds of the temple. The Golden Pavilion, or Kinkaku, is a three-story building on the grounds of the temple. . . .

History

This temple was originally built in 1397 to serve as a retirement villa for Shogun Ashikaga Yoshimitsu, as part of his estate then known as Kiyama. It was his son who converted the building into a Zen temple of the Rinzai school. The temple was burned down several times during the Onin War. . . .

Garden

The Golden Pavilion has a magnificent Japanese garden right outside of it. The pond in front of it is called Kyoko-chi. There are many islands and stones on the pond that represent the Buddhist creation story. . . .

Figure 2. Example of Wikipedia document (translation of Japanese).

We developed a guidance system based on document retrieval from the KBs called the “Dialogue Navigator for Kyoto City”. There is an example dialogue with the system in [Figure 3](#).

-
- U1: Please tell me about the Golden Pavilion. (Information query)
 S1: The Golden Pavilion is one of the buildings in the Rokuon-ji in Kyoto, and is the main attraction of the temple sites. The entire pavilion except the basement floor is covered with pure gold leaf.
 U2: When was it built? (WH-Question)
 S2: The Golden Pavilion was originally built in 1397 to serve as a retirement villa for Shogun Yoshimitsu.
 U3: Then, please tell me its history. (Information query)
-

.....

Figure 3. Example dialogue of Dialogue Navigator for Kyoto City.

2.3 Backend Retrieval System

We adopted a standard vector space model to calculate the matching score between a user utterance (= ASR result) and the document in the KB. That is, the vector of the document $\mathbf{d} = (x_1, x_2, \dots, x_n)^T$ was created by using the occurrence counts of nouns in the document by the section unit. The vector

for the user utterance $\mathbf{W} = (w_1, w_2, \dots, w_n)^T$ is also created from the ASR result. Here, x_i and w_i are occurrence counts for noun i . The matching score $Match(\mathbf{W}, \mathbf{d})$ is calculated as the product of these two vectors. The ASR confidence measure $CM(w_i)$ (Lee et al., 2004) is also used as a weight for the occurrence count

$$Match(\mathbf{W}, \mathbf{d}) = \sum_i^n x_i \cdot w_i \cdot CM(w_i). \quad (2.1)$$

2.4 Backend Question-Answering System

A QA system typically consists of a question classifier module that determines the type of question and an answer extraction module that generates answer candidates from the KB and selects the most appropriate candidates using some scoring function (Ravichandran and Hovy, 2002; NIST and DARPA, 2003).

Our system is able to handle six types of wh-questions: “person name”, “place”, “date”, “length/height”, “price” and “access information”. Heuristic rules that consist of 114 Japanese cue phrases were hand-crafted in this work to classify the types of the user utterances. For instance, the input is determined to be a question of the person name type when the cue phrase “*Daredesuka* (who)” is included in the ASR result, and is determined to be a question of the price type when “*Ikuradesuka* (how much)” is included. Each rule maps the input to the corresponding type². Six types of NEs that correspond to the target question types were labeled a priori. We used the Japanese NE tagger CaboCha (T. Kudo, 2003) to detect “person name”, “place”, “date”, “length/height” and “price”. “Access information” was manually labeled for the Tourist information KB.

We implemented an answer extraction module that consists of commonly-used procedures. The system extracts NEs or answer candidates NE_i that correspond to the wh-question type from the retrieved documents. It then computes the score of QA QA_SCORE for each answer candidate NE_i using the following three features. Here, $Sent_i$ denotes the sentence containing NE_i and $Bunsetsu_i$ denotes a set of *bunsetsu*³ that have a dependency relationship with the *bunsetsu* containing NE_i .

- $CM(CP_i)$: The ASR CM of the cue phrase CP_i (such as “who” or “where”) used to classify the question type, which corresponds to the type of NE_i .
- MS_i : The number of times nouns in the input wh-question appear in $Sent_i$.

- MDC_i : The number of times nouns in the input wh-question appear in $Bunsetsu_i$.

Then, the score of QA QA_SCORE for NE_i is calculated as

$$QA_SCORE_{NE_i} = CM(CP_i) \cdot MS_i \cdot MDC_i. \quad (2.2)$$

In the baseline system for the preliminary analysis, the sentence containing the NE_i with the highest score (product of *Match* of the document containing NE_i and $QA_SCORE_{NE_i}$) is used for an answer.

2.5 Use of N-Best Hypotheses of ASR and Contextual Information for Generating Responses

Errors are inevitable in large vocabulary continuous speech recognition. The retrieval result would be severely damaged if some important information was not correctly recognized. Even if the first-best hypothesis includes an error, the correct recognition result may be included in the N-best hypotheses. We thus use the N-best hypotheses of the ASR result to create a search query and extract an answer.

Users of interactive retrieval systems tend to make utterances that include anaphoric expressions⁴. In these cases, it is impossible to extract the correct answer by only using the current utterance. Since the deterministic anaphora resolution (Matsuda and Fukumoto, 2006) is complex and prone to errors, stochastic matching is used in information retrieval, and we adopt a strategy that concatenates contextual information or nouns in the user's previous utterances to generate a search query (Murata et al., 2006). The simplest way is to use all the utterances made by the current user. However, this might also add inappropriate contexts because the topic might have changed during the session. De Boni (Boni and Manandhar, 2005) proposed an algorithm for detecting topics based on similarity of question sequences in a question-answering task with typed text input. We track the topic using metadata from the KB or the title of the document. The topic is tracked using the current document that have been focused on, which usually correspond to a sightseeing spot or Wikipedia entry. For example, as long as the user is apparently talking about the Golden Pavilion, the topic is fixed to the "Golden Pavilion". Thus, the occurrence counts of nouns within the context (weighted by their ASR CM) are incorporated when generating a search query W_i .

2.6 Field Test of Trial System

We carried out a field test at Kyoto University museum. Users ranged in a wide variety of ages from children to seniors and apparently had little expe-

rience with using spoken dialogue systems. During an exhibition that lasted three months, 2,564 dialogue sessions were collected. A trigram language model for the ASR system was trained using the KB, a dialogue corpus from a different domain, and Web texts (Misu and Kawahara, 006a). The vocabulary size of the model was 55,000. The average word accuracy for the information queries and the wh-questions was 72.0%.

We constructed a test set using 1,416 in-domain utterances that include 1,084 information queries as well as 332 wh-questions collected over a particular period of time (the first one third of the period). The average length (number of words) of the utterances was 4.8. These utterances were manually transcribed and labeled with the correct answers (= documents for information queries or answer NEs for wh-questions).

3. Optimization of Dialogue Management in Document Retrieval System

3.1 Choices in Generating Responses

In this chapter, among the many choices available in the system, we focus on choices related to the generation of responses and confirmations. Confirmation is indispensable to avoid inappropriate documents from being presented especially when the score for retrieval is low. It may also be “safer” to present the entire document than to present a specific answer to the user’s wh-question, when the score for answer extraction is low. For example, in the example dialogue in [Figure 3](#), if the system cannot find the exact answer of “1,397” for U2, the system can present a document about the history of the pavilion that may include information about the construction.

This kind of choices in conventional studies were made based on combinations of empirical knowledge, such as the ASR performance and the task type. However, hand-crafting heuristic rules is usually costly, and subtle changes in choices can seriously affect the performance of the whole system. Therefore, we propose a formulation where the above choices are optimized through on-line learning.

3.2 Optimization of Responses based on Bayes Risk

In the following subsections, we review the Bayes risk-based dialogue management that we have proposed in (Misu and Kawahara, 2009).

Bayes risk $L(d_j|\mathbf{W})$ is minimized in general pattern classification to determine the optimal class d_j for an input $\mathbf{W}$. In the Bayes classifier this is defined

by

$$L(d_j|\mathbf{W}) = \sum_{i=1}^n l(d_j|d_i)p(d_i|\mathbf{W}), \quad (2.3)$$

where $(d_1, d_2, \dots, d_n)$ denotes the given classes and $p(d_i|\mathbf{W})$ denotes the posterior probability for class d_i of $\mathbf{W}$. $l(d_j|d_i)$ is the loss function and represents the loss of predicting class d_j when the true class is d_i .

These classes $(d_1, d_2, \dots, d_n)$ in our document retrieval task correspond to all documents. We assume the loss function among classes is the same, and we extend the framework to reward (negative loss; $l(d_j|d_i) < 0$) appropriate classifications:

$$l(d_j|d_i) = \begin{cases} -Rwd & \text{if } j = i, \\ Penalty & \text{otherwise.} \end{cases} \quad (2.4)$$

As a result, from Equation (2.3), we obtain the Bayes risk $L(d_j|\mathbf{W})$ to determine document d_j for input W :

$$L(d_j|\mathbf{W}) = -Rwd * p(d_j|\mathbf{W}) + Penalty * (1 - p(d_j|\mathbf{W})). \quad (2.5)$$

In the spoken dialogue system, there are several choices in the manner of response or action to the user's request. Thus, we can define the Bayes risk for each response candidate. The *Rwd* and *Penalty* values are determined depending on the manner of response, and are defined by the degree of benefit to the user based on the correct information presentation and the loss caused by redundant time:

$$L(Res_i(d_j)|\mathbf{W}) = -Rwd_{Res_i} * p(d_j|\mathbf{W}) + Penalty_{Res_i} * (1 - p(d_j|\mathbf{W})). \quad (2.6)$$

The optimal choice is made by selecting the response that has the minimal amount of risk.

The relationship between the proposed method and conventional cost/reward functions used in dialogue management is as follows: The dialogue management proposed by Niimi (Niimi and Kobayashi, 1996) can be thought of as a case where the reward function is constant, and the penalty differs depending on the manner of confirmation, i.e., the explicit or implicit confirmation. The dual cost method (Dohsaka et al., 2003) takes into consideration the cost of several methods of presentation as a penalty, but the reward is constant.

3.3 Generation of Response Candidates

Bayes risk-based dialogue management (Misu and Kawahara, 2009) is accomplished by comparing the possible responses hypothesized by varying the

conditions for generating the search queries for KB retrieval and the manner of response and then selecting an appropriate response from the set of these responses. This chapter focuses on response generation, and we do not deal with optimization in search query generation in (Misu and Kawahara, 2009). That is, among the retrieved candidate documents by varying the manner in which the N-best hypotheses of ASR are used (the 1st, 2nd, or 3rd hypothesis, or all of them) and choosing whether to use contextual information, the candidate document is fixed to the hypothesis with the maximum likelihood (=matching score) of retrieval.

The possible response set **Res** includes answering $Ans(d_i)$, presentation $Pres(d_i)$, confirmation $Conf(d_i)$, and rejection $Rej(d_i)$. $Ans(d_i)$ denotes the user's specific wh-question being answered, which is generated by extracting one specific sentence that includes an answer named entity (NE) to the wh-question. $Pres(d_i)$ denotes a simple presentation of document d_i , which is actually made by summarizing it. $Conf(d_i)$ is an explicit confirmation⁵ for presenting document d_i . Rej denotes a rejection: the system gives up making a response from document d_i and request the user for a rephrasal. This flow is illustrated in Figure 4.

3.4 Definition of Bayes Risk for Candidate Response

The dialogue management we propose is accomplished by comparing and then selecting from possible responses hypothesized by varying the condition. We define the Bayes risk based on the reward for success, the penalty for failure, and the probability of success, which is approximated by the confidence measure of the document matching (Section 3.5), for response candidates. That is, a reward is given depending on the manner of response (Rwd_{Ret} or Rwd_{QA}) when the system presents an appropriate response. On the other hand, when the system presents an incorrect response, a penalty is given based on extraneous time, which is approximated by the total number of sentences in all turns before the appropriate information is obtained. For example, the penalty for a confirmation is 2 {system's confirmation + user's approval}, and that of a rejection is 1 {system's rejection}. When the system presents incorrect information, the penalty for a failure *FailurePenalty* (FP) is calculated, which consists of an improper presentation, the user's correction, and the system's request for rephrasal. Penalty for additional sentences to complete a task (*AddSent*) is also given as extraneous time before accessing the appropriate document when the user rephrases a information query/wh-question. The value of *AddSent* is calculated as an expected number of additional sentences before accessing the correct response assuming the probability for success by rephrasal was p . The

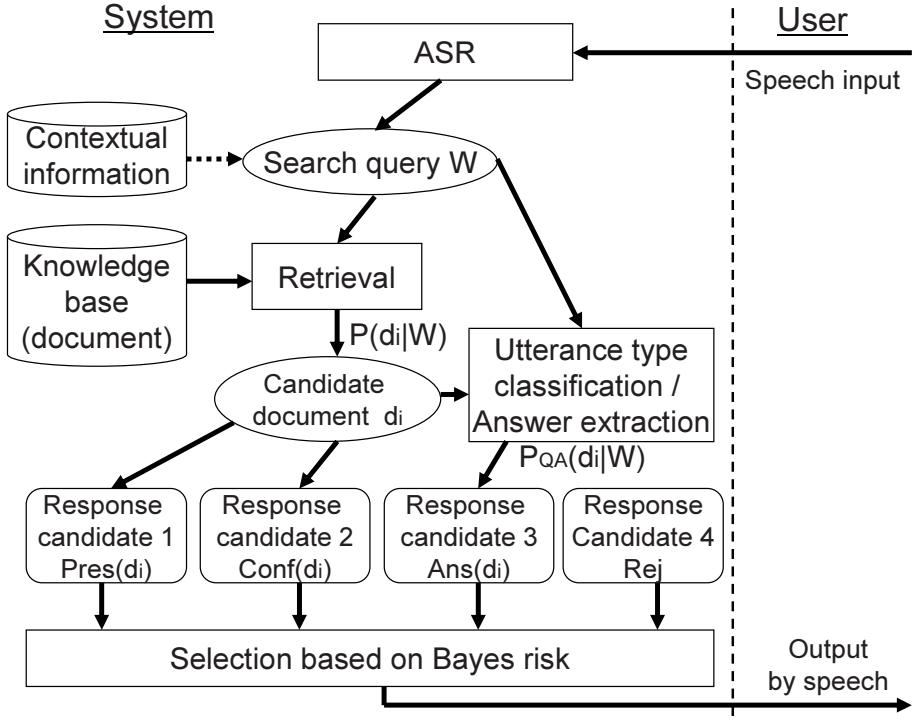


Figure 4. Overview of Bayes risk-based dialogue management.

AddSent for a retrieval is calculated as

$$\begin{aligned}
 AddSent &= FP + p * 1 + (1 - p)(FP + p * 1 + (1 - p)(\dots)) \\
 &\cong \frac{(FP + p)}{p}.
 \end{aligned} \tag{2.7}$$

In the experiment described in this chapter, we use the success rate in the field trial (Misu and Kawahara, 2007). In particular, $p = 0.6$ is used. Thus, *AddSent* depends on variable *FP*.

The Bayes risk for the response candidates is formulated as follows using the success rate of document retrieval $p(d_i|\mathbf{W})$, success rate of answer extraction $p_{QA}(d_i|\mathbf{W})$, and the reward pair (Rwd_{Ret} and Rwd_{QA} ; $Rwd_{Ret} < Rwd_{QA}$) for successful presentations as well as the *FP* for inappropriate presentations.

■ **Answering wh-question using document d_i**

$$\begin{aligned}
 Risk(Ans(d_i)) &= -Rwd_{QA} * p_{QA}(d_i|\mathbf{W}) \\
 &\quad + (FP + AddSent) * (1 - p_{QA}(d_i|\mathbf{W}))
 \end{aligned} \tag{2.8}$$

■ **Presentation of document d_i (without confirmation)**

$$\begin{aligned} Risk(Pres(d_i)) &= -Rwd_{Ret} * p(d_i|\mathbf{W}) \\ &\quad + (FP + AddSent) * (1 - p(d_i|\mathbf{W})) \end{aligned} \quad (2.9)$$

■ **Confirmation for presenting document d_i**

$$\begin{aligned} Risk(Conf(d_i)) &= (-Rwd_{Ret} + 2) * p(d_i|\mathbf{W}) \\ &\quad + (2 + AddSent) * (1 - p(d_i|\mathbf{W})) \end{aligned} \quad (2.10)$$

■ **Rejection**

Since the success rate is 0 in this case, $Risk(Rej)$ is given as follows.

$$Risk(Rej) = 1 + AddSent \quad (2.11)$$

Figure 5 shows the relation between success rates and risks for response candidates. The risks of four response candidates are illustrated. Note that, the x-axis is $p(d_i|\mathbf{W})$ for $Pres(d_i)$, $Conf(d_i)$ and $p_{QA}(d_i|\mathbf{W})$ for $Ans(d_i)$. The optimal response candidate is determined for $p(d_i|\mathbf{W})$ and $p_{QA}(d_i|\mathbf{W})$ as shown in the bold line.

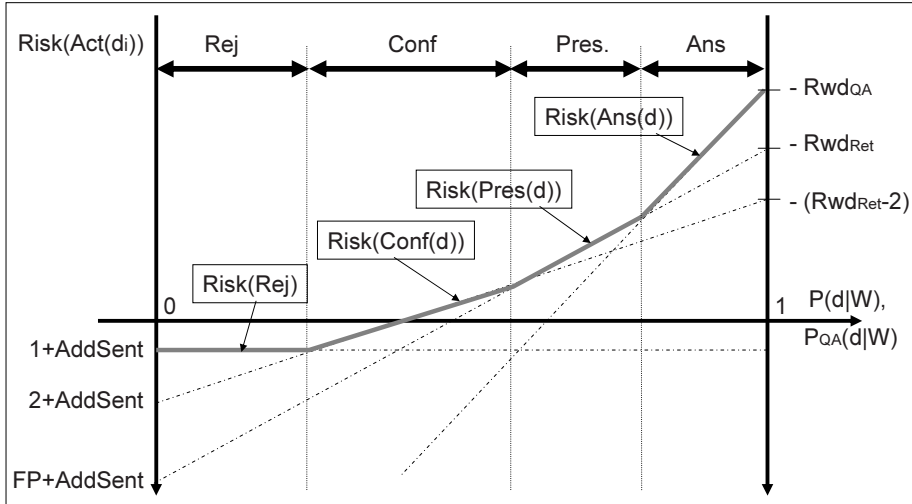


Figure 5. Success rates v.s. risks for response candidates.

Figure 6 shows an example of calculating a Bayes risk (where $FR = 6$, $Rwd_{Ret} = 5$, $Rwd_{QA} = 30$). In this example, since the answer to the user's question does not exist in the knowledge base, the score of answer extraction is low. Therefore, the system chooses a confirmation before presenting the entire document.

User utterance: When did the shogun order to build the temple?

(Previous query:) Tell me about the Silver Pavilion.

Response candidates:

Document with the largest score:

→ $p(\text{Silver Pavilion history}) = 0.4$

→ $p_{QA}(\text{Silver Pavilion history}) = 0.2$: In 1485

- $Risk(Ans(\text{Silver Pavilion history}; \text{In1485})) = 8.4$

- $Risk(Pres(\text{Silver Pavilion history})) = 6.4$

- **$Risk(\text{Conf}(\text{Silver Pavilion history})) = 4.8$**

* Rejection

- $Risk(Rej) = 9.0$

↓

Response: $\text{Conf}(\text{Silver Pavilion history})$

“Do you want to know the history of the Silver Pavilion?”

Figure 6. Example of Bayes risk calculation.

3.5 Confidence Measure of Information Retrieval and Question-Answering

Documents are retrieved using the matching module described in Section 2.3. The matching score $Match(\mathbf{W}, \mathbf{d}_i)$ is then transformed into a confidence measure $p(d_i)$ using a logistic sigmoid function. This is used as an approximation of $p(d_i|\mathbf{W})$ ⁶:

$$p(d_i) = \frac{1}{1 + \exp\{-\theta_1 * Match(\mathbf{W}, \mathbf{d}) - \theta_2\}}. \quad (2.12)$$

Here, θ_1 and θ_2 are parameters of the sigmoid function ($\theta_1 > 0$). The score of question-answering $QAScore$ is also transformed into a likelihood $p_{QA}(d_i|\mathbf{W})$ using another sigmoid function which is defined using another parameters of θ_3 and θ_4 ($\Theta = (\theta_1, \dots, \theta_4)$).

4. Online Learning of Bayes Risk-based Dialogue Management

4.1 Parameter Optimization by Online Learning

The Bayes risk-based dialogue strategy is trained by updating the parameters of sigmoid functions $\Theta = (\theta_1, \dots, \theta_4)$ so as to appropriately estimate the success rate of retrieval and question-answering. For tractable inputs, the system will learn to present documents or answers more efficiently. In contrast,

for intractable inputs, such as erroneous or out-of-system inputs, the system will learn to make confirmations or gives up as quickly as possible (appropriate action for such queries is “rejection”). Thus, training with several dialogue sessions should lead to optimal decisions being made considering the current success rate of retrieval.

The proposed method is also expected to adapt to changes in the data, by periodically conducting parameter updates. This is one of the advantages of using the proposed method, as compared to the previous works (Levin and Pieraccini, 2006; Horvitz and Paek, 2006).

The training procedure can be described in four steps.

- 1 (At each step t) Generate response candidates $Pres(d_i)$, $Conf(d_i)$, $Ans(d_i)$, and Rej from document d_i that has the largest likelihood $p(d_i)$.
- 2 Generate response $Res_t(d_i)$ (or select response candidate with minimum risk) for d_i and calculate actual reward/penalty.
- 3 Update parameters Θ . θ_1 and θ_2 are updated when the input is an information query, and θ_3 and θ_4 is updated for wh-questions. This is elaborated in the following subsections.
- 4 Return to step 1: $t \leftarrow t + 1$.

4.2 Optimization using Maximum Likelihood Estimation

We adopt the maximum likelihood estimation for learning the parameters Θ used for probability function (Kurita, 1994). Let the set of the learning samples be $\{ \langle Match_p, C_p \rangle \mid p = 1, \dots, t \}$, where a teaching signal C_p is given as a binary of 1 (success) or 0 (failure). If we assume the output of Equation (2.12) (relabeled as z_p) as an estimate of the conditional probability given an input $Match_p$, the log-likelihood l for the samples is given by the following cross entropy error function:

$$l = \sum_{p=1}^t \{ C_p \ln z_p + (1 - C_p) \ln(1 - z_p) \}. \quad (2.13)$$

The maximum likelihood estimate (MLE) of the weights is computed as one that maximizes this log-likelihood for previous t samples (Kurita, 1994). The MLE weights are used as Θ^{t+1} .

As an algorithm to solve the maximum likelihood equations numerically, we adopt the Fisher’s scoring algorithm that considers the variance in the score via the second derivative of the log of the likelihood function with respect to

Θ (or Fisher information that is often used in natural gradient approaches of RL (Peters and Schaal, 2008)). This is a type of Newton's method, and the MLE is calculated quickly (in less than five seconds) by matrix calculation. We demonstrate that the optimal value is obtained with a small number of samples.

4.3 Optimization using Steepest Descent

The parameters Θ can be optimized by the steepest descent method that simply uses the first derivative. The parameters Θ are updated using the following equation in order to minimize the mean square error between the estimated risk and the actual reward/penalty:

$$\Theta^{t+1} = \Theta^t + \delta \frac{\partial}{\partial \Theta} (ARP - Risk(Res_t(d_i)))^2. \quad (2.14)$$

Here, δ is a learning rate, which is empirically set to 0.001.

4.4 Online Learning Method using Reinforcement Learning

The optimal decisions can also be obtained using reinforcement learning (RL)⁸. The goal of the online learning using RL is to estimate the value $Q(S, A)$ of each response (or action) $\mathbf{A} = (Pres(d_i), Conf(d_i), Ans(d_i), Rej)$ for state space. In a document retrieval task, since the matching score $Match(W, d)$, which corresponds to the state space S in this task, can take any positive number, we need to train the value $Q(S, A)$ for the continuous state space. We thus represent the values of responses for the current state by a function approximation system instead of a lookup table (Singh et al., 2002). It should be noted that the POMDP solution technique using belief states with a delayed reward (e.g. (Williams and Young, 2007)) is similar to a RL for the continuous state space.

We approximate $Q(S, A)$ with triangle functions τ given by

$$\tau_m(S) = \begin{cases} 1 - |\frac{S}{\lambda} - m\lambda| & \text{if } |\frac{S}{\lambda} - m\lambda| < 1, \\ 0 & \text{otherwise,} \end{cases} \quad (2.15)$$

and values $\mathbf{V}_A = (V_A^0, V_A^\lambda, \dots, V_A^{n\lambda})$ for grid points. Thus, the value function $Q(S, A)$ is represented as follows:

$$Q(S, A) = \begin{cases} \sum_{m=0}^n V_A^{m\lambda} \cdot \tau_m(S) & \text{if } S < n\lambda, \\ V_A^{n\lambda} & \text{if } S \geq n\lambda. \end{cases} \quad (2.16)$$

Here, λ denotes the grid size and n denotes the number of grid points. [Figure 7](#) illustrates an example of a value function $Q(S, a)$ for an action a , where five

grid points and triangle functions $n = 4$ are used to approximate the function ($\mathbf{V}_A = (-4, 2, 5, 6, 8)$ are used.). There is another plane whose x-axis is $QAScore$; $Q(QAScore, Ans(d_i))$ is calculated in the plane. The optimal choice is made by selecting the response that has the minimum value of $Q(S, A)$ or $Q(QAScore, A)$.

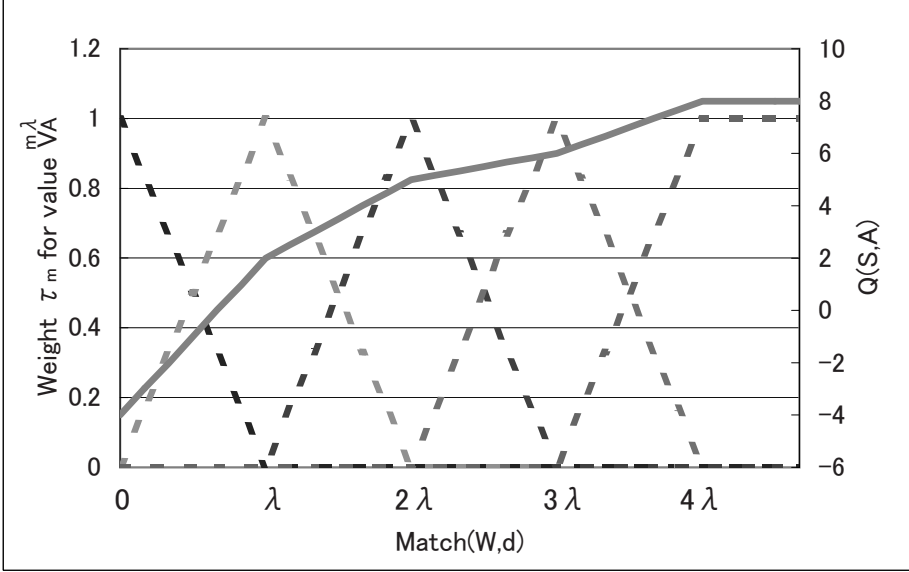


Figure 7. Example of a value function.

The values $\mathbf{V}$ are updated through online learning by the following procedures: For each step t , the system generates an action to execute d_e^t based on the ϵ -greedy strategy. That is, the best response that has the minimum value is selected for a probability $1 - \epsilon$, and responses is randomly selected for a probability ϵ , which was set to 0.2. Value parameters $\mathbf{V}_{a_e}$ of the selected response a_e^t were updated using the temporal difference (TD) algorithm:

$$\begin{aligned} V_{a_e}^{n\lambda (t+1)} &= V_{a_e}^{n\lambda (t)} + \delta \text{TDError} \frac{\partial Q(S, a_e)}{\partial V_{a_e}^{n\lambda}} \\ &= V_{a_e}^{n\lambda (t)} + \delta (R_{a_e} - Q(S, a_e)) \cdot \tau_n(S). \end{aligned} \quad (2.17)$$

Here, R_{a_e} denotes the actual reward/penalty for the selected response a_e . The parameters of λ , n and δ were empirically set to $\lambda = 1.5$, $n = 6$ and $\delta = 0.001$ based on the result of a preliminary experiment.

5. Evaluation of Online Learning Methods

We evaluated the online learning methods. The set of 1,416 utterances (1,084 information queries and 332 wh-questions) is used in this evaluation. We trained the dialogue strategy by optimizing the parameters. We evaluated the improvement by using a 10-fold cross validation by splitting the utterance set into ten (set-1, $\dots$, set-10), that is, one set was used as a test set to evaluate performance, and the other nine were used as training data. Since the method using RL has a random factor in the ϵ -greedy strategy, the result of the method is an average of the 10 trials.

The evaluation measures were the success rate and the average number of sentences for information access. We regard a retrieval as successful if the system presented (or confirmed to present) the appropriate response for the utterance. The actual reward/penalty ARP (or R) is obtained by assigning 1 into $p(d_i|\mathbf{W})$ of equations in Section 3.4 for success and 0 for failure, for the response candidates⁹. We rewarded correct presentation by 10 ($Rwd_{Ret} = 10$) and correct question-answering by 30 ($Rwd_{QA} = 30$) considering difference in the number of samples in the test set. The FP was set to 6 based on typical recovery patterns observed in the field trial (Misu and Kawahara, 2007). All parameters (Θ or $\mathbf{V}$) were initialized to zero.

Figure 8 shows the relationship between the number of steps t for learning and the success rate of information access, and Figure 9 shows the relationship between the number of steps t and the average number of expected ARP per query obtained by the step t strategy at that time¹⁰. A small number of ARP implies a better dialogue strategy. By using Bayes risk-based strategy using ML estimation using Fisher’s scoring algorithm (BRML), we could achieve a significantly higher performance than that of the steepest descent method (BRSD). The eventual performance of BRML was almost comparable to that of the method using RL with grid points (RLG).

We then evaluated the performance in terms of convergence speed. The BRML was converged very quickly with almost 50 samples. This convergence speed is one of the advantages of the BRML method such as when developing a new system or adapting it to changes in the tendency in the data. In contrast, the convergence speed of the RLG was considerably slower than that of the BRML requiring 500 steps. Of course other techniques, such as the natural gradient approach (Peters and Schaal, 2008) may improve the speed, but training by RLG requires a large number of iterations, especially when dealing with a continuous state space. One reason for this is that the RLG considers each response action as an independent one using no a priori knowledge about the dependency between responses. However, this assumption is not true (at least between “Presentation”, “Confirmation” and “Rejection”). For example, if the system is rewarded by “Confirmation”, it is supposed to obtain a better reward

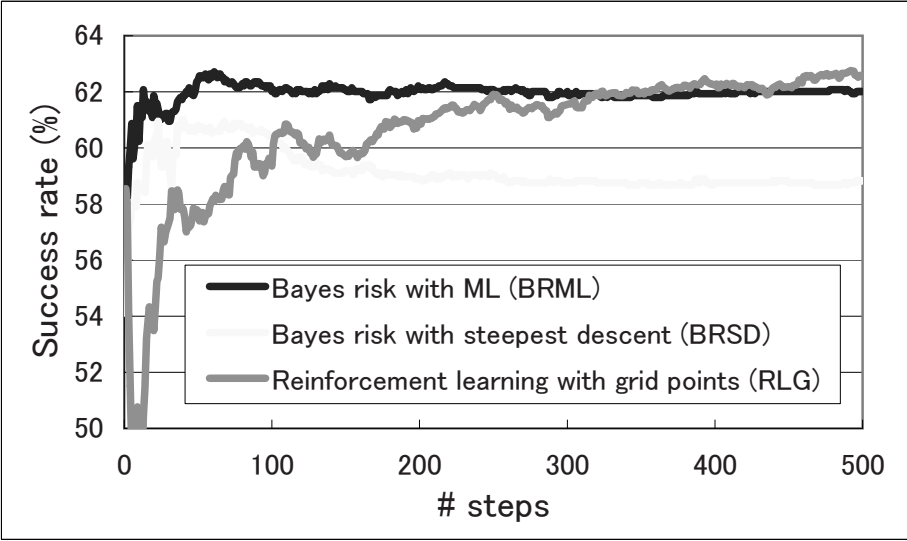


Figure 8. Number of step vs. Success rate of information access.

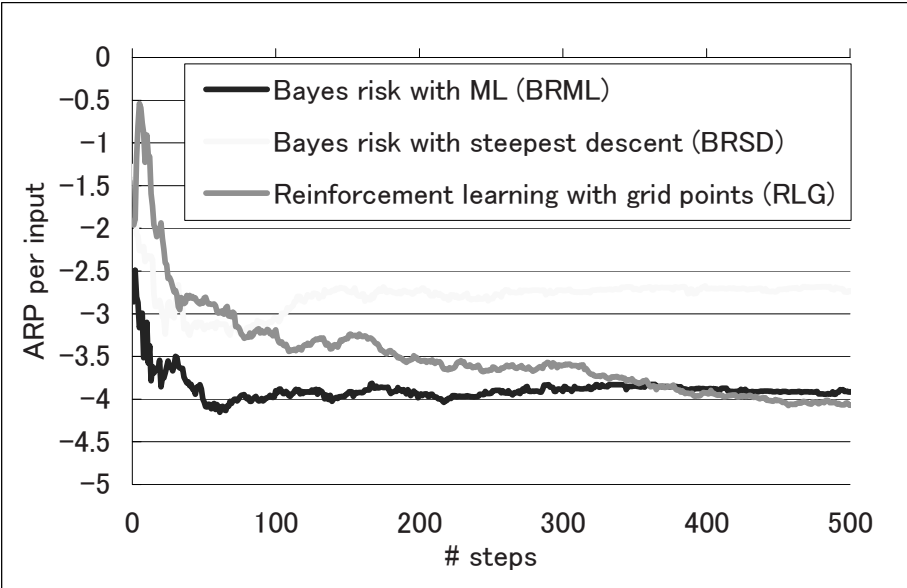


Figure 9. Number of step vs. Actual reward/penalty (ARP).

by “Presentation”. In contrast, if it is penalized by confirmation, the penalty is supposed to be less with rejection. In the method using Bayes risk, the values

of responses are optimized simultaneously in one step via the estimation of the success rate of retrieval. Thus, the method can estimate the risk of response with a fewer number of parameters. For these reasons, we consider that the training by BRML was converged with a small number of steps.

The target of the optimization in BRML is parameters of the logistic sigmoid function that estimate posterior probability of success, and it does not depend on the values of reward and penalty. This means that the optimality of the dialogue strategy by the proposed method is guaranteed. For example, if we replace the rewards by $Rwd_{Ret} = 0$ and $Rwd_{QA} = 0$, we will obtain a dialogue strategy that minimize the number of sentences in all turns before the appropriate information is obtained without parameter re-tuning. This property is an important advantage over the other approaches that require the whole training process using the re-tuned parameters.

6. Conclusions

We have proposed an online learning method of dialogue framework to generate an optimal response based on Bayes risk. Experimental evaluations by real user utterances demonstrated that the optimal dialogue strategy can be obtained with a small number of training samples. Although we only implemented and evaluated a simple explicit confirmation that asks the user whether the retrieved document is a correct one or not, the proposed method is expected to incorporate more various responses in document retrieval tasks, such as a clarification request and an implicit confirmation.

We used only two parameters of the matching score and bias for the logistic regression (Equation (2.12)) to estimate the success rate, but this can be easily extended to incorporate various feature parameters, such as a difference of score (margin) with the second best candidate or a system's previous response. The Bayes risk-based strategies presented in this chapter assume a sooner reward, and cannot be directly applied to a dialogue task where a reward/penalty is given as a delayed reward. However, we can optimize the entire dialogue by introducing a cumulative future reward and the optimization process of WF-STs (Hori et al., 2008).

Notes

1. <http://trec.nist.gov/data/qamain.html>
2. This method is expected to realize high precision for our task, but high recall is not expected. We thus back off inputs to the information query mode in case they are not classified as questions.
3. *Bunsetsu* is defined as a basic unit of Japanese grammar and it consists of a content word (or a sequence of nouns) followed by function words. We conducted the dependency structure analysis on all sentences in the knowledge base.
4. Demonstratives (e.g. "it") are typically omitted in Japanese. And, the utterance U2 in Figure 3 usually becomes "When was build".
5. We adopted an explicit confirmation such as "Do you want to know the *document's title*?"

6. This corresponds to a logistic regression of the success rate.
7. We regard a retrieval as successful if the system presented (or confirmed) the appropriate document/NE for the query.
8. In this task, a reward or a penalty is given as a sooner reward. This problem corresponds to a multi-armed bandit problem with a continuous state space.
9. These values were calculated using the manually labeled correct responses.
10. The response with the minimum risk is selected in the Bayes risk-based strategies and the response with the minimum value is selected in the strategy using RL.

References

- Bohus, D., Langner, B., Raux, A. Black, A., and Rudnicky, M. E. A. (2006). Online Supervised Learning of Non-Understanding Recovery Policies. In *Proceedings of IEEE/ACL Workshop on Spoken Language Technology (SLT)*, pages 170–173.
- Boni, M. D. and Manandhar, S. (2005). Implementing Clarification Dialogues in Open Domain Question Answering. *Natural Language Engineering*, 11(4):343–361.
- Brøndsted, T., Larsen, L., Lindberg, B., Rasmussen, M., Tan, Z., and Xu, H. (2006). Distributed Speech Recognition for Information Retrieval on Mobile Devices. In *Proceedings of Workshop on Speech in Mobile and Pervasive Environments (SiMPE)*.
- Chen, B., Chen, Y., Chang, C., and Chen, H. (2005). Speech Retrieval of Mandarin Broadcast News via Mobile Devices. In *Proceedings of Interspeech*, pages 109–112.
- Dohsaka, K., Yasuda, N., and Aikawa, K. (2003). Efficient Spoken Dialogue Control Depending on the Speech Recognition Rate and System’s Database. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*.
- Hori, C., Ohtake, K., Misu, T., Kashioka, H., and Nakamura, S. (2008). Dialog Management using Weighted Finite-State Transducers. In *Proceedings of Interspeech*, pages 211–214.
- Horvitz, E. and Paek, T. (2006). Complementary Computing: Policies for Transferring Callers from Dialog Systems to Human Receptionists. *User Modeling and User Adapted Interaction*, 17(1-2):159 – 182.
- Kim, D., Sim, H., Kim, K., Kim, J., Kim, H., and Sung, J. (2008a). Effects of User Model on POMDP-based Dialogue Systems. In *Proceedings of Interspeech*, pages 1169–1172.
- Kim, K., Lee, C., Jung, S., and Lee, G. (2008b). A Frame-based Probabilistic Framework for Spoken Dialog Management using Dialog Examples. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.
- Komatani, K., Ueno, S., Kawahara, T., and Okuno, H. (2005). User Modeling in Spoken Dialogue Systems to Generate Flexible Guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183.

- Kurita, T. (1994). Iterative Weighted Least Squares Algorithms for Neural Networks Classifiers. *New Generation Computing*, 12:375–394.
- Lamel, L., Bennacef, S., Gauvain, J.-L., Dartigues, H., and Temem, J. N. (2002). User Evaluation of the MASK Kiosk. *Speech Communication*, 38(1).
- Lamel, L., Rosset, S., Gauvain, J., and Bennacef, S. (1999). The LIMSI ARISE System for Train Travel Information. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Lee, A., Shikano, K., and Kawahara, T. (2004). Real-Time Word Confidence Scoring using Local Posterior Probabilities on Tree Trellis Search. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing ICASSP*, pages 793–796.
- Lemon, O. and Pietquin, O. (2007). Machine Learning for Spoken Dialogue Systems. In *Proceedings of Interspeech*, pages 247–255.
- Levin, E. and Pieraccini, R. (2006). Value-based Optimal Decision for Dialog Systems. In *Proceedings of IEEE/ACL Workshop on Spoken Language Technology (SLT)*, pages 198–201.
- Levin, E., Pieraccini, R., and Eckert, W. (2000). A Stochastic Model of Human-machine Interaction for Learning Dialog Strategies. *IEEE Trans. on Speech and Audio Processing*, 8:11–23.
- Matsuda, M. and Fukumoto, J. (2006). Answering Question of IAD Task using Reference Resolution of Follow-up Questions. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pages 414–421.
- Misu, T. and Kawahara, T. (2006a). A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proceedings of Interspeech*, pages 9–12.
- Misu, T. and Kawahara, T. (2006b). Dialogue Strategy to Clarify User's Queries for Document Retrieval System with Speech Interface. *Speech Communication*, 48(9):1137–1150.
- Misu, T. and Kawahara, T. (2007). Speech-based Interactive Information Guidance System using Question-Answering Technique. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Misu, T. and Kawahara, T. (2009). Bayes Risk-based Dialogue Management for Document Retrieval System with Speech Interface. *Speech Communication*, 52(1):61–71.
- Murata, M., Utiyama, M., and Isahara, H. (2006). Japanese Question-Answering System for Contextual Questions using Simple Connection Method, Decreased Adding with Multiple Answers, and Selection by Ratio. In *Proceedings of Asian Information Retrieval Symposium*, pages 601–607.

- Niimi, Y. and Kobayashi, Y. (1996). A Dialog Control Strategy Based on the Reliability of Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*.
- NIST and DARPA (2003). The twelfth Text REtrieval Conference (TREC 2003). In *NIST Special Publication SP 500-255*.
- Pan, Y. and Lee, L. (2007). TYPE-II Dialogue Systems for Information Access from Unstructured Knowledge Sources. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 544–549.
- Peters, J. and Schaal, S. (2008). Natural Actor-Critic. *Neurocomputing*, 71(7-9):1180–1190.
- Potamianos, A., Ammicht, E., and Kuo, H. (2000). Dialogue Management in the Bell Labs Communicator System. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*.
- Raux, A., Langner, B., Black, A., and Eskenazi, M. (2005). Let's Go Public! Taking a Spoken Dialog System to the Real World. In *Proceedings of Interspeech*.
- Ravichandran, D. and Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting of the ACL*.
- Reithinger, N. and Sonntag, D. (2005). An Integration Framework for a Mobile Multimodal Dialogue System Accessing the Semantic Web. In *Proceedings of Interspeech*.
- Rosset, S., Galibert, O., Illouz, G., and Max, A. (2006). Integrating Spoken Dialog and Question Answering: the Ritel Project. In *Proceedings of Interspeech*, pages 1914–1917.
- Roy, N., Pineau, J., and Thrun, S. (2000). Spoken Dialogue Management using Probabilistic Reasoning. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 93–100.
- Rudnicky, A., Bennett, C., Black, A., Chotomongcol, A., Lenzo, K., Oh, A., and Singh (2000). Tasks and Domain Specific Modelling in the Carnegie Mellon Communicator System. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, volume 2.
- Seneff, S. and Polifroni, J. (2000). Dialogue Management in the Mercury Flight Reservation System. In *Proceedings of ANLP-NAACL 2000, Satellite Workshop*.
- Singh, S., Litman, D., Kearns, M., and Walker, M. (2002). Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJ-Fun System. *Journal of Artificial Intelligence Research*, 16:105–133.
- Sturm, J., Os, E., and Boves, L. (1999). Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE system. In *Proceedings of ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*.

- T. Kudo, Y. M. (2003). Fast Methods for Kernel-Based Text Analysis. In *Proceedings of the 42nd Annual Meeting of the ACL*.
- Williams, J. and Young, S. (2007). Scaling POMDPs for Spoken Dialog Management. *IEEE Trans. on Speech and Audio Processing*, 15(7):2116–2129.
- Young, S., Schatzmann, J., Weilhammer, K., and Ye, H. (2007). The Hidden Information State Approach to Dialog Management. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., and Hetherington, L. (2000). Jupiter: A Telephone-based Conversational Interface for Waeather Information. *IEEE Trans. on Speech and Audio Processing*, 8(1):85–96.

Chapter 3

TOWARDS FINE-GRAIN USER-SIMULATION FOR SPOKEN DIALOGUE SYSTEMS

Ramón López-Cózar

*Dept. of Languages and Computer Systems, University of Granada
Granada, Spain*

rlopezc@ugr.es

David Griol

*Dept. of Computer Science, University Carlos III de Madrid
Madrid, Spain*

dgriol@inf.uc3m.es

Gonzalo Espejo, Zoraida Callejas, Nieves Ábalos

*Dept. of Languages and Computer Systems, University of Granada
Granada, Spain*

gonaep@correo.ugr.es, zoraida@ugr.es, nayade@correo.ugr.es

Abstract

Continuous advances in the field of spoken dialogue systems make the processes of design, implementation and evaluation of these systems more and more complex. To solve problems emerging from this complexity, a technique which has attracted increasing interest during the last decades is based on the automatic generation of dialogues between the system and a user simulator, which is another system that represents human interactions with the dialogue system. This chapter describes the main methodologies and techniques developed to create user simulators, and presents a discussion of their main characteristics and the benefits that they provide for the development, improvement and assessment of this kind of systems. Additionally, we propose a user simulation technique to test the performance of spoken dialogue systems. The technique is based on a novel approach to simulating different levels of user cooperativeness, which

allows carrying out a more detailed system assessment. In the experiments we have evaluated a spoken dialogue system designed for the fast food domain. The evaluation has focused on the performance of the speech recogniser, semantic analyser and dialogue manager of this system. The results show that the technique provides relevant information to obtain a solid evaluation of the system, enabling us to find problems in these modules which cannot be observed taking into account just one cooperativeness level.

Keywords: User modelling; Evaluation methodologies.

1. Introduction

The design of dialogue systems is a complex task that generally requires the use of expert knowledge acquired in the development of previous systems, including tests taken with users interacting with the system. The development of these systems is usually an iterative process in which different prototypes are released and tested with real users (Nielsen, 1993). From these tests, objective and subjective measures can be obtained concerning the appropriateness of several aspects of the system (Möller, 2004). The tests provide a basis for refinements of the prototype systems until eventually a system is obtained which is as perfect as possible in terms of correct functioning and user satisfaction. However, employing user studies to support the development process is very expensive and time consuming. The use of techniques like Wizard of Oz (Dow et al., 1986; Carbinì et al., 2006) for system design reduces the costs by avoiding the need for a functional prototype for testing purposes, but unfortunately it does not avoid the need for constant human intervention to obtain useful results. For these reasons, during the last decade many research groups have been attempting to find a way to automate these processes, leading to the appearance of the first user simulators. These simulators are automatic systems that represent human interactions with the dialogue system to be tested.

Research in techniques for user modelling has a long history within the fields of natural language processing and spoken dialogue systems. The main purpose of a user simulator is to improve the main characteristics of a spoken dialogue system through the generation of corpora of interactions between the system and the simulator (Möller et al., 2006). Collecting large samples of interactions with real users is an expensive process in terms of time and effort. Moreover, each time changes are made to the system it is necessary to collect more data in order to evaluate the changes. The user simulator makes it possible to generate a large number of dialogues in a very simple way. Therefore, these techniques contribute positively to the development of dialogue systems, reduce the time and effort that would be needed for their evaluation and also allow to adapt them to deal with individual user needs and preferences.

Simulated data can be used to evaluate different aspects of a dialogue system, particularly at the earlier stages of development, or to determine the effects of changes to the system's functionalities. For example, in order to evaluate the consequences of the choice of a particular confirmation strategy on transaction duration or user satisfaction, simulations can be done using different strategies and the resulting data can be analyzed and compared. Another example is the introduction of errors or unpredicted answers in order to evaluate the capacity of the dialogue manager to react to unexpected situations.

A second usage is to support the automatic learning of optimal dialogue strategies using reinforcement learning, given that large amounts of data are required for a systematic exploration of the dialogue state space. Corpora of simulated data are extremely valuable for this purpose, considering the costs of collecting data from real users. In any case, it is possible that the optimal strategy may not be present in a corpus of dialogues gathered from real users, so additional simulated data may enable additional alternative choices in the state space to be explored (Schatzmann et al., 2005).

According to the level of abstraction at which the dialogue is modelled, it is possible to find in the literature user simulators working at the word level (López-Cózar et al., 2003; Araki et al., 1997; Pietquin and Beaufort, 2005) or the intention level (Griol et al., 2008; Jung et al., 2009). Working at the word level, the input to the simulator is either the words in text format or the user utterances. Words in text format allow testing the performance of the spoken language understanding (SLU) component of the dialogue system, and that of the dialogue manager in dealing with ill-formed sentences. Using utterances (voice samples files) enables deeper checking of the robustness of the system. For example, it allows testing the performance of techniques at the ASR level to deal with noisy conditions. It also enables to test the performance of the SLU module in dealing with ASR errors, and that of the dialogue manager in dealing with SLU errors.

If the simulator works at the intention level, it receives as input abstract representations of the semantics of sentences, for example frames (Eckert et al., 1997; Levin et al., 2000; Scheffler and Young, 2002; Scheffler and Young, 2001). Hence, it is not possible to check the performance of the speech recogniser or the SLU component, but only that of the dialogue manager. This strategy is useful, however, to address the problem of data sparseness and to optimise dialogue management strategies.

In this chapter we summarize the main characteristics and advantages of user simulation techniques and present a technique to enhance a rule-based user simulator previously developed by including different levels of user cooperativeness. Providing a fine-grained scale of user cooperativeness makes it possible to carry out a detailed evaluation of the performance of a spoken dia-

logue system, assessing its ability to deal with responses that do not necessarily match every system prompt.

Our study has been evaluated by means of a Spanish dialogue system called Saplen (López-Cózar et al., 1997), designed to answer customers queries and register product orders in a fast-food restaurant. Our user simulator has been used to improve the system by identifying problems in the performance of the speech recogniser, semantic analyser and dialogue manager. Moreover, the evaluation results provide valuable information about how to best tune the dialogue management strategies and language models for speech recognition to meet the needs of real users.

The remainder of this chapter is organised as follows. Section 2 discusses previous studies on user simulator for spoken dialogue systems. Section 3 presents our two user simulators: the initial one and an enhanced simulator which implements a fine-grained scale user cooperativeness to better evaluate spoken dialogue systems. Section 4 presents the experiments carried out employing the enhanced simulator to evaluate the performance of the Saplen dialogue system. Section 5 discusses the experimental results, the findings by employing three types of user cooperativeness, and possibilities for future work. Finally, Section 6 presents the conclusions.

2. Related Work

The implementation of user simulators has been carried out using mainly two techniques: rule-based methods (Chung, 2004; Komatani et al., 2003; Lin and Lee, 2001; López-Cózar et al., 2003) and corpus-based approaches (Schatzmann et al., 2006; Griol et al., 2008). There are also in the literature hybrid techniques which combine features of these two approaches. For example, Cuayáhuatl et al. (2005) present a technique that combines a goal-oriented method with the bigram model using (Hidden Markov Models) HMMs, whereas Georgila et al. (2005) extend this study using n-grams.

It is also possible to classify the different approaches with regard to the level of abstraction at which they model dialogue. This can be at the acoustic level, the word level or the intention level. The latter is a particularly useful compressed representation of human-computer interaction. Intentions cannot be observed, but they can be described using speech-act and dialogue-act theory (Searle, 1969; Traum, 1999; Bunt, 1981). For dialogue modelling, simulation at the intention level is the most convenient, since the effects of recognition and understanding errors can be modelled and the intricacies of natural language generation can be avoided (Young, 2002).

In this section we explain the main features of rule-based and corpus-based approaches and discuss a number of user simulators representative of each

type. The end of the section discusses issues concerning the evaluation of user simulators.

2.1 Rule-based User Simulators

Using the rule-based approach, the designer of a user simulator creates a set of rules that decide the behaviour of the simulator. The advantage of this approach is in the certainty of the reactions of the simulator, which enables the designer to have complete control over the experiments. An initial example can be found in the study presented in (Araki et al., 1997), which proposes to evaluate spoken dialogue systems using a system-to-system interaction method that automatically generates dialogues. An additional system, called coordinator, includes linguistic noise in the interaction in order to simulate speech recognition errors in the communication channel, as can be observed in Figure 1. The evaluation measured the system's robustness in sorting out problems in the communication, focusing on task completion and the number of dialogue turns per dialogue for a given word error rate.

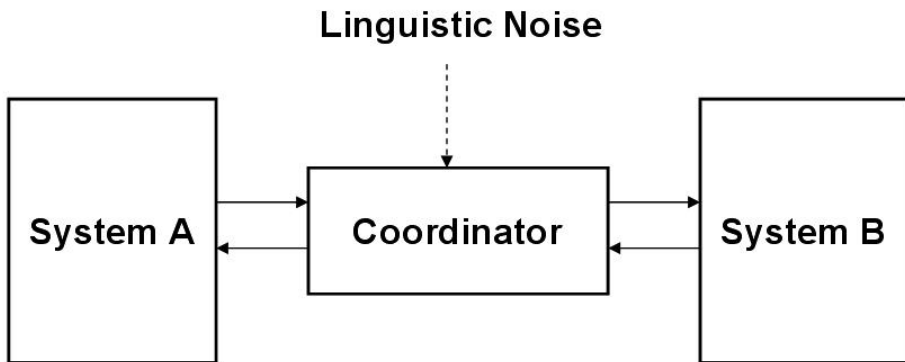


Figure 1. Concept of system-to-system dialogue with linguistic noise (Araki et al., 1997).

Another technique, described in (Chung, 2004), allows two types of simulated output (text and speech) receiving as input simulated semantic frames (Figure 2). In the experiments carried out in the restaurant information domain, the authors generated 2,000 dialogues in text mode, which were particularly useful for extending the coverage of the NL parser, and to diagnose problems overlooked in the rule-based mechanisms for context tracking.

In order to check the n-gram language models employed by the speech recogniser (bigrams and trigrams), the authors generated 36 dialogues in speech mode. Of these dialogues, 29 were completed without errors, with the correct desired data set achieved. Regarding the erroneous dialogues, three of them showed problems due to ASR errors, whereas four presented errors with

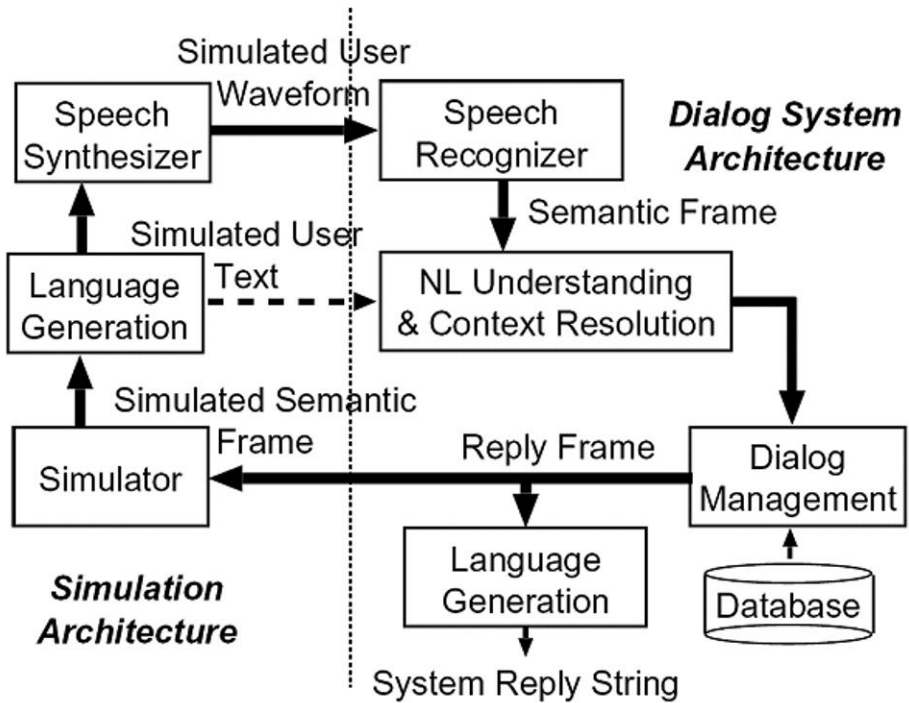


Figure 2. SDS integrated with user simulator (Chung, 2004).

the parsing and the context tracking mechanisms. The simulator was applied to provide restaurants' multiple constraints in single turns, for which typically empty data sets were retrieved. In earlier development cycles, these experiments were crucial to find combinations of constraints that yielded problematic responses of the system.

Filisko and Seneff (2006) propose an advanced rule-based user simulator that could be configured by the developer in order to show different behaviours in the dialogues. The authors were interested in checking error recovery mechanisms of a SDS working in the flight reservation domain. The goal was to acquire out-of-vocabulary city names by means of subdialogues in which the user had to speak-and-spell city names, e.g., "Nelson N E L S O N". In order to make its responses as realistic as possible, the simulator combined segments of user utterances available from a speech corpus for the application domain. To combine the segments, the simulator employed a set of utterance templates. This way, when generating a response, it chose one template, filled the existing gaps (for example, departure and destination cities) and produced the complete sentence (Figure 3).

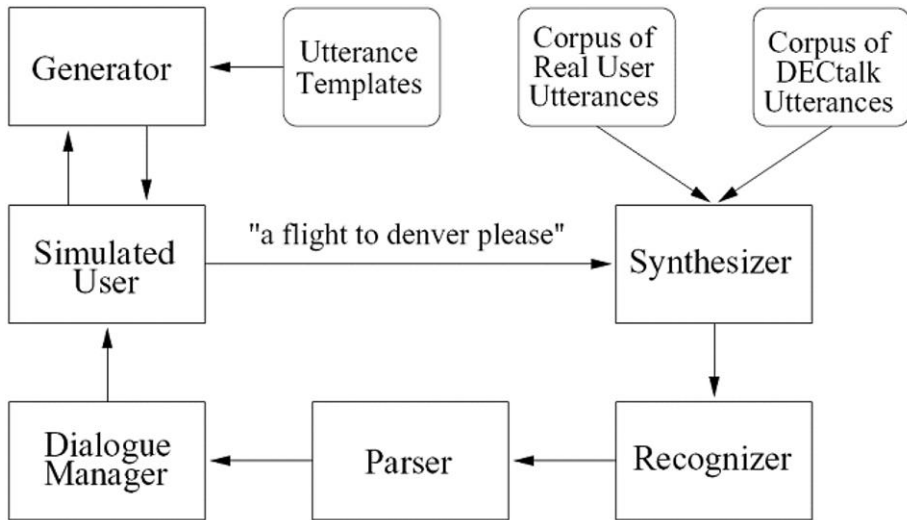


Figure 3. Generation of a user simulator's response (Filisko and Seneff, 2006).

2.2 Corpus-based User Simulators

Using the corpus-based (or data-based) approach, the simulator uses probabilistic methods to generate a response for each system's prompt, with the advantage that the uncertainty of the method can better reflect the unexpected behaviours of users interacting with the system. The advantage of this approach lies in its simplicity and in that it is totally domain independent. The main disadvantage, however, is that it may be too limited to give a realistic simulated behaviour because, although user actions are dependent on the previous system action, they should also be consistent throughout the dialogue as a whole.

Statistical models of user behaviour have been suggested as the solution to the lack of data when training and evaluating dialogue strategies. Using this approach, the dialogue manager can explore the space of possible dialogue situations and learn new potentially better strategies. The most extended methodology for machine-learning of dialogue strategies consists of modeling human-computer interaction as an optimization problem using Markov Decision Process (MDP) and reinforcement learning methods.

Eckert et al. (1997) introduce the use of statistical models to predict the next user action by means of an n-gram model. The proposed model has the advantage of being both statistical and task-independent. Its main weakness is that it approximates the complete history of the dialogue by only a bigram model. In (Levin and Pieraccini, 1997), the bigram model was modified by

considering only a set of possible user answers following a given system action (the Levin model). Both models have the drawback of considering that every user response depends only on the previous system turn. Therefore, the user simulator can change objectives continuously or repeat information previously provided.

In the case of advanced dialogue systems, the possible paths through the dialogue state space are not known in advance and the specification of all possible transitions is not possible. Partially Observable MDPs (POMDPs) outperform MDP-based dialogue strategies since they provide an explicit representation of uncertainty. However, the state space is too big for an exact POMDP optimization and currently there are no methods for exhaustively searching the complete state space of a dialogue system in which the state space is emergent rather than predetermined. This issue has been addressed by constraining the state space to a manageable size and by focusing on task-oriented systems in which the goal is to elicit a finite (generally fairly small) set of values from the user to fill the slots in a form.

One possible way to address some of these issues is to collect and analyze vast amounts of data covering the different ways in which users interact with a system and the different choices that can be applied in dialogue management. However, controlling all these factors with real users in actual interactions would be a daunting, if not impossible task. A more efficient method for collecting data under controlled conditions would be to simulate interactions in which the various user and system factors can be systematically manipulated.

Scheffler and Young (1999) propose a graph-based model. The arcs of the network symbolize actions, and each node represents user decisions (choice points). In-depth knowledge of the task and great manual effort are necessary for the specification of all possible dialogue paths. Pietquin and Beaufort (2005) combined characteristics of the models proposed in (Scheffler and Young, 1999) and (Levin and Pieraccini, 1997). The main objective was to reduce the manual effort necessary for the construction of the networks. A Bayesian network was suggested for user modelling. All model parameters were hand-selected.

Georgila et al. (2005) propose to use HMMs, defining a more detailed description of the states and considering an extended representation of the history of the dialogue. Dialogue is described as a sequence of Information States (Bos et al., 2003). Two different methodologies are described to select the next user action given a history of information states. The first method uses n -grams (Eckert et al., 1997) with values of n from 2 to 5 to consider a longer history of the dialogue. The best results were obtained with 4-grams. The second methodology is based on the use of a linear combination of 290 characteristics to calculate the probability of every action for a specific state.

Georgila et al. (2006) present a data-based user simulator which takes into account advanced n-grams to consider that user actions are conditioned by the *current status* of tasks, and not only by speech acts and tasks. One example of this type of current status is whether or not the information needed to perform a given action has been confirmed. The simulator actions were decided considering a probability distribution learned in the training. The authors carried out experiments with the Communicator 2001 corpus (dialogues concerning flights, hotels, and car reservation). The goal of the evaluation was to measure how “human-like” the behaviour of the simulator was, considering as measures the expected accuracy, precision, recall and perplexity.

Cuayáhuitl et al. (2005) propose a method for dialogue simulation based on HMMs in which both user and system behaviours are simulated. Instead of training only a generic HMM model to simulate any type of dialogue, the dialogues of an initial corpus are grouped according to the different objectives. A submodel is trained for each one of the objectives, and a bigram model is used to predict the sequence of objectives.

In (Schatzmann et al., 2007a; Schatzmann et al., 2007b), a technique for user simulation based on explicit representations of the user goal and the user agenda is presented. The user agenda is a structure that contains the pending user dialogue acts that are needed to elicit the information specified in the goal. This model formalizes human-machine dialogues at a semantic level as a sequence of states and dialogue acts. An EM-based algorithm is used to estimate optimal parameter values iteratively.

Wang et al. (2005) address the issue of generating language model training data during the initial stages of dialogue system development by means of a user simulator. The proposed simulation technique obtains the probability model via an interplay between a probabilistic user model and a dialogue system that answers queries for a restaurants information domain. The main goal was to find a way to acquire language model training material in the absence of any in-domain real user data. The dialogues that were obtained by means of the interaction between the user simulator and the dialogue system are then used to acquire adequate coverage of the possible syntactic and semantic patterns to train both the recognizer and the natural language system. Experimental results verify that the resulting data from user simulation runs are much more refined than the original data set, both in terms of the semantic content of the sentences (i.e., different types of queries) as well as the probability distribution of the within-class values (e.g., cuisine types, neighbourhood names, etc.).

Griol et al. (2008) present a statistical approach to develop a user simulator for learning optimal dialogue strategies. The user simulator replaces the functions performed in a dialogue system by the ASR and the NLU modules, as shown in Figure 4. The user answers are generated by taking into account the information provided by the simulator throughout the history of the

dialogue (user register), the last system turn, and the objective(s) predefined for the dialogue. A labelled corpus of dialogues is used to estimate the user model, which is based on a classification methodology. An error simulator is used to introduce errors and confidence measures from which it is possible to adapt the error simulator module to the operation of any ASR and NLU modules. The results of its application to learn a dialogue model for the DIHANA project (Griol et al., 2006) demonstrate that the coverage of the dialogue manager is increased by incorporating the successful simulated dialogues and that the number of unseen situations can be reduced. A study of the evolution of the strategy followed by the dialogue manager shows how it modifies its strategy by detecting new correct answers that were not defined in the initial strategy.

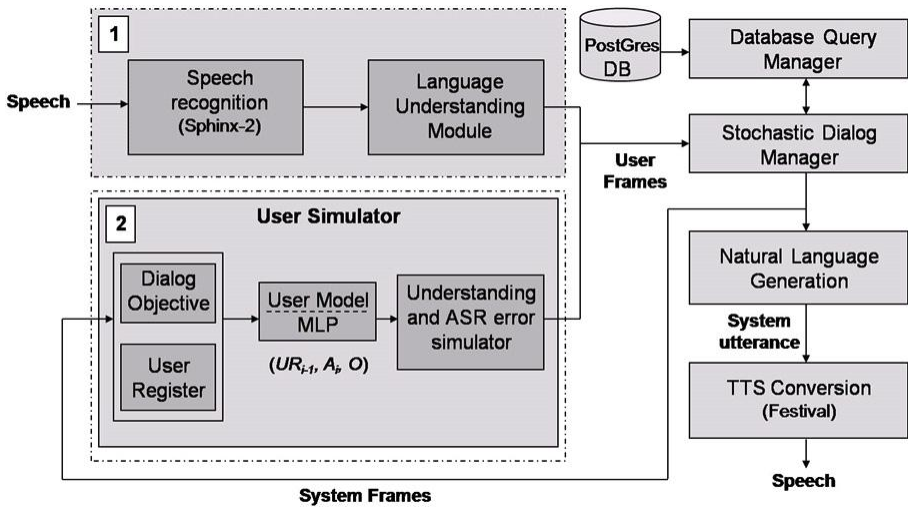


Figure 4. Architecture of the DIHANA dialogue system. (1) Interaction with real users. (2) Operation with the user simulator (Torres et al., 2008).

A data-driven user simulation technique for simulating user intention and utterance is introduced in (Jung et al., 2009). The user intention modeling and generating method uses a linear-chain conditional random field, and a two-phase data-driven domain-specific user utterance simulation method and a linguistic knowledge-based ASR channel simulation method. Different evaluation metrics were introduced to measure the quality of user simulation at intention and utterance. The main conclusions obtained from experimentation with a dialogue system for car navigation indicate that the user simulator was easy to set up and showed similar tendencies to real human users.

2.3 Hybrid User Simulators

As stated above, there also exist hybrid techniques which combine features of rule-based and corpus-based approaches. One example is in (Torres et al., 2008), where the user simulator interacts with a dialogue system that provides information about train schedules and other services (Bonafonte et al., 2000). The system's dialogue manager uses a stochastic dialogue model that is a bi-gram model of dialogue acts. Using this model, the dialogue manager selects the following state taking into account the last user turn and its current system state. The user simulator proposed in this work is a version of this dialogue manager, which has been modified to play the user role. It uses the same bi-gram model of dialogue acts. Using this model, the user simulator selects the following user action depending only on the last system action, as in (Eckert et al., 1997). Additional information (rules and restrictions that depend on the user goals) is included in the model to achieve the cooperation of the user and the consistency of the dialogues. Figure 5 shows the block diagram of the dialogue system extended with the user simulator modules. A user dialogue manager (UDM) and a user reply generator (URG) are the components of the user simulator. These modules are very similar to their corresponding modules on the system side (SDM and SRG).

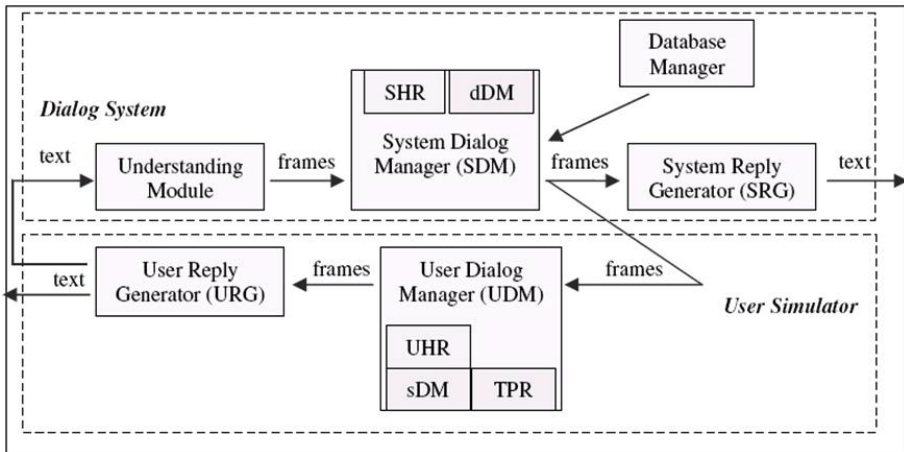


Figure 5. Architecture of the user simulator proposed to interact with the BASURDE dialogue system.(Torres et al. 2008).

2.4 Evaluation of User Simulators

There are no generally accepted criteria for what constitutes a good user simulator. Typically used methods are adopted from other research fields

such as Information Retrieval and Machine Learning. A first classification divides evaluation techniques into direct evaluation methods and indirect methods (Young, 2002).

2.4.1 Direct Methods. Direct methods evaluate the user simulator by measuring the *quality* of its predictions. Typically, the *Recall* measure has been used to take into account how many actions in the real response are predicted correctly, whereas the *Precision* measure has been used to consider the proportion of correct actions among all the predicted actions.

For example, Schatzmann et al. (2005) show the results of precision and recall obtained from the evaluation of different user simulators. The maximum values are located around 35%. One drawback of these measures is that they consider a high penalty for the actions that are unseen in the responses on the simulator, although they could be potentially provided by a real user.

Another example is (Scheffler and Young, 2001), which defines evaluation features at three dimensions: high-level features (dialogue and turn lengths), dialogue style (speech-act frequency; proportion of goal-directed actions, grounding, formalities, and unrecognized actions; proportion of information provided, reprovided, requested and rerequested), and dialogue efficiency (goal completion rates and times). The simulation presented in (Schatzmann et al., 2007a) was evaluated by testing the similarity between real and simulated data by means of statistical measures (dialogue length, task completion rate and dialogue performance).

In (Georgila et al., 2006), the use of *Perplexity* for the evaluation of user simulator is introduced. It determines whether the simulated dialogues contain sequences of actions that are similar to those contained in the real dialogues. In (Cuayáhuitl et al., 2005), the comparison between the simulated corpus and a corpus acquired with real users is carried out by training a HMM with each corpus and then measuring the similarity between the two corpora on the basis of the distance between the two HMM.

The aim of the work described in (Ai and Litman, 2008) is to extend the previous work to evaluate the extent to which state-of-the-art user simulators can mimic human user behaviours and how well they can replace human users in a variety of tasks. Schatzmann et al. propose a set of evaluation measures to assess the quality of simulated corpora. These evaluation measures have proven to be sufficient to discern simulated from real dialogues (Griol et al., 2009).

Since this multiple-measure approach does not offer an easily reportable statistic indicating the quality of a user simulation, Williams (2007) proposes a single measure for evaluating and rank-ordering user simulations based on the divergence between the simulated and real users' performance. In their study, they recruited human judges to assess the quality of three user simula-

tors. Judges were asked to read the transcripts of the dialogues between the computer tutoring system and the simulators and to rate the dialogues on a 5-point scale from different perspectives. They first assessed human judges' abilities in distinguishing real from simulated users, finding that it is hard for human judges to reach good agreement on the ratings. However, these ratings provide consistent ranking on the quality of the real and the user simulator models. They concluded that this ranking model can be used to quickly assess the quality of a new simulation model without manual efforts by ranking the new model against the traditional ones.

2.4.2 Indirect Methods. The main objective of indirect methods of evaluation is to measure the *utility* of the user simulator within the framework of the operation of the complete system. These methods try to evaluate the operation of the dialogue strategy learned by means of the simulator. This evaluation is usually carried out by verifying the operation of the new strategy through a new interaction with the simulator. Then, the initial strategy is compared with the learned one using the simulator. The main problem with this evaluation resides in the dependence of the acquired corpus on the user model.

For example, Schatzmann et al. (2005) present a series of experiments that investigate the effect of the user simulator on simulation-based reinforcement learning of dialogue strategies. The results indicate that the choice of the user model has a significant impact on the learned strategy. The results also demonstrate that a strategy learned with a high-quality user model generalizes well to other types of user models. Lemon and Liu (2007) extend this work by evaluating only one type of stochastic user simulation but with different types of users and under different environmental conditions. This study concludes that dialogue policies trained in high-noise conditions perform significantly better than those trained for low-noise conditions.

Ai et al. (2007) evaluate what kind of user simulator is suitable for developing a training corpus for using Markov Decision Processes (MDPs) to automatically learn dialogue strategies. Three different user simulators, which were trained from a real corpus and simulate the word level, were evaluated using a speech-enabled Intelligent Tutoring System that helps students understand qualitative physics questions. The first model, called Probabilistic Model (PM), is meant to capture realistic student behavior in a probabilistic way. The second model, called Total Random Model (TRM) ignores what the current question is or what feedback is given and randomly picks one utterance from all the utterances in the entire candidate answer set. The third model, called Restricted Random Model (RRM) differs from the PM in that given a certain tutor question and a tutor feedback, it chooses to give certain, uncertain, neutral, or mixed answer with equal probability. The results of their work suggest that with sparse training data, a model that aims to randomly explore more

dialogue state spaces with certain constraints can actually outperform a more complex model that simulates realistic user behaviours in a statistical way.

3. Our User Simulators

This section of the chapter focuses on our initial user simulator and its application to evaluate the performance of the Saplen dialogue system. It addresses improvements made in the simulator to create an enhanced version. This version implements a novel technique to carry out a detailed evaluation of the performance of spoken dialogue systems, assessing their ability to deal with responses that do not necessarily match every system prompt.

3.1 The Initial User Simulator

In a previous study (López-Cózar et al., 2003) we developed a user simulator which is the basis for the enhanced user simulator discussed in this chapter. The purpose of the initial simulator was to interact automatically with a spoken dialogue system in order to create a corpus of dialogues that could be used for testing the performance of the system. We carried out experiments with this simulator employing the Saplen dialogue system, previously designed in our lab to provide fast food information and register product orders (López-Cózar et al., 2002). As can be observed in [Figure 6](#), the simulator receives the current prompt generated by the dialogue system as well as the frame(s) obtained from the analysis of the previous response of the simulator. Each response of the simulator is an utterance (voice samples file) recorded by a client of the fast food restaurant, which is taken from a speech database.

The interaction between the simulator and the dialogue system is carried out by means of a set of scenarios that indicate the goals the simulator must try to achieve during the interaction. For example, a scenario may specify that the simulator must order one ham sandwich, one large beer and one chocolate milkshake, and provide the telephone number, postal code and address of a user. After the simulator has selected the appropriate scenario goal, e.g., `<POSTAL_CODE> = "18001"`, it retrieves an utterance from the speech database for which the associated frame matches the selected goal, e.g., "my postal code is 18001". This utterance is the response of the simulator.

In this version of the simulator, the generated responses always fulfilled the system expectations, because the level of user cooperativeness was the highest in all cases. For example, if the system prompted for the user's address, the simulator provided an address; if the system prompted for the user's telephone number, the simulator provided a telephone number and so on, as can be observed in the following sample dialogue:

- (1) Saplen: What would you like to have?
- (2) User simulator : Three ham sandwiches.

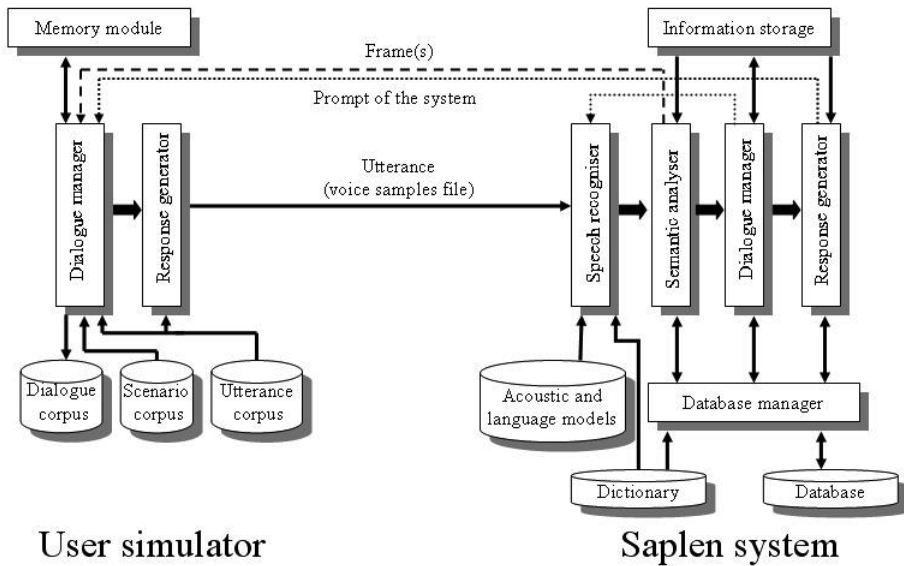


Figure 6. Interaction between the initial version of the user simulator and the Saplen system.

- (3) Saplen: I have to confirm your order. Did you say six ham sandwiches? Please say yes or no.
- (4) User simulator : No.
- (5) Saplen: OK. How many sandwiches did you say?
- (6) User simulator : Three.
- (7) Saplen: OK. So that's three ham sandwiches. Please say your telephone number.
- (8) User simulator : 9 5 8 2 7 5 3 6 0.
- (9) Saplen: I need to confirm your telephone number. Did you say 9, 5, 8, 2, 7, 5, 3, 6, 0? Please answer yes or no.
- (10) User simulator : Yes.
- (11) Saplen: OK. Telephone number 9, 5, 8, 2, 7, 5, 3, 6, 0. Please say your postal code.
- (12) User simulator: 1 8 0 0 1.
- (13) Saplen: OK. Postal code 1, 8, 0, 0, 1. Please say your address, for example, Elm street, number 13, second floor, apartment letter e.
- (14) User simulator : Andalucia street number 58 first floor letter e.
- (15) Saplen: I need to confirm your address. Did you say Andalucia street, number 58, first floor letter e? Please answer yes or no.
- (16) User simulator : Yes.
- (17) Saplen: OK. Address Andalucia street, number 58, first floor letter

e. Would you like to drink something?

(18) User simulator : No.

(19) Saplen: OK. I need to confirm your order. I think you ordered three ham sandwiches. Is this correct? Please answer yes or no.

(20) User simulator : Yes.

(21) Saplen: OK. The price is 12 euros. Is this OK for you? Please answer yes or no.

(22) User simulator : Yes.

(23) Saplen: OK. The delivery service will take 30 minutes approximately to get to your address. Is this OK for you? Please answer yes or no.

(24) User simulator : Yes.

(25) Saplen: OK. Thank you very much for using our automatic telephone service. Good night.

An advantage of this kind of interaction between the simulator and the dialogue system is that it makes it possible to estimate the system performance interacting with users who behave “optimally” as they provide exactly the data requested by the system in each interaction. In fact, this initial user simulator was useful to improve the performance of the Saplen system by identifying problems in the performance of the speech recogniser, semantic analyser and dialogue manager. Moreover, the evaluation results provided valuable information about how to best tune the dialogue management strategies and language models for speech recognition to meet the needs of real users (López-Cózar and Callejas, 2005). However, a drawback of the “optimal” behaviour exhibited by the simulator may not be very realistic.

3.2 The Enhanced User Simulator

In order to address this drawback of the initial simulator, we have developed an *enhanced user simulator* which shows a more realistic behaviour. To do so, this simulator implements three cooperativeness levels: *low*, *medium* and *high*. The *high* level corresponds to the behaviour of the initial user simulator. To implement the *medium* cooperativeness the simulator does not always respond with the kind of utterance requested by the dialogue system, although it always generates responses that are appropriate for the system prompts. For example, the simulator may repeat a piece of data when the system prompts to confirm the data using a *Yes/No* response. This behaviour can be observed in the following sample dialogue:

(1) Saplen: What would you like to have?

(2) User simulator: Three ham sandwiches.

(3) Saplen: I have to confirm your order. Did you say six ham sandwiches? Please say yes or no.

(4) User simulator: Three.

(5) Saplen: OK. So that's three ham sandwiches. Please say your telephone number.

This behaviour of the simulator can be useful to model experienced users who know how to make the interaction go faster by providing in advance the data that the system will require in subsequent prompts, e.g. uttering directly 'Three' instead of 'No' in turn (4). Simulating this type of user makes it possible to evaluate the ability and "intelligence" of the dialogue system to deal appropriately with responses that do not match every system prompt exactly.

To implement the *low* cooperativeness the simulator generates appropriate responses for some system prompts, but for others it generates responses that are completely inappropriate. This behaviour can be observed in the following sample dialogue:

(1) Saplen: What would you like to have?

(2) User simulator: Andalucia street number 58 first floor letter e.

(3) Saplen: I need to confirm your address. Did you say Andalucia street, number 58, first floor letter e? Please answer yes or no.

(4) User simulator: Yes.

(5) Saplen: OK. Address Andalucia street, number 58, first floor letter e. Would you like anything to eat?

The main advantage of the enhanced user simulator is that it makes it possible to model real users who because of inexperience or being nervous may answer some prompts with unexpected utterances. Therefore, this type of simulation allows us to test the robustness of the speech recogniser and semantic analyser to deal with these utterances. In addition, it can be useful to test the ability of the dialogue manager to appropriately handle data items that are not expected. For example, if the address data were provided unexpectedly in advance, the dialogue manager should not prompt for these data later.

4. Experiments

The goal of the experiments has been to study whether the dialogues generated employing the enhanced user simulator provide a better basis for the system assessment than the initial simulator.

The evaluation has been carried out in terms of word accuracy (WA), sentence understanding (SU) and task completion (TC).

WA is the proportion of correctly recognised words, which is computed taking into account the number of words inserted, deleted and substituted by the speech recogniser. SU is the proportion of correctly understood sentences regardless of the possible speech recognition errors. In other words, we say that there is sentence understanding if the semantic representation obtained by

the semantic analyser is correct even though some words might have been incorrectly recognised. TC is the proportion of successful dialogues, i.e., the percentage of dialogues in which the simulator achieves all the scenario goals. In order to avoid excessively long dialogues between the system and the simulator, which would not be accepted by real users, the simulator cancelled the interaction with the system if the total number of turns (i.e. of system plus user simulator turns) exceeded a threshold set to 30 turns. The reason for using this maximum limit is that, taking into account the structure of the scenarios as well as the dialogue management strategy of the Saplen system, a dialogue without any error correction requires 15 turns in total to make the orders, provide the user data (post code, telephone number, and address) and answer the confirmation prompts generated by the system. Considering 30 as an interaction limit means that we permitted $30 - 15 = 15$ correction turns per dialogue, which we consider an ample margin for the user simulator to correct possible errors. Cancelled dialogues are not considered successful and thus decrease the TC rate.

4.1 Speech Database and Scenario Corpus

To carry out the experiments we have employed a speech database that we collected in a fast food restaurant and contains around 800 dialogues between clients and restaurant assistants. The database is comprised of 18 sentence types: product orders, telephone numbers, postal codes, addresses, queries, confirmations, amounts, food names, ingredients, drink names, sizes, flavours, temperatures, street names, building numbers, building floors, apartment letters, and error indications.

Selecting at random 5,500 client utterances among the 18 sentence types in the database, we have created two utterance corpora, one for training the language models and the other for testing them, ensuring that no training utterances were included in the testing corpus. Both corpora include the orthographic transcriptions of the utterances as well as their corresponding semantic representations (frames). One half of the utterances that the simulator employs to correct system errors and to confirm data are used for training and the other half are used for testing. These utterances are not used as scenario goals given that they are scenario-independent.

To automatically generate dialogues between the Saplen system and the proposed user simulator we have designed 50 scenarios. The scenario goals are selected by choosing frames at random in the test utterance corpus corresponding to product orders, telephone numbers, postal codes and addresses.

Employing the simulator we have generated 20 dialogues per each scenario, cooperativeness level and type of language model for speech recogni-

tion, which makes a total of $20 \times 50 \times 3 \times 2 = 6,000$ dialogues. These dialogues have been saved in log files for further evaluation.

4.2 Language Models for Speech Recognition

The Saplen system was configured to use two different kinds of language model for speech recognition: one was based on 17 prompt-dependent language models (PDLMs), in the form of word bigrams (Rabiner et al., 1996), whilst the other was based on one prompt-independent language model (PILM), also a word bigram. Both kinds of language model have been used in previous studies (López-Cózar and Callejas, 2005).

The problem with the PDLMs is that they provide very poor recognition results if the users respond to system prompts with a type of utterance that does not match the active grammar (e.g. an address when the system prompted for a telephone number). This happens because the utterances are analysed employing a grammar compiled from utterances of a different type. Therefore, this language model is not appropriate to provide users with a natural interaction that enables them to answer system prompts with utterances that do not strictly match what the system requires, which is something that they would probably do when interacting with a human operator.

Contrary to what happens with the PDLMs, the PILM permits the recognition of any kind of utterance within the domain, which helps to provide users with more flexible interaction. However, the accuracy is in general lower than with the PDLMs given that the vocabulary is much larger and there are many more types of utterance to be considered.

5. Results

Table 1 shows the average results obtained for the three levels of user cooperativeness and the two language models for speech recognition in terms of word accuracy (WA), sentence understanding (SU) and task completion (TC). As can be observed, better performance is achieved for higher cooperativeness levels regardless of the language model employed. The differences in the performance are more clearly observed when the PDLMs are employed. When the cooperativeness is *high* the simulator always provides responses that match the current system prompt. Therefore using the PDLMs each utterance is analysed employing the appropriate recognition grammar.

The scores decrease when the level of cooperativeness is *medium* or *low* given that for these conditions the simulator sometimes provides utterances that do not match the current system prompt. According to the results set out in the table, it can be said that the system should only employ the PDLMs if the cooperativeness of real users were *high*, thus achieving $TC = 70.56\%$. For the other cooperativeness levels the performance would be very poor.

Table 1. Evaluation results (in %) for the three levels of user cooperativeness employing PDLMs and PILM.

	PDLMs			PILMs		
	WA	SU	TC	WA	SU	TC
Cooperat. level						
High	90.05	85.18	70.56	75.40	57.86	11.67
Medium	70.56	70.76	21.67	76.60	55.71	5.56
Low	43.69	56.82	11.13	77.87	53.28	4.47

When the PILM is employed the values for WA and SU are similar for the three level of cooperativeness. The reason is that regardless of the level, the simulator responses always match the recognition grammar, as it is compiled from training utterances permitted for all the system prompts. Taking into account the results set out in the table, it can be said that the system performance is totally unacceptable for the PILM regardless of the cooperativeness level, since TC is 11.67% as the greatest.

These results show that the system should only employ the PDLMs if the cooperativeness of the real users were *high* since the system performance can be considered more or less acceptable (TC = 70,56%). The results also show that when the PILM is used the system performance is too poor for an interaction with real users regardless of the type of language model employed, since TC is too low (only 11.67% in the best case).

5.1 Detection of Problems in the Performance of the Dialogue System

The main objective of the experiments was to carry out an evaluation of the Saplen system in order to identify problems with the speech recogniser, semantic analyser or dialogue manager, to fix those and thus increase the system's robustness to deal with a variety of users. To do this we have considered the log files created during the dialogue between the simulator and the Saplen system, have focused on the dialogues with very low values for the evaluation measures, and have analysed these to find the reasons for the unacceptable system performance.

5.1.1 Findings for the *High* Cooperativeness Level. When the PDLMs are employed, the utterances that the simulator generates as responses always match the active speech recognition grammars, which cause WA to be quite high (90.05%). The 10% word error rate is caused by three factors. One is that some 'Yes/No' answers to confirmation prompts are misrecognised, e.g. the word 'sí' (yes) is sometimes substituted by the word 'te'. One possible solution to this problem might be to change the language model

so that it only allows recognising 'sí' or 'No', and ask the user explicitly to utter any of these two words.

Another reason is that there are problems recognising some addresses for which not all data items are recognised. The third reason is that there are many recognition errors if the speakers have strong southern Spanish accents, as they usually do not pronounce the final 's' of plural words, which causes the recognition of the singular form of substantives and adjectives instead of plurals. Given that these errors in the number correspondence do not affect the semantics of the utterances, most of the product orders are correctly understood even though some words are incorrectly recognised.

When the PILM is employed, there are also many speech recognition errors in the responses to system confirmation prompts, especially if these are uttered by speakers with strong southern Spanish accents. Given that these users omit the final 's' of plural words, as discussed above, the acoustic similarity increases and thus the word 'no' is often substituted by the word 'dos' (two), 'uno' (one) or 'error', whereas the word 'sí' (yes) is often substituted by the word 'seis' (six). Moreover in many cases the words 'sí' and 'no' are discarded by the semantic analyser of the system as their confidence scores are smaller than the lowest confidence threshold employed (set to 0.3). Because of these problems there are many repetitive confirmation turns to get data confirmed, which lengthens the dialogues and causes some of these to be cancelled as the interaction limit (30 turns) is reached before all the scenario goals are achieved. A possible solution to this problem is not to use the PILM for confirmations. Instead, the system should use the PDLM specific for confirmations and force the user utter either 'Yes' or 'No'.

5.1.2 Findings for the *Medium* Cooperativeness Level.

When the PDLMs are employed the WA for the *medium* cooperativeness is lower than for the *high* cooperativeness (70.56% vs. 90.05%). The reason is that in addition to facing the problems discussed in the previous section, in this case the Saplen system has to overcome the problem that the sentences uttered to answer confirmation prompts are not permitted by the active grammars.

If the cooperativeness is *medium*, the simulator answers confirmation prompts by repeating the data that the dialogue system is trying to confirm, although it always prompts for a 'Yes/No' response. The problem identified in the analysis is that the grammar employed to recognise responses to confirmation prompts was initially created considering only users who would utter either a confirmation, a negation or an error indication (i.e. high cooperativeness). Because of this, the system confirmation strategy employing the PDLMs fails for the medium cooperativeness as the responses of the simulator are not permitted by the grammar. On the contrary, product orders, telephone numbers, postal codes and addresses are more or less well understood, although

the errors in gender/number correspondences and those for some addresses discussed above also occur in these dialogues. As a consequence of all the problems the average TC employing the PDLs is 21.67%, which is obviously too low to consider the system performance acceptable for real users behaving with *medium* cooperativeness.

When the PILM is employed WA is 76.6%, which is very similar to that obtained for the *low* cooperativeness (77.87%) employing the same language model. One reason for this low rate is the large amount of errors in gender/number correspondences (e.g. 'verdes' (green) substituted by 'verde'). As commented above, these errors happen especially when the words are uttered by speakers with strong southern Spanish accents. This problem suggests that we should model plurals as pronunciation variants, as they do not have an influence on the concept accuracy.

Also the same problems detected for the *high* cooperativeness with the recognition of some addresses are found for the *medium* cooperativeness, which means that the system needs to employ extra turns to get and confirm all the data items in the addresses.

In addition we observe a problem in the confirmation strategy that is not observed for the *high* cooperativeness and that is particularly noticeable in the confirmation of telephone numbers. If the cooperativeness is *medium* the simulator confirms the data by repeating the telephone number instead of generating the 'Yes/No' response requested by the system. To have a telephone number correctly understood, the Saplen system requires on the one hand that all its digits are recognised with confidence scores greater than the higher confidence threshold (set to 0.5). On the other hand, the system requires an implicit confirmation from the user when it includes the recognised number if the prompt is to get the postal code.

According to the method employed to assign confidence scores to frame slots, the confidence score of a slot that contains a telephone number is the lowest confidence score of the digits. For example, the confidence score for the recognition hypothesis "nine (0.5684) five (0.9652) eight (0.5647) one (0.5894) two (0.6954) three (0.9654) three (0.4362) four (0.6584) five (0.5898)" would be 0.4362.

Because of all these factors, the system has problems confirming some telephone numbers, especially when these are uttered by speakers with strong southern Spanish accents. The reason is that employing a telephone number to confirm a telephone number tends to require another confirmation, given that it is likely that at least one digit is misrecognised or recognised with low confidence. This problem provokes repetitive dialogue turns that lengthens the dialogues, causing some of these to be cancelled as the interaction limit is reached. Again, this problem might be addressed by using the specific lan-

guage model for confirmations, which allows recognising only either 'Yes' or 'No'.

5.1.3 Findings for the *Low* Cooperativeness Level. When the PDLMs are employed, the WA is very low (43.69%). One reason for this is the high number of errors in the confirmation turns given that these users do not answer confirmation prompts with 'Yes/No' responses but they repeat the data the system is trying to confirm, as it happened with the *medium* cooperativeness. Another reason is that for these users the simulator selects at random the kind of utterance to answer system prompts to enter product orders, telephone numbers, postal codes or addresses. This decreases the average TC employing the PDLMs to 11.13%, which is obviously too low to consider the system performance acceptable for real users with low cooperativeness.

When the PILM is employed, the value of WA (77.87%) is very similar to that obtained for the other cooperativeness levels (75.4% and 76.6%), given that in the three cases the same kind of language model is employed throughout the whole dialogue regardless of the system prompt. Consequently, the SU rate (53.28%) is also similar to that for the other cooperativeness levels (57.86% and 55.71%).

As discussed before, the behaviour of the simulator for the *low* cooperativeness is very similar to that for the *medium* cooperativeness, with the difference that the former features a random selection of utterances to answer system prompts to enter product orders, telephone numbers, postal codes and addresses. Because of this difference, the interaction for the *low* cooperativeness reveals a problem in the semantic analyser of the Saplen system which is not observed for the other two levels: in some cases telephone numbers are correctly recognised but are understood as postal codes, while postal codes are correctly recognised but understood as telephone numbers.

The reason is that the system employs its current prompt to differentiate between both kinds of utterance. Therefore, when it prompts to get a telephone number, it considers that the recognised sequence of digits is a telephone number. Similarly, when it prompts for a postal code, it assumes that the sequence is a postal code, and when it prompts for a building number, it considers that the digit sequence is a building number.

This simple understanding method works well if the cooperativeness level is *high* or *medium* and the simulator produces the expected kind of utterance. However, if the cooperativeness level is *low*, the simulator may answer prompts to enter a telephone number with a product order, telephone number, postal code or address, which causes the possible confusion if the postal code is randomly selected.

To solve this problem it could be possible to include knowledge in the semantic rules about the different format of telephone numbers (nine digits) and

postal codes (five digits). This way, the semantic analyser could guess whether the utterance is a telephone number or a postal code regardless of the prompt, and the system could ask the user to confirm the guess if the utterance type does not match the current prompt.

5.2 Future Work

Future work to improve the proposed technique includes studying alternative methods to simulate more precisely the behaviour of real users. One possibility would be to set the level of user cooperativeness dynamically as the dialogue evolves. In the current set up this selection is made beforehand and the setting remains fixed throughout all the dialogue. A different strategy would be to consider that a real user may change his cooperativeness depending on the success of the interaction. For example, the cooperativeness of the simulator could be set to *low* at the beginning of the dialogue and it could be changed to *medium* or *high* dynamically as long as the system restricts the interaction flexibility as an attempt to recover from understanding problems.

We also plan to enable the simulator's ability to model the users' changes of mind. In our application domain these changes may be related to modifications in the ordered products, which will be useful to test the system functionality that handles the product orders.

6. Conclusions

This chapter has provided a survey on how to test spoken dialogue systems by means of *user simulators*. These simulators are systems designed to interact automatically with spoken dialogue system, replacing more or less faithfully the behaviour of users interacting with dialogue systems. This type of simulator has attracted the interest of the research community in the last decade, as they allow evaluating dialogue systems in a very simple way, employing a high number of dialogues that can be automatically generated. Therefore, the simulators reduce the time and effort required for the evaluation of the dialogue systems each time the systems are modified.

The chapter has surveyed the recent literature on user simulators for spoken dialogue systems focusing on the main implementation approaches (rule- and corpus-based as well as hybrid methods), addressing as well performance evaluation (direct and indirect evaluation methods). The survey also discussed a possible classification of user simulators in terms of the level of abstraction at which they model dialogue (acoustic, word or the intention levels).

The chapter has discussed as well our own contributions to this field, presenting firstly our initial simulator and then an enhanced version of it that enables assessing spoken dialogue systems by employing different levels of user cooperativeness. The improved simulator has been employed to test the

performance of a previously developed dialogue system (Saplen) using two front-ends for speech recognition: one based on 17 prompt-dependent language models (PDLMs) and the other based on one prompt-independent language model (PILM). We have described the speech database, the scenario corpus and the language models for speech recognition employed in the experiments. The results show that Saplen should only employ the PDLMs if the cooperativeness of the real users is *high* since the system performance can be considered more or less acceptable. Moreover, when the PILM is used the system performance is too poor for an interaction with real users regardless of the type of language model employed.

Focussing on the dialogues with very low values for the evaluation measures, we found the main problems in the performance of the dialogue system for the three cooperativeness levels considered, which let us think of possible ways to improve the system's performance in terms of speech recognition, spoken language understanding and dialogue management. The chapter finished presenting possibilities for future work in order to improve the performance of the enhanced user simulator, concretely to study alternative methods to simulate more precisely the behaviour of real users, and to provide the simulator with the ability to model changes of mind of the users.

Acknowledgments

This research has been funded by the Spanish project HADA TIN2007-64718.

References

- Ai, H. and Litman, D. (2008). Assessing Dialog System User Simulation Evaluation Measures Using Human Judges. In *Proceedings of the 46th Conference Association for Computational Linguistics (ACL'08)*, Ohio, USA.
- Ai, H., Tetreault, J., and Litman, D. (2007). Comparing User Simulation Models for Dialog Strategy Learning. In *Proceedings of Human Language Technologies 2007: the Conference of the North American Chapter of the Association For Computational Linguistics*, pages 1–4, Morristown, NJ, USA.
- Araki, M., Watanabe, T., and Doshita, S. (1997). Evaluating Dialogue Strategies for Recovering from Misunderstandings. In *Proceedings of IJCAI Workshop on Collaboration Cooperation and Conflict in Dialogue Systems*, pages 13–18.
- Bonafonte, A., Aibar, P., Castell, E., Lleida, E., Mariño, J., Sanchís, E., and Torres, M. I. (2000). Desarrollo de un sistema de diálogo oral en dominios restringidos. In *Proceedings of Primeras Jornadas en Tecnología del Habla*, Sevilla (Spain).

- Bos, J., Klein, E., Lemon, O., and Oka, T. (2003). DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124, Sapporo (Japan).
- Bunt, H. (1981). *Rules for the Interpretation, Evaluation and Generation of Dialogue Acts*. IPO anual progress report (16). Technische Universiteit Eindhoven.
- Carbini, S., Delphin-Poulat, L., Perron, L., and Viallet, J. (2006). From a Wizard of Oz Experiments to a Real Time Speech and Gesture Multimodal Interface. *Proceedings of IEEE Signal Processing*, 86(12):3559–3577.
- Chung, G. (2004). Developing a Flexible Spoken Dialog System Using Simulation. In *Proceedings of the 42nd Annual Meeting of the ACL'04*, pages 63–70.
- Cuayáhuítl, H., Renals, S., Lemon, O., and Shimodaira, H. (2005). Human-Computer Dialogue Simulation Using Hidden Markov Models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'05)*, pages 290–295, San Juan (Puerto Rico).
- Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J., and Gandy, M. (1986). Wizard of Oz Support throughout an Iterative Design Process. *Proceedings of IEEE Pervasive Computing*, 4(8):18–26.
- Eckert, W., Levin, E., and Pieraccini, R. (1997). User Modeling for Spoken Dialogue System Evaluation. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'97)*, pages 80–87, Santa Barbara (USA).
- Filisko, E. and Seneff, S. (2006). Learning Decision Models in Spoken Dialogue Systems Via User Simulation. In *Proceedings of AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*.
- Georgila, K., Henderson, J., and Lemon, O. (2005). Learning User Simulations for Information State Update Dialogue Systems. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech'05)*, pages 893–896, Lisbon (Portugal).
- Georgila, K., Henderson, J., and Lemon, O. (2006). User Simulation for Spoken Dialogue Systems: Learning and Evaluation. In *Proceedings of the 9th International Conference on Spoken Language Processing (Inter-speech/ICSLP)*, pages 1065–1068, Pittsburgh (USA).
- Griol, D., Callejas, Z., and López-Cózar, R. (2009). A Comparison between Dialog Corpora Acquired with Real and Simulated Users. In *Proceedings of the SIGDIAL 2009 Conference*, pages 326–332, London, UK. Association for Computational Linguistics.
- Griol, D., Hurtado, L. F., Segarra, E., and Sanchis, E. (2008). A Statistical Approach to Spoken Dialog Systems Design and Evaluation. *Speech Communication*, 50(8-9):666–682.

- Griol, D., Torres, F., Hurtado, L., Grau, S., García, F., Sanchis, E., and Segarra, E. (2006). A Dialog System for the DIHANA Project. In *Proceedings of International Conference Speech and Computer (SPECOM'06)*, pages 131–136, Saint Petersburg (Russia).
- Jung, S., Lee, C., Kim, K., Jeong, M., and Lee, G. G. (2009). Data-Driven User Simulation for Automated Evaluation of Spoken Dialogue Systems. *Computer Speech and Language*, 23:479–509.
- Komatani, K., Ueno, S., Kawahara, T., and Okuno, H. (2003). Flexible Guidance Generation using User Model in Spoken Dialogue Systems. In *Proceedings of the 41st Annual Meeting of the ACL'03*.
- Lemon, O. and Liu, X. (2007). Dialogue Policy Learning for Combinations of Noise and User Simulation: Transfer Results. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 55–58, Antwerp (Belgium).
- Levin, E. and Pieraccini, R. (1997). A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 1883–1896, Rhodes (Greece).
- Levin, E., Pieraccini, R., and Eckert, W. (2000). A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. In *IEEE Transactions on Speech and Audio Processing*, volume 8(1), pages 11–23.
- Lin, B. and Lee, L. (2001). Computer Aided Analysis and Design for Spoken Dialogue Systems based on Quantitative Simulations. *IEEE Trans Speech Audio Process*, 9(5):534–548.
- López-Cózar, R. and Callejas, Z. (2005). Combining Language Models in the Input Interface of a Spoken Dialogue System. *Computer Speech and Language*, 20:420–440.
- López-Cózar, R., de la Torre, A., Segura, J., and Rubio, A. (2003). Assessment of Dialogue Systems by Means of a New Simulation Technique. *Speech Communication*, 40(3):387–407.
- López-Cózar, R., García, P., Díaz, J., and Rubio, A. J. (1997). A Voice Activated Dialogue System for Fast-Food Restaurant Applications. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 1783–1786, Rhodes (Greece).
- López-Cózar, R., la Torre, A. D., Segura, J., Rubio, A., and López-Soler, J. (2002). A New Method for Testing Dialogue Systems based on Simulations of Real-World Conditions. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*, pages 305–308.
- Möller, S. (2004). *Quality of Telephone-based Spoken Dialogue Systems*. Springer.
- Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., and Reithinger, N. (2006). MeMo: towards Automatic Us-

- ability Evaluation of Spoken Dialogue Services by User Error Simulations. In *Proceedings of Interspeech '06*, pages 1786–1789, Pittsburgh (USA).
- Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann.
- Pietquin, O. and Beaufort, R. (2005). Comparing ASR Modeling Methods for Spoken Dialogue Simulation and Optimal Strategy Learning. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05)*, pages 861–864, Lisbon (Portugal).
- Rabiner, L., Juang, B., and Lee, C. (1996). An Overview of Automatic Speech Recognition. In Publishers, K. A., editor, *Automatic Speech and speaker Recognition: Advanced Topic*, pages 1–30.
- Schatzmann, J., Georgila, K., and Young, S. (2005). Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 45–54, Lisbon (Portugal).
- Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007a). Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 149–152, Rochester, NY (USA).
- Schatzmann, J., Thomson, B., and Young, S. (2007b). Statistical User Simulation with a Hidden Agenda. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 273–282, Antwerp (Belgium).
- Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. In *Knowledge Engineering Review*, volume 21(2), pages 97–126.
- Scheffler, K. and Young, S. (1999). Simulation of Human-Machine Dialogues. Technical report, CUED/F-INFENG/TR 355, Cambridge University Engineering Dept., Cambridge (UK).
- Scheffler, K. and Young, S. (2001). Corpus-based Dialogue Simulation for Automatic Strategy Learning and Evaluation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001). Workshop on Adaptation in Dialogue Systems*, pages 64–70, Pittsburgh (USA).
- Scheffler, K. and Young, S. (2002). Automatic Learning of Dialogue Strategy Using Dialogue Simulation and Reinforcement Learning. In *Proceedings of International Conference on Human Language Technology (HLT'02)*, pages 12–18, San Diego (USA).
- Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge (UK).
- Torres, F., Sanchis, E., and Segarra, E. (2008). User Simulation in a Stochastic Dialog System. *Computer Speech and Language*, 22(3):230–255.

- Traum, D. (1999). *Speech Acts for Dialogue Agents*. M. Wooldridge and A. Rao (eds.), Foundations of Rational Agency, Dordrecht, Kluwer.
- Wang, C., Seneff, S., and Chung, G. (2005). Language Model Data Filtering via User Simulation and Dialogue Resynthesis. In *Proceedings of Interspeech'05*, pages 21–24, Lisbon (Portugal).
- Williams, J. (2007). A Method for Evaluating and Comparing User Simulations: The Cramer-von Mises Divergence. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.
- Young, S. (2002). The Statistical Approach to the Design of Spoken Dialogue Systems. Technical report, CUED/F-INFENG/TR.433, Cambridge University Engineering Department, Cambridge (UK).

Chapter 4

SALIENT FEATURES FOR ANGER RECOGNITION IN GERMAN AND ENGLISH IVR PORTALS

Tim Polzehl

*Technischen Universität Berlin, Deutsche Telekom Laboratories
Berlin, Germany*

tim.polzehl@telekom.de

Alexander Schmitt

*Institute of Information Technology, University of Ulm
Ulm, Germany*

alexander.schmitt@uni-ulm.de

Florian Metze

*Language Technologies Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA*

fmetze@cs.cmu.edu

Abstract

Anger recognition in speech dialogue systems can help to enhance human computer interaction. In this chapter we report on the setup and performance optimization techniques for successful anger classification using acoustic cues. We evaluate the performance of a broad variety of features on both a German and an American English voice portal database which contain “real” (i.e. non-acted) continuous speech of narrow-band quality. Starting with a large-scale feature extraction, we determine optimal sets of feature combinations for each language, by applying an Information-Gain based ranking scheme. Analyzing the ranking we notice that a large proportion of the most promising features for both databases are derived from MFCC and loudness. In contrast to this similarity also pitch features proved importance for the English database. We further calculate classification scores for our setups using discriminative training and Support-Vector Machine classification. The developed systems show that anger

recognition in both English and German language can be processed very similarly reaching comparable results.

Keywords: Emotion recognition; IVR dialogue anger recognition; Acoustic and prosodic feature selection and classification; Affective speech.

1. Introduction

Detecting emotions in human computer interaction is gaining more and more attention in the speech research community. Moreover, classifying human emotions by means of automated speech analysis is achieving a performance, which makes deployment possible. Emotion detection in Interactive Voice Response (IVR) dialogue systems can be used to monitor quality of service or to adapt emphatic dialogue strategies (Yacoub et al., 2003; Shafran et al., 2003).

Especially anger recognition can deliver useful information to both the customer and the carrier of IVR platforms. It indicates potentially problematic turns or slots to the carrier so he can monitor and refine the system. It can further serve as trigger to switch between tailored dialogue strategies for emotional conditions to better react to the user's behavior (Metze et al., 2008; Burkhardt et al., 2005b). Some carriers have also been experimenting with re-routing the customers to the assistance of a human operator when problems occur. Problems and uncertainties arise from the imbalance in complexity between human computer interaction and models trained for these interactions. The difficulty is to capture the various and diverse patterns of human expression that convey emotional information by automated measurements.

In this chapter we analyze the importance of different acoustic and prosodic measurements, i.e. we examine expressive patterns that are based on vocal intonation. Applying our anger recognition system (Polzehl et al., 2009) we capture these expressions extracting low-level audio descriptors, e.g. pitch, loudness, MFCC, spectrals, formants and intensity. In a next step statistics are applied to the descriptors. These statistics mostly encompass moments, extrema, linear regression coefficients and ranges of the respective acoustic contours. We gain insight into the importance of our features by ranking them according to their Information-Gain Ratio. Looking at high-ranked features we report on their distribution and numbers in total as well as in relation to each other. We compare our features on an English and a German corpus both containing telephony conversations with IVR systems. Finally, we obtain classification scores from the assorted sets by means of discriminative classification by training Support Vector Machines. Applying cross-validation to our training data we estimate optimal parameter settings and give results on a separated hold-out set.

2. Related Work

Other systems also model the course of acoustic contours by dynamic methods (Vlasenko and Wendemuth, 2007). However, static modeling, as described in Section 5, outperforms dynamic modeling in almost all recent and current works on anger recognition. Also lexical and contextual information has been applied for the present task. Lexical features model the information given by the spoken utterances or word hypotheses obtained from automatic speech recognition (ASR). Lee (Lee and Narayanan, 2005; Lee et al., 2008) calculates the class-dependent discriminative power of a word using the self-mutual information criterion. Basing on this criterion he introduces the “emotional salience” of a word with respect to a class. Following his theory, a salient word is the one which appears more often in one class than in other classes. In order to give a word-dependent, class-independent score of emotional salience he applies a weighted summation over all classes. The higher the emotional salience of a word the more discriminative it is. Expanding the basic unit from separated words to phrases Metze et al. (2009) include contextual word information by calculating emotional salience of n-grams. Also the use of n-grams directly had been proposed (Steidl, 2009; Shafran and Mohri, 2005). Other linguistic features, e.g. the part-of-speech (POS) or bag-of-words (BOW) representations are reported on by Schuller et al. (2009). Finally, some systems also include models of contextual dialogue information (Lee and Narayanan, 2005; Liscombe et al., 2005; Schmitt et al., 2009). These features also comprise, the barge-in heuristic of a user, repetitions, ASR errors or the dialogue history.

Note that for any linguistic consideration the transcripts of the spoken words are needed. Systems have to be equipped with an ASR component or transcripts have to be added manually. The aim of the present contribution is to describe a system that works with acoustic cues only, independent from the ASR component.

3. Overview of Database Conditions

When comparing existing works on anger recognition one has to be aware of essential conditions in the underlying database design. The most restricted database settings would certainly have prearranged sentences performed by professional speakers (one at a time) recorded in audio studios tolerating almost no background noise and performing close capturing of speech signals. Real life speech does not have any of these settings.

Offering as much as 97% accuracy for recognition of angry utterances in a 7 class recognition test performed by humans the TU Berlin EMO-DB (Burkhardt et al., 2005a) bases on speech produced by German speaking professional actors. Here it is important to mention that the database contains

10 pre-selected sentences all of which are conditioned to be interpretable in 6 different emotions and neutral speech. All recordings have wideband quality. Classifying for all emotions and neutral speech automatically (Schuller, 2006) resulted in 92% accuracy. For this experiment he chose only a subset of the EMO-DB speech data that, judged by humans, exceeded a recognition rate of 80% and a naturalness evaluation value of 60%. Eventually, 12% of all utterances selected contained angry speech. He implemented a high number of acoustic audio descriptors such as intensity, pitch, formants, Mel-frequency Cepstral Coefficients (MFCCs), harmonics to noise ratio (HNR), and further information on duration and spectral slope. He compared different classification algorithms and obtained best scores with Support Vector Machines (SVM).

A further anger recognition experiment was carried out on the DES database (Enberg and Hansen, 1996) which contains mostly read Dutch speech and also includes free text passages. All recordings are of wide band quality as well. The main difference to the EMO-DB is that the linguistic content had not been controlled entirely during recordings. The people chose their words according to individual topics. The accuracy for human anger recognition for this corpus resulted in 75%. This accuracy bases on a five class recognition test. The approach by Schuller results in 81% accuracy when classifying for all emotions. Voting for the class that has the highest prior probability would reach an accuracy of 31% only.

It is essentially important to note that also these results are based on acted speech data, containing consciously produced emotions, performed by professional speakers. Human recognition rates were obtained by comparing impressions of the labelers during the perception test with the intended emotions of actors' performances. In cases where there is no professional performance, i.e. when looking at natural speech utterances, we need to rely on the labels of the testers only. To obtain a measurement for consistency of such corpora the inter labeler agreement measurement can be applied. It is the ratio of the chance level corrected proportion of times that the labelers agree to the maximum proportion of times that the labelers could agree. The inter labeler agreement of two labelers is given by Cohen's Kappa. We apply Davies extension of Cohen's Kappa (Davies and Fleiss, 1982) for multiple labelers to give a value of coherence among the labelers.

Lee and Narayanan (2005) as well as Batliner (2000) used realistic IVR speech data. These experiments use call center data of narrow-band quality. Also the classification tasks were facilitated. Both applied binary classification, i.e. Batliner discriminates angry from neutral speech, Lee and Narayanan classify for negative versus non-negative utterances. Given a two class task it is even more important to know the prior probability of class distribution. Batliner obtains an overall accuracy of 69% using Linear Discriminative Clas-

sification (LDC). Unfortunately no class distribution or inter labeler agreement for his corpus is given. Lee and Narayanan reached a gender dependent accuracy of 81% for female and 82% for male speakers. He measured inter labeler agreement with 0.45 for male and 0.47 for female speakers, which can be interpreted as moderate agreement. For both gender classes, constant voting for the non-negative class would mean to achieve roughly 75% accuracy already and - without any classification - outperforms the results obtained by Batliner.

Note that, given any class distribution skewness, the accuracy measurement allows for false bias since it is influenced by the majority class to a greater extent than it is influenced by other classes. If a model of the majority class yields better scores than other models for other non-majority classes and the class distribution is not balanced the resulting accuracy measurement gives overestimated figures. In the present case of anger recognition such an inequality in model performance is often the case. We therefore emphasize the general use of balanced performance measurements, such as the f1-measure, which will be discussed in Section 8.

4. Selected Corpora

Nearly all studies on anger recognition are based on a singular corpus making a generalization of the results difficult. Our aim in this study is to compare the performance of different features when trained and tested on different languages. Both of the databases we used do have background noise, recordings do include cross- and off-talk, speakers are free in choice of words and would never enunciate words as clearly as trained speakers do.

The German database roughly captures 21 hours recordings from a German Interactive Voice Response (IVR) portal offering telephone-related services and assistance in troubleshooting. The data can be subdivided into 4683 dialogs, averaging 5.8 turns per dialog. For each turn, 3 labelers assigned one of the following labels: *not angry*, *not sure*, *slightly angry*, *clear anger*, *clear rage* or marked the turns as *non applicable* when encountering garbage. The labels were mapped onto two cover classes by clustering according to a threshold over the average of all voters' labels as described in (Burkhardt et al., 2009). Following Davies extension of Cohen's Kappa (Davies and Fleiss, 1982) for multiple labelers we obtain a value of $\kappa = 0.52$ which corresponds to moderate inter labeler agreement (Steidl et al., 2005). Finally, our training setup contains 1761 angry turns and 2502 non-angry turns. The test setup includes 190 angry turns and 302 non-angry turns which roughly corresponds to a 40/60 split of anger/non-anger distribution in the sets. The average turn length after cleaning out initial and final pauses results in 1.8 seconds.

The English database originates from a US-American IVR portal capable of fixing Internet-related problems jointly with the caller. Three labelers divided

the corpus into *angry*, *annoyed* and *non-angry* utterances. The final label was defined based on majority voting resulting in 90.2% neutral, 5.1% garbage, 3.4% annoyed and 0.7% angry utterances. 0.6% of the samples in the corpus were sorted out since all three raters had different opinions. While the number of angry and annoyed utterances seems very low, 429 calls (i.e. 22.4% of all dialogues) contained annoyed or angry utterances. In order to be able to compare results of both corpora we matched the conditions of the English database to the conditions of the German database, i.e. we collapsed *annoyed* and *angry* to *angry* and created a test and training set according to the 40/60 split. The resulting training set consists of 1396 non-angry and 931 angry turns while the final test set comprises 164 non-angry utterances and 81 utterances of the anger class. The inter labeler agreement in the final set results in $\kappa = 0.63$, which also resembles moderate agreement. The average turn length after cleaning out initial and final pauses is approx. 0.8 seconds. Details of both corpora are listed in [Table 1](#).

Table 1. Database comparison of both corpora.

	German	English
Domain	Mobile	Internet Support
Number of Dialogs in Total	4682	1911
Duration in Total	21h	10h
Average Number of Turns per Dialog	5.7	11.88
Number of Raters	3	3
Speech Quality	Narrow-band	Narrow-band
Deployed Subsets for Anger Recognition		
Number of Anger Turns in Train set	1761	931
Number of Non-Anger Turns in Train set	2502	1396
Number of Anger Turns in Test set	190	81
Number of Non-Anger Turns in Test set	302	164
Average Utterance Length in Seconds ^a	1.80	0.84
Average Duration Anger in Seconds	3.27 ± 2.27	1.87 ± 0.61
Average Duration Non-Anger in Seconds	2.91 ± 2.16	1.57 ± 0.66
Cohen's Extended Kappa	0.52	0.63

^a without initial or final turn pauses

5. Prosodic and Acoustic Modeling

Our prosodic and acoustic feature definition provides a broad variety of information about vocal expression patterns that can be useful when classifying speech metadata. Our approach is structured into two consecutive steps. In the first step an audio descriptor extraction unit processes the raw audio format

and provides speech descriptors. In the second step a statistics unit calculates various statistics on both the descriptors and certain sub-segments of them.

5.1 Audio Descriptor Extraction

All descriptors are extracted using 10ms frame shift. For any windowing we used Gaussian windows. The resulting audio descriptors can be sub-divided into 7 groups: *pitch*, *loudness*, *MFCC*, *spectrals*, *formants*, *intensity* and *other*.

5.1.1 Pitch. Starting with the group of perceptually motivated measurements we extract *pitch* by autocorrelation as described in (Boersma and Weenink, 2009). To avoid octave jumps in pitch estimation we post-process a range of possible pitch values using relative thresholds between voiced and unvoiced candidates. Remaining octave confusions between sub-segments of a turn are further processed by a rule-based path finding algorithm. In order to normalize for the absolute height of different speakers we convert pitch into the semitone domain using the mean pitch as reference value for a whole turn. As pitch is not defined for unvoiced segments we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares.

5.1.2 Loudness. Another perceptively motivated measurement is the *loudness* as defined by (Fastl and Zwicker, 2005). This measurement operates on a Bark filtered version of the spectrum and finally integrates the filter coefficients into a single loudness value in some units per frame. In contrast to pitch this measurement is always defined so we obtain a continuous descriptor contour.

5.1.3 MFCC. We further filter the spectrum into Mel domain units. After filtering a discrete cosine transformation (DCT) gives the values of the Mel frequency cepstral coefficients (*MFCC*). We place the filter centers in equally spaced intervals of 100 Mel distance. We compute a number of 16 MFC coefficients and keep the zero coefficient. Although MFCCs are most commonly used in speech recognition tasks they often give excellent performance in anger recognition tasks as well (Nobuo and Yasunari, 2007).

5.1.4 Spectrals. Other features drawn from the cepstral representation of the speech signal are the center of spectral mass gravity (spectral centroid) and the 95% roll-off point of spectral energy. Both features capture aspects related to the spectral slope (also called the spectral tilt) and correspond to perceptual impression of sharpness and brightness of sounds (Fastl and Zwicker, 2005). Another measurement drawn from spectral representation is the magnitude of spectral change over time, also known as spectral flux. The more abruptly changes in the spectrum occur the higher the magnitude of

this measurement. Grouping the center of spectral mass, the roll-off point and the spectral flux together these features will be referred to as *spectrals* in the following experiments.

5.1.5 Formants. Further, we extract 5 formants center frequencies and estimate the respective bandwidths. Looking for formants up to approx 3.5kHz we apply a pre-emphasis of 6dB/octave before computing LPC coefficients after Burg as given by (Press et al., 1992). We extract formants only for voiced regions as the algorithm yields reasonable results for voiced segments only.

5.1.6 Intensity. Taken directly from the speech signal we extract the contour of *intensity*. Taking the square of discrete amplitude values we convert every windowed frame's energy into dB scale relative to the auditory threshold pressure. To avoid any DC offset we subtract the mean pressure before calculation.

5.1.7 Others. Referred to as *other* features we calculate the Harmonics-to-Noise Ratio (HNR). Similar to pitch processing this measurement is taken from the autocorrelation domain. The HNR estimates the amount of harmonicity in the signal by means of periodicity detection. Also this measurement is calculated for voiced segments only. Next, we add a single coefficient for the correlation between pitch and intensity as an individual feature. Examining the signal amplitude we calculate the Zero-Crossing-Rate and estimate the average amplitude over the turn. Finally, taken from the relation of pitched and non-pitched speech segments we calculate durational or rhythm-related features such as pause lengths and the average expansion of voiced segments.

5.2 Statistic Feature Definition

The statistic unit derives means, moments of first to fourth order, extrema and ranges from the respective contours in the first place. Resulting features are e.g.: the standard deviation of the pitch, the average loudness level, the mean of a Mel frequency cepstral coefficient, the maximum change in spectral flux, the range of variations in bandwidths of a certain formant, the distribution skewness of intensity level or the minimal level of harmonicity.

Special statistics are then applied to certain descriptors such as pitch and loudness. These descriptors are examined with a linear regression analysis. We also include the error coefficient from the analysis in order to have an estimation of linearity of the contours.

Furthermore, pitch, loudness and intensity are additionally processed by a Discrete Cosine Transformation (DCT). Applying DCT to these contours di-

rectly we model their spectral composition. There exist different norms of DCT calculation. We refer to a DCT type III which is defined as:

$$X_k = \frac{1}{2}x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N}n\left(k + \frac{1}{2}\right)\right] \quad k = 0, \dots, N-1. \quad (4.1)$$

A high correlation of a contour with the lower coefficients indicates a rather slowly moving time behavior while mid-range coefficients would rather correlate with fast moving audio descriptors. Higher order coefficients would correlate with micro-prosodic movements of the respective curves, which corresponds to a kind of shimmer in the power magnitude or jitter in pitch movement.

A crucial task is the time normalization. Dealing with IVR speech we usually deal with very short utterances that often have command-like style. We suppose, every turn is a short utterance of one prosodic entity. Consequently we calculate our statistics to account for whole utterances. Note that this seems suboptimal for longer utterances. We keep this approach for the current status of experiments due to our corpus design.

In order to exploit the temporal behavior at a certain point in time we append Delta coefficients of first (Δ) and second ($\Delta\Delta$) order and calculate statistics on them alike.

As already mentioned, some features tend to give meaningful values only when applied to specific segments. We therefore developed an extended version of the speech-silence detection proposed by (Rabiner and Sambur, 1975). After having found voiced points we move to the very first and the very last point now looking for adjacent areas of relatively high zero-crossing rates. Also any non-voiced segment in between the outer borders is classified into high and low zero-crossing regions corresponding to unvoiced or silent speech segments. Eventually, we calculate features on the basis of voiced and/or unvoiced sounds both separately and jointly. In order to capture magnitudes of voiced to unvoiced relations we also compute these quotients as ratio measurements. We apply it to audio descriptors such as intensity and loudness to obtain:

- ratio of mean of unvoiced to mean of voiced points;
- ratio of median of unvoiced to median of voiced points;
- ratio of maximum of unvoiced to maximum of voiced points.

In some utterances we notice an absence of unvoiced sounds. In fact the English database includes less unvoiced sounds than the German does. This can be due many reasons. First, standard English usually entails a lower level

of pressure when producing unvoiced sounds, e.g. fricatives and especially the glottal "h" sound. Also the phonological strong aspiration is normally expected to occur with less pressure in English (Wells, 1982). Thus in English language these sounds may be harder to detect from ZCR and our detection algorithm may fail. Secondly, this can refer to a difference in speaking style. The average utterance length of English samples shows nearly half the length of German utterances. This could indicate a more command-like speaking style, i.e. omitting words that are not necessary, consequently being less outspoken. After all, 16% of all utterances in the German training set and 22% of all utterances in the German test set were of no unvoiced sound share. For the English database these figures raised to 27% and 33% respectively. The longer the turns the more reasonable this measurement can be applied.

In total, we obtain some 1450 features. Table 2 shows the different audio descriptors and the number of features calculated from them. Table 2 also shows figures of f1 performance, which will be discussed in the Section 6 and Section 8. Note that the different number of features can take bias on the performance comparison. Further insight can be gained when examining individual feature performance, e.g. produced by a feature ranking scheme as proposed in Section 6.

Table 2. Feature groups and performance on the German and English database.

Feature Group	Number of Features	f1 Performance on German DB	f1 Performance on English DB
pitch	240	67.7	72.9
loudness	171	68.3	71.2
MFCC	612	68.6	68.4
spectrals	75	68.4	69.1
formants	180	68.4	67.8
intensity	171	68.5	73.5
other	10	56.2	67.2

6. Feature Ranking

In order to gain insight about which of our features are most suitable for the given classification task we apply a filter-based ranking scheme, i.e. the Information-Gain-Ratio (IGR) (Duda et al., 2000). This measure evaluates the gain in information that a single feature contributes in adding up to an average amount of information needed to classify for all classes. It is based on the Shannon Entropy H (Shannon, 1948) for a class distribution $P(p_1, \dots, p_K)$ of P samples which is measured in bit unit and defined as

$$H = - \sum_{i=1}^K p_i \cdot \log_2(p_i). \quad (4.2)$$

Now let Ψ be the totality of our samples and $\Psi_i \in \Psi$ the subset of elements that belongs to class index i . The average information needed in order to classify a sample out of Ψ into a class $i_1 \dots i_K$ is given by

$$H(\Psi) = - \sum_{i=1}^K p_i \cdot \log_2(p_i) \quad \text{with} \quad p_i = \frac{|\Psi_i|}{|\Psi|}. \quad (4.3)$$

To estimate the contribution of a single feature every unique value is taken as partition point. For non-discrete features a discretization has to be executed. Let $\Psi_{x,j}$ with $j = 1 \dots J$ bins be the partition blocks of Ψ_x , holding values of a single feature x , the amount of information contributed by this feature is given by

$$H(\Psi|x) = \sum_{j=1}^J \frac{|\Psi_{x,j}|}{|\Psi|} \cdot H(\Psi_{x,j}). \quad (4.4)$$

The Information Gain (IG) of a feature is then given as its contribution to reach the average needed information for classification:

$$IG(\Psi, x) = H(\Psi) - H(\Psi|x). \quad (4.5)$$

The IGR accounts for the fact that information gain is biased towards features with high number of individual values in their span. IGR normalizes the information gain by the amount of total information that can be drawn out of J splits:

$$IGR(x, \Psi) = \frac{IG(\Psi, x)}{H(\frac{|\Psi_{x,1}|}{|\Psi|}, \dots, \frac{|\Psi_{x,J}|}{|\Psi|})}. \quad (4.6)$$

Table 3 presents the 20 top-ranked features for the English and the German corpus according to IGR. To obtain a more general and independent ranking we performed 10-fold cross validation as described in Section 8.1. The ranking presented accounts for the average ranking throughout the folds.

For the English database almost all features are of loudness descriptor origin predominantly capturing the moments of the contour or its maximum and range applied to the original contour, not its derivatives. The picture is much more diverse when we look at the German ranks. Although the loudness features that are present are of the same kind as those on the English set we note also formant, MFCC and intensity descriptors.

Table 3. Top-20 ranked features for German and English databases.

<i>German Database</i>	<i>English Database</i>
intensity DCT coeff ₂	loudness max
loudness std	loudness std of voiced points
loudness max	loudness std
5th formant bandwidth std	loudness mean
5th formant std	loudness inter-quartile range
intensity err. of lin.reg over voiced points of $\Delta\Delta$	loudness mean voiced points
loudness std of voiced points	intensity skewness of voiced points
loudness DCT coeff ₁ of $\Delta\Delta$	loudness inter-quartile range of voiced points of Δ
intensity err. lin.reg over voiced points of Δ	loudness median
loudness inter-quartile range	loudness median over voiced points
loudness DCT coeff ₂ of Δ	loudness DCT coeff ₁ 6
MFCC coeff ₁ 5 std over whole utterance	loudness std voiced points of Δ
loudness mean voiced points	loudness DCT coeff ₂ 6
pitch lin.reg over Δ	loudness DCT coeff ₁ 2
MFCC coeff ₁ min of voiced segments	loudness max unvoiced points
pitch mean of $\Delta\Delta$	intensity lin.reg. of voiced points of $\Delta\Delta$
pitch mean	loudness DCT coeff ₂ 0
loudness DCT coeff ₁ 1	loudness DCT coeff ₃ 0
loudness inter-quartile range of voiced points	loudness max of Δ
MFCC max of coeff ₁ 0 of voiced segments	loudness DCT coeff ₁ of Δ

7. Normalization

In order to compare observations from different (normal) distributions we apply *z-normalization*, which results are also known as *z-scores*, *normal scores* or *standardized variables* (Duda et al., 2000). New values are derived by subtracting a population mean from its own score and then dividing it by the population standard deviation. Each feature results in a normalized value having a mean of zero and unit standard deviation.

Let $\mu(x)$ be the mean of a feature population and $\sigma(x)$ be its standard deviation. A standardized feature value $\tilde{x}_i$ is then given by

$$\tilde{x}_i = \frac{x_i - \mu(x)}{\sigma(x)}. \quad (4.7)$$

The resulting distance is also called *Mahalanobis distance* and measures the distance from x to $\mu(x)$ in units of standard deviations.

Other normalizing steps are taken in the feature extraction unit, e.g. pitch is converted in semitones relative to the turn mean value. Thus we normalize for different heights of different speakers' voices. Also intensity is normalized by a fixed relation to the auditory threshold in dB. For more information see Section 5.

8. Classification

When classifying patterns into classes there are, besides issues of normalization, three major choices, i.e. the evaluation strategy, the desired measure for evaluation and the applied classification algorithm. Normally we strive to obtain results that are not only valid for current experimental setups but also for any unknown data. In other words, we want to maximize the external validity. Doing so, the most common technique is the cross validation. Also the measurement of classification success is crucial. Unbalanced class sizes and random chance probabilities often blur the actual accomplished classification success with statistical bias. Finally, the choice of the classification algorithm has to be in line with computational resources, real-time requirements and the size of training examples.

8.1 Cross Validation

To avoid over-fitting of our classification algorithm we apply a 10-fold cross validation on the training set. I.e. we partition all our training data in 10 fixed, equally sized splits, each mirroring the class distribution of the whole training set. Now we run the evaluation on 10 folds. Each fold treats 9 out of 10 splits as training material and uses one split for testing. For each of the 10 passes a different split is chosen as test split. Since a test split has never been included in the training material the 10 estimations provide independent estimations. After all folds have been processed the resulting estimations are averaged. Our partitions are designed in a speaker independent way, i.e. a speakers in the test split of a fold never occurred in the training material of that fold. This procedure gives advantage over the traditional definition of one global training set and one global test set, where all the material in the training set serves for building one global model and the evaluation is done processing the unseen global test set once. By applying cross validation we obtain a better impression

about how successful the classifier operates when encountering different test splits. The traditional way provides only one such result. However, in order to be comparable to our former systems we additionally keep a fixed holdout set (global test set) for evaluation.

8.2 Evaluation Measurement

In order to compare results from different feature sets we calculate classification success using the *f1*-measurement. In information retrieval in general the concepts of precision and recall are essential estimates to analyze classification results. The recall of a class measures how many examples - out of all examples of a class at hand - were effectively classified into the right class. The precision on the other hand considers the classified examples and counts how many of these examples - that were already classified into the class at hand - actually belong to that class. Note that we can always reach a recall of 100% by simply collecting all examples into one class. However, this class would have the worst precision. On the other hand we could act overcautiously and only assign an example to a class when we are absolutely sure of doing the right allocation. In this case we would result in a high precision in that class but we would probably reject a lot of examples that originally belong to the class. What we want to achieve is a classification of both high recall and high precision. The *F-measure* is one measurement capable of dealing with the problem (Witten and Frank, 2005). It accounts for the harmonic mean of both precision and recall of a given class.

Let TP be the number of true positive and TN be the number of true negative examples that are correctly classified into the respective positive (Anger) or negative (Non-Anger) class. A false positive FP occurs when the outcome is incorrectly predicted as positive when it is actually negative. A false negative FN occurs when the outcome is incorrectly predicted as negative when it is actually positive. Then the F-measure of a class is given by:

$$F = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \quad (4.8)$$

Since we normally want to classify examples into more than only one class we need to consider the sizes of training examples in each class. The often used accuracy measurement would allow for false bias since it is influenced by the majority class to a greater extent than by other classes. If we have an unbalanced class distribution and the model of the majority class yields better scores than other models for other non-majority classes the resulting accuracy measurement gives overestimated figures. Since our class distribution is unbalanced and our models tend to fit the majority class to a greater extent we use the *f1*-measurement for final classification success estimation. The *f1*

is defined as the arithmetic (unweighted) mean of F-measures from all data classes, i.e.:

$$f1 = \frac{F_{anger} + F_{non-anger}}{2}. \quad (4.9)$$

The f1 accounts for a balanced estimation of overall success.

8.3 Classification Algorithm

To obtain f1-scores we used Support Vector Machines (SVM) (Vapnik and Cortes, 1995). One reason for this choice is that SVMs are proven to yield good results for small data sets. SVMs view data as sets of vectors in a multi-dimensional space. The task of the algorithm is to find the hyper-plane in that separates the binary classes and provides a maximal margin in between the vectors from different classes. Maximizing the corridor between the hyper-plane and the data points the classification provides a high degree of generalization. Furthermore, the algorithm defines the hyper-plane by means of support-vectors, which are to be selected out of all data vectors. Although the training run can be very costly in the test phase only those vectors that had been selected as support vectors are computationally relevant. SVMs can be extended to non-linear feature spaces by passing the original data points through kernel functions, which according to our experiments leads to little improvement in terms of classification scores but rises costs drastically at the same time. The choice of the best kernel function can only be done experimentally. We use a linear kernel function for the present experiments.

9. Experiments and Results

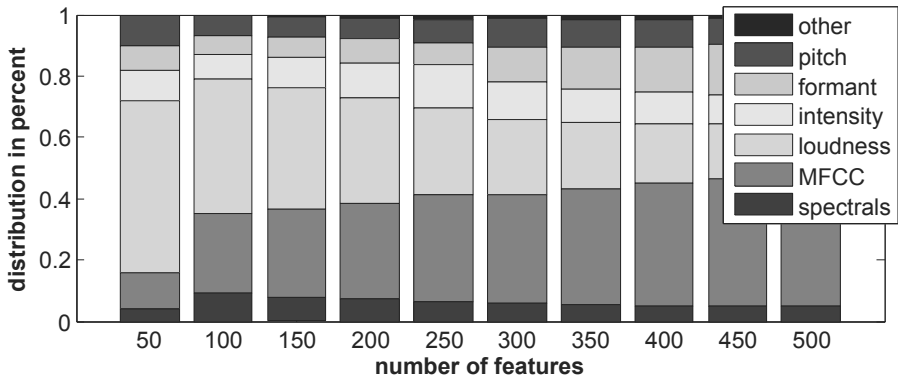
9.1 Analyzing Feature Distributions

Figure 1 shows the relative distributions of the feature sets grouped to their audio descriptor's origin when expanding the feature space from 50 top-ranked features to 500 top-ranked features. The number of features in total is 1450. Comparing ranks we notice that the top 50 ranks of the English database are occupied by intensity, spectrals and predominantly loudness features only. Pitch, formants and MFCC descriptors are not generating top rank features within the top 100 ranks. However, beyond this point pitch features become much more important for the English database than for the German.

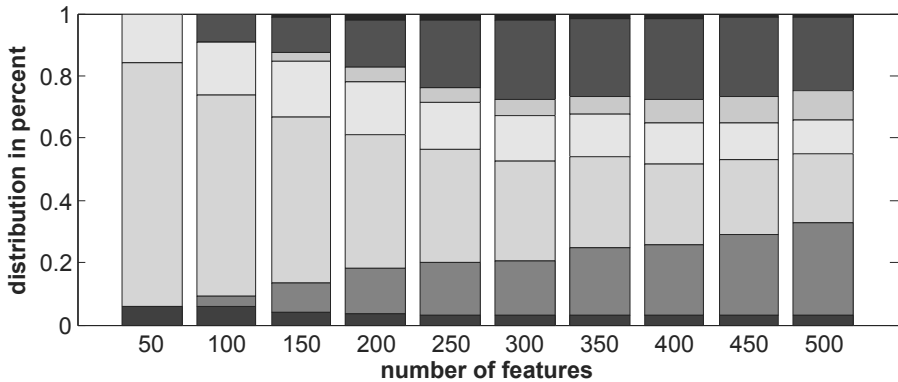
Table 2 already suggests that the different audio descriptor groups are of more equally weighted importance for the German database than they are for the English one. Also the feature distribution in the top-ranks suggest a more heterogeneous distribution in the German set. In general it seems as for the German set loudness and MFCCs are building the most important descriptors.

The more features the more important become MFCC contours. Note that also here the absolute number of MFCCs features affects the distribution more and more when the feature space expands.

As we expand the feature space for the English database three descriptor groups are of most importance: loudness, MFCC and pitch. Also cross-comparing the languages it seems that loudness is of higher impact for the English database as there are consistently more loudness features among all sizes of feature spaces for the English database. On the opposite, MFCC descriptors are more important to the German database. Note that these charts do not tell about how good the classification would be. This issue is discussed in the following section.



(a) German corpus



(b) English corpus

Figure 1. Feature group distribution group from top-50 until top-500 ranked features.

9.2 Optimal Feature Sets

Table 2 already gives the f1-measurements for classification of features derived from separated audio descriptor groups. To obtain better results we are now looking for the best combination of features from all descriptor groups. We thus need to find the parameters, i.e. the optimal number of features and the actual features included in that set. We make use of the IGR ranking again. Moving along the top ranks we incrementally admit a rising number of top-ranked features for classification. As expected, the scores predominantly rise as we start expanding the feature space. At a certain point no relevant information seem to be added when including more top-ranked features. In this phase, the scores seem to remain at a certain level showing some jitter. After including even more features we notice an overall decrease of performance again. Figure 2 shows the development of f1-measurement by incremental expansion.

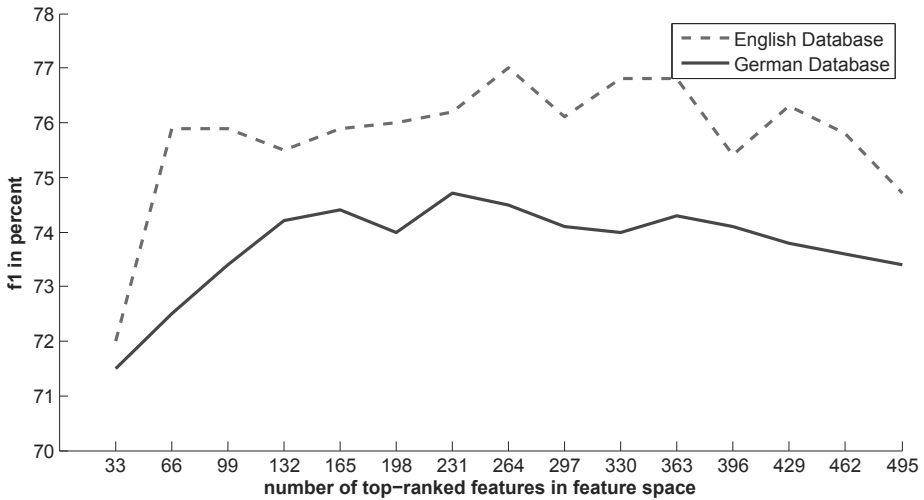


Figure 2. Determination of optimal feature set size.

The optimal number of top-ranked features to include into the feature space resulted in 231 for the German database and 264 for the English database. Looking at Figure 1 once more we can clearly see that on basis of the English database there is a higher number of pitch and loudness features in the top 250 feature space whereas in the German database more MFCC features can be found.

Note that the saw-like shape of the graphs in Figure 2 indicate a non-optimal ranking since some feature inclusions seem to harm the performance. This is because the IGR filter uses heuristics to estimate the gain of information a single feature offers. Also any feature combination effects are not considered

by the filter, although it is known that features that are less discriminative can develop high discrimination when integrated into a set. However, in the present experiments the IGR filter estimates the gain of single features independent from any set effects. Regarding the magnitude of the jitter we can see that it is as low as approx. 1% which after all proves a generally reasonable ranking. Regarding other requirements, as to lower to computational costs, one could also stop at the first local maximum of the f1 curves resulting in a reduced feature set of 66 features for the English and 165 features for the German database without losing more than 1% f1.

9.3 Optimal Classification

In a final step we adjusted the complexity of our classification algorithm which results in a best score of 78.2 f1 for the English and 74.7 f1 for the German database. All parameter settings were obtained by cross-validation evaluation. Previous studies on both corpora yielded a much lower performance compared to our new findings. The former system described in (Schmitt et al., 2009) with the English database reached 72.6% f1 while the system described in (Burkhardt et al., 2009) developed for the German database reached 70% f1. The performance gain on the training set of respectively 5.6% and 4.7% f1 in our study can be attributed to the employment of the enhanced feature sets and the feature selection by IGR filtering.

Applied to the holdout sets we obtain figures presented in Table 4. For both languages the models capture more of Non-Anger information than of Anger information. Consequently the F-measure of the Anger class is always lower than the one of the Non-Anger class. We also see a better recall of Anger in English language. At the same time we see a better precision in German classification. After all, the overall performance of the final systems proved to be equivalently high.

Note, that for given the class distribution of roughly 33/66 split in the test set constant classification into the majority class would result in approx. 40% f1.

Table 4. System performance figures on test sets.

Database	Class	Recall	Precision	F-measure	f1-measure
German	Non-Anger	88.9%	84.9%	86.7%	77.2%
	Anger	63.7%	72.0%	67.6%	
English	Non-Anger	82.3%	86.0%	84.1%	77.0%
	Anger	72.8%	67.0%	69.8%	

10. Discussion

Looking at classification scores we observe comparable results for both languages. Also the number of features needed for a reasonable anger classification seems similar. The difference in performance between test and train sets indicates that the calculated scores are reliable. Absolute scores prove a good overall classification success. Computational complexity can be reduced considerably without losing much classification score. However, we found differences in feature space setup in between the databases. The following section presents possible explanations and discussion questions.

10.0.1 Signal Quality. One hypothesis for explaining the differences in feature distribution could be that callers may have dialed in via different transmission channels using different encoding paradigms. While the English database mostly comprises calls that were routed through land line connections the German database accounts for a greater share of mobile telephony transmission channels. Since fixed line connections transmit usually less compressed speech it can be assumed that there is more information retained in it. An analysis of the impact of information detail level on a anger recognition task remains to future research, as more information transmitted in total does not automatically mean more relevant information among total information.

10.0.2 Speech Length. Another hypothesis for explaining the differences in the results could be the discrepancy in average turn length. The turn length can have a huge effect on statistics when applying a static feature length classification strategy. To estimate the impact of the average turn length we subsampled the German database to match the English average turn length. We processed the sub-samples analogously to the original database. As a result we obtain major differences in the ranking list when operating on the shorter subset. While MFCC features account for roughly 35% in the original ranked German feature set the number drops to 22% on the subset. Accordingly, this figure becomes closer to the figure of 18% when working on the English corpus. Consequently we can hypothesize that the longer the turn the more important the MFCC features become. A possible explanation could be the increasing independence of the MFCC from the spoken context when drawing features on turn length. Though 70% of the MFCC features on the original set are also among the top ranked features on the subset the differences seem to be concentrated on the features from voiced speech parts. Also the higher MFCC coefficients seem to be affected from replacement. Further experiments showing the impact of inter-lingual feature replacements can be found in (Polzehl et al., 2010).

On the other hand, loudness and pitch features tend to remain on the original ranks when manipulating the average turn length. After all, we still observe a large difference between the German and the English databases when looking at pitch features. Sub-sampling did not have any significant effect. Consequently this difference is not correlated with the average turn length of the database. Further, there is no clear effect from band limitation coming from different transmission channels. Pitch estimation by autocorrelation reconstructs the pitch into similar intervals for both databases. On the basis of these findings we can further hypothesize that there might exist a larger difference in emotional pitch usage in between German and English language at a linguistic level. In English language pitch variations might have a generally larger connection to vocal anger expression than in German.

10.0.3 Speech Transcription. Finally, the procedures of training the labelers and the more precise differences in IVR design and dialogue domain could be considered as possible factors of influence as well. Also, as the English database offers a higher value of inter labeler agreement we would expect a better classification score for it. After all, though the classification results on the training sets mirror this difference they seem very balanced when classifying on the test sets. However, a difference in performance between test and training sets which accounts for less than 4% seems to indicate reasonable and reliable results for our anger recognition system on both corpora.

11. Conclusions

On the basis of the selected corpora, we have shown that detecting angry utterances from IVR speech by acoustic measurements in English language is similar to detecting those utterances in German language. We have set up a large variety of acoustic and prosodic features. After applying IGR filter based ranking we compared the distribution of the features for both languages. Working with both languages we determine an absolute optimum when including 231 (German database) and 264 (English database) top-ranked features into the feature space. With respect of the maximum feature set size of 1450 these numbers are very close.

When choosing the optimum number of features for each language, the relative importance of feature groups is also similar. Features derived from filtering in the spectral domain, e.g. MFCC, loudness, seem to be most promising for both databases. They account for more than 50% of all features. However, MFCCs occur more frequently under the top-ranked features when operating on the German database, while operating on the English database loudness features are more frequently among top ranks.

Another difference lies within the impact of pitch features. Although they are not among the top 50 features they become more and more important

when including up to 300 features for the English language. They account for roughly 25% when trained on the English database while the number is as small as roughly 10% when trained on the German corpus.

In terms of classification scores we obtain equally high f1 scores of approx. 77% for both languages. The classification baseline, which is given by constant majority class voting, is approx. 40% f1. Our results clearly outperform the baseline and improve previous versions of our anger recognition system. Moreover, the overall level of anger recognition preformance in IVR speech opens up deployment possibilities for real life applications.

Acknowledgments

The authors wish to thank Prof. Sebastian Möller, Prof. Wolfgang Minker, Dr. Felix Burkhardt, Dr. Joachim Stegmann, Dr. David Sündermann, Dr. Roberto Pieraccini and Dr. Jackson Liscombe for their encouragement and support. The authors also wish to thank their colleagues at Quality and Usability Lab of Technische Universität Berlin / Deutsche Telekom Laboratories, the Dialogue Systems Group of the Institute of Information Technology at University of Ulm and the Language Technologies Institute of Carnegie Mellon University, Pittsburgh for support and insightful discussions.

The research described in this chapter was supported by the Transregional Collaborative Research Center SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

- Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In *Proceedings of ISCA Workshop on Speech and Emotion*.
- Boersma, P. and Weenink, D. (2009). Praat: Doing Phonetics by Computer. <http://www.praat.org/>.
- Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., and Huber, R. (2009). Detecting Real Life Anger. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Burkhardt, F., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005a). A Database of German Emotional Speech. In *Proceedings of Interspeech*. ISCA.
- Burkhardt, F., van Ballegooy, M., and Huber, R. (2005b). An Emotion-Aware Voice Portal. In *Proceedings of Electronic Speech Signal Processing ESSP*.
- Davies, M. and Fleiss, J. (1982). Measuring Agreement for Multinomial Data. volume 38.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. John Wiley & Sons, 2nd edition.

- Enberg, I. S. and Hansen, A. V. (1996). Documentation of the Danish Emotional Speech Database. Technical report, Aalborg University, Denmark.
- Fastl, H. and Zwicker, E. (2005). *Psychoacoustics: Facts and Models*. Springer, Berlin, 3rd edition.
- Lee, C. M. and Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing*, 13:293–303.
- Lee, F.-M., Li, L.-H., and Huang, R.-Y. (2008). Recognizing Low/High Anger in Speech for Call Centers. In *Proceedings of International Conference on Signal Processing, Robotics and Automation*, pages 171–176. World Scientific and Engineering Academy and Society (WSEAS).
- Liscombe, J., Riccardi, G., and Hakkani-Tür, D. (2005). Using Context to Improve Emotion Detection in Spoken Dialog Systems. In *Proceedings of Interspeech*, pages 1845–1848.
- Metze, F., Englert, R., Bub, U., Burkhardt, F., and Stegmann, J. (2008). Getting Closer: Tailored Human-Computer Speech Dialog. *Universal Access in the Information Society*.
- Metze, F., Polzehl, T., and Wagner, M. (2009). Fusion of Acoustic and Linguistic Speech Features for Emotion Detection. In *Proceedings of International Conference on Semantic Computing (ICSC)*.
- Nobuo, S. and Yasunari, O. (2007). Emotion Recognition using Mel-Frequency Cepstral Coefficients. *Information and Media Technologies*, 2:835–848.
- Polzehl, T., Schmitt, A., and Metze, F. (2010). Approaching Multi-Lingual Emotion Recognition from Speech - On Language Dependency of Acoustic/Prosodic Features for Anger Detection. In *SpeechProsody*, Chicago, U.S.A.
- Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., and Metze, F. (2009). Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. In *Proceedings of Interspeech*.
- Press, W.H. and Teukolsky, W., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge, University Press, 2nd edition.
- Rabiner, L. and Sambur, M. R. (1975). An Algorithm for Determining the End-points of Isolated Utterances. *The Bell System Technical Journal*, 56:297–315.
- Schmitt, A., Heinroth, T., and Liscombe, J. (2009). On NoMatches, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection? In *SIG-DIAL Meeting on Discourse and Dialogue*, London, UK. Association for Computational Linguistics.
- Schuller, B. (2006). *Automatische Emotionserkennung aus Sprachlicher und manueller Interaktion*. Dissertation, Technische Universität München, München.

- Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2009). Emotion Recognition from Speech: Putting ASR in the Loop. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Shafran, I. and Mohri, M. (2005). A Comparison of Classifiers for Detecting Emotion from Speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Shafran, I., Riley, M., and Mohri, M. (2003). Voice Signatures. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*, pages 31–36.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27.
- Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. PhD thesis.
- Steidl, S., Levit, M., Batliner, A., Nöth, E., and Niemann, H. (2005). "Of All Things the Measure is Man" - Classification of Emotions and Inter-Labeler Consistency. In IEEE, editor, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 317–320.
- Vapnik, V. and Cortes, C. (1995). Support Vector Networks. *Machine Learning*, 20:273–297.
- Vlasenko, B. and Wendemuth, A. (2007). Tuning Hidden Markov Model for Speech Emotion Recognition. In 36. *Deutsche Jahrestagung für Akustik (DAGA)*.
- Wells, J. C. (1982). *Accents of English*, volume 1–3. Cambridge University Press.
- Witten, I. and Frank, F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*.
- Yacoub, S., Simske, S., Lin, X., and Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. In *Eurospeech, Geneva*, pages 1–4.

Chapter 5

PSYCHOMIME CLASSIFICATION AND VISUALIZATION USING A SELF-ORGANIZING MAP FOR IMPLEMENTING EMOTIONAL SPOKEN DIALOGUE SYSTEM

Yoshiaki Kurosawa, Kazuya Mera, Toshiyuki Takezawa
*Graduate School of Information Sciences, Hiroshima City University
Hiroshima, Japan*

{ kurosawa, mera, takezawa }@ls.info.hiroshima-cu.ac.jp

Abstract We aim to automatically and appropriately classify Japanese psycho-mimes that represent users' emotional aspects, such as “*pokapoka*” and “*tikutiku*,” and visualize the classification results as a map by using a self-organizing map (SOM) algorithm as the basis of the implementing a spoken dialogue application with emotional agents. Dealing with psychomimes and visualizing their classification has become increasingly important because they reflect the speaker's emotions and frequently appear in communication, particularly in Japanese. However, it is difficult to communicate meaning to people who do not understand psychomimes because they are not directly perceivable. We experimentally classified psychomimes with SOM and represented the results as maps with significant three-vector dimensions, i.e., three verb classes assigned by a verb thesaurus dictionary. The experimental results demonstrated that our method was effective in accomplishing a precision rate for classification that was higher than 80% in almost every group attained by setting an adequate threshold according to the distribution of the SOM node data. Moreover, we demonstrated the importance of selecting not one or two verb classes but three to depict the classification results as a map that can express subtle differences in emotion.

Keywords: Onomatopoeia classification and visualization; Verb thesaurus.

1. Introduction

Recent developments in speech recognition technology have made it possible to apply various studies to the development of useful applications. When applied to task-oriented systems, these technologies have played key roles, including an operator that can accept hotel reservations (Price, 1990; Allen et al., 1995) and a guide that can give users directions (Takahashi and Takezawa, 2002; Nishimura et al., 2003; Gruenstein et al., 2006). There have been additional studies that have dealt with the development of emotional agents as well as task-oriented systems (Nisimura et al., 2006; Mera et al., 2010). These studies have included models for interpreting the agent's feelings and emotional factors. The users of the system seem to be able to communicate with the agents as if they were humans because they can talk and dance through these models.

However, agents that are capable of displaying such emotions are still in the phase of development in many respects and are confronted by many problems. One of these is that they can only deal with certain grammatical categories (such as nouns and verbs) in the speaker's utterances. Moreover, even if the proper categories are being used, they can only deal with words that have been manually given an emotional value beforehand. In other words, agents cannot interpret several emotional phrases (such as those in onomatopoeias or mimetic words, particularly, psychomimes) that represent the speaker's psychological or emotional states. Agents that cannot interpret these may irritate system users because psychomimes are frequently used in Japanese. Therefore, we need to develop agents that can interpret these in Japanese at the very least. Agents face an additional challenge in that psychomimes can be generated continuously and thus their meanings subsequently change, both continuously and contextually. We must be able to process such changes.

We focused on an artificial neural network algorithm, the self-organizing map (SOM) (Kohonen, 2001), and used it to classify psychomimes. The reason we used SOM was that it could classify data without having to prepare supervised information and it could depict the results as a visually meaningful map. This visual representation is very important because psychomimes represent speaker's emotions where we cannot directly observe precise meanings as will be described in the following section. Moreover, many sources have reported SOM can effectively classify given data (such as corpora) into several classes. Although these reports have certainly let us understand that SOM can effectively be used for various targets such as noun classification, it was not clear whether the algorithm could deal with psychomimes in the same way. Here, we discuss how effective SOM was in classifying psychomimes in our experiments.

This chapter is organized as follows. We briefly explain studies related to psychomimes and emotional spoken dialogue systems in Section 2. We summarize SOM and our method of using it in Section 3 and discuss our experimental evaluation in Section 4. Conclusions and our future work are presented in Section 5.

2. Psychomimes and Emotional Spoken Dialogue Systems

This section first defines psychomimes and compares them to the broader category of onomatopoeia. We then discuss related studies on emotional spoken dialogue systems that are necessary to process psychomimes.

2.1 Onomatopoeias and Psychomimes

Onomatopoeia, including psychomimes, represents figures of speech in which words represent sounds and objects states. Japanese particularly like to use onomatopoeia. Two examples of onomatopoeia that represent pain are

- | | | | | |
|----|-------------------|-------------------|-----------------|---------------|
| 1. | <i>Watashi no</i> | <i>yubi ga</i> | <i>tikutiku</i> | <i>itamu.</i> |
| | (noun + particle) | (noun + particle) | (psychomime) | (verb) |
| | = My | finger | | prickles. |
| | | | | |
| 2. | <i>Watashi no</i> | <i>yubi ga</i> | <i>zukizuki</i> | <i>itamu.</i> |
| | (noun + particle) | (noun + particle) | (psychomime) | (verb) |
| | = My | finger | | throbs. |

We cannot accurately determine the meanings of these two sentences because they have two different forms of onomatopoeia in Japanese and their translated sentences have different verbs in English. That is, we cannot understand them without processing the forms of onomatopoeia that represent emotional or psychological states. Therefore, an emotional spoken dialogue system must be capable of processing them, at least in Japanese.

Moreover, we need to categorize the forms of onomatopoeia into several groups according to how easily they can be processed. For example, it is easy to discern the meanings of sound-symbolic words because the word representation is based on sounds: *gangan* is comprehensible by hitting an object such as a tin bucket because its meaning is based on the sound. It is also easy to discern the meanings of general mimetic words because the word representation depends on directly perceivable objects, events, and states; *bukabuka* is visually able to be perceived because it is based on being buggy or too big. In

contrast, psychomimes like *tikutiku* and *zukizuki* in sentences (1) and (2) are difficult to discern because they are related to the speaker's emotional state and are not directly accessible to anyone except the speaker. Hence, we must develop a method of processing such figures of speech that can be implemented in an emotional spoken dialogue system.

We must also note that new forms of onomatopoeia are often generated, and that their meanings are not static but rather changeable. Thus, we must be able to deal with the characteristics of psychomimes in the same way.

2.2 Emotional Spoken Dialogue Systems

To the best of our knowledge, there have been few studies related to technology that can detect feelings. One such study, however, describes the Meipu system (Mera et al., 2010). Meipu is an animated agent who can interpret users' spoken input and respond to it by moving or speaking (Figure 1). She behaves like a human on the basis of human-emotion modeling or the emotion-generating calculations (EGCs) (Mera, 2003).



Figure 1. Animated agent called Meipu. It can respond to range of spoken utterances and perform various actions such as dancing and greeting with emotion based on emotion generating calculations.

An agent's emotion in EGCs is calculated with respect to each event or sentence. An event is analyzed as a case frame representation according to defined equations that consist of two or three grammatical terms: subject, object, and predicate. The representation is then divided into two categories, e.g., "pleasure" or "displeasure" using favorite values (FVs). An FV is the degree of like/dislike for the terms; it is given a real number on a scale of $[-1.0, 1.0]$ based on the results of a questionnaire.

The basic concept behind EGCs is a case frame representation, although we have made Meipu's actions correspond to her emotions. She can easily comprehend sentences (1) and (2) as shown in Section 2.1, in English because the verbs 'prickle' and 'throb' refer to emotional states and they are one of the EGCs grammatical terms (the predicate). However, she cannot comprehend psychomimes because they do not correspond to any EGCs terms in Japanese. That is, the possibility of calculation depends on the grammatical terms in the

defined case frame. The psychomimes in Japanese do not fulfill this condition. In addition, EGCs cannot automatically give a real number to newly generated or recently changed psychomimes. We therefore need to expand this method to process psychomimes.

We used an SOM algorithm to solve the problems related to emotional processing, to analyze the characteristics of psychomimes and to express the subtle differences in meaning.

3. Self-Organizing Map

3.1 What is SOM?

A self-organizing map (SOM) is an artificial neural network algorithm, first proposed by Teuvo Kohonen, in which a learning process could be automatically and appropriately performed without having to prepare supervised information (Kohonen, 2001).

This process is carried out between an input layer and a competitive layer, as shown in Figure 2. The former is for a vector input represented by vectors such as co-occurrence frequencies in target documents, and the latter is for learning and for depicting the result as a map.

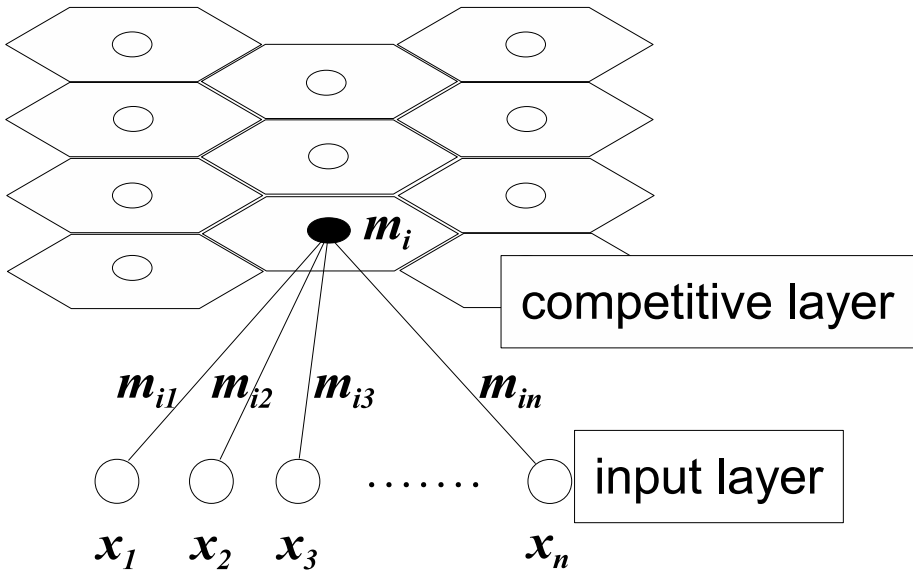


Figure 2. Basic concept behind the SOM algorithm. The x is an input vector, $x = \{x_1, x_2, \dots, x_n\}$. m is the model or reference vector from the i th node, like a neuron, drawn in an hexagonal shape. Each component represents a kind of weight attached to each input component in the input vector and the changes during the learning process.

In the learning stage, the SOM first tries to find special node c , the winner node, according to the equation below. In other words, it is capable of finding a best-match reference node, m_c , which is the most similar to input vector x in all nodes m_i :

$$\forall i, \|x - m_c\| \leq \|x - m_i\|. \quad (5.1)$$

Next, the SOM renews each component m_i in the selected vector and adjacent vectors around it so as to make themselves become more similar than the previous state according to

$$\begin{aligned} m_i(t+1) &= m_i(t) + h_{ci}(t)(x(t) - m_i(t)) & \text{if } \forall i \in N_c(t), \\ m_i(t+1) &= m_i(t) & \text{otherwise,} \end{aligned} \quad (5.2)$$

where $N_c(t)$ is a region called the “neighborhood” near a winner node in step or time t , which is an integer on a discrete-time scale, and $h_{ci}(t)$ is a neighborhood function for detecting the region. We use the following simple equation instead of the Gaussian function

$$\begin{aligned} h_{ci}(t) &= a(t), \\ a(t) &= a(0) \left(1 - \frac{t}{T}\right) & (0 < a(t) < 1), \end{aligned} \quad (5.3)$$

where value $a(t)$ is defined with a learning rate. This value is usually decreasing in time because it depends on t and T , which means the number of total steps for learning. Therefore, $h_{ci}(t)$ is also decreasing.

This means that the SOM processes not only the winner node but also some neighboring nodes, which become narrow and gradually decrease in number as iteration progresses. That is, the learning scope of SOM is broader in the earlier process and narrower in the latter process.

Let us consider some container data composed of two dimensions: width and height. If the width is broader and the height is lower, we call that container a dish or pan. If the width is narrower and the height is higher, on the other hand, we call that container a cup or glass. Now, suppose we have a case where an input vector looks like a pan, as shown in [Figure 3](#).

The input enables SOM to determine a winner node and temporarily select six neighborhood nodes. After this selection, the value of each selected area is updated according to Equation (5.2). For example, if the height of each node is greater than that of the winner, it decreases, while it increases if the height of each node is lower than that of the winner. After this update, the SOM receives the next input and repeats the process.

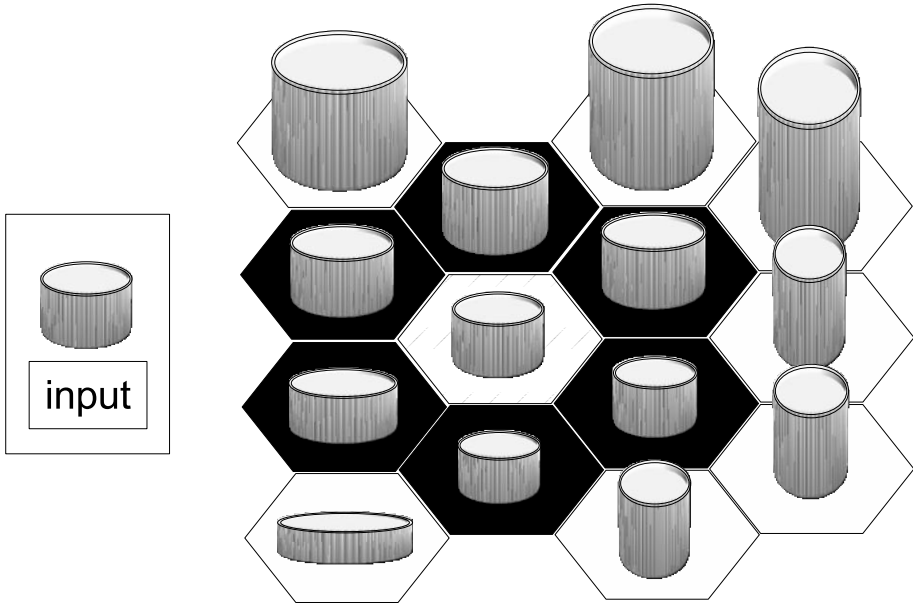


Figure 3. SOM sample case where the input is performed. The area with diagonal lines indicates winner node. The area with filled nodes near the winner indicates the neighborhood.

As a result, all the nodes gradually become similar, as seen in Figure 4. In other words, after the learning process, specific areas are composed of similar nodes with almost identical container-like characteristics.

As previously mentioned and seen from the Equation (5.3), the neighborhood region narrows and gradually decreases in size as iteration progresses. For this reason, the nodes around the center of the region are repeatedly affected, while the ones farther from the center are almost never affected. Such a distribution is then acquired, as seen in Figure 5, where we should regard a value or values next to 0.0 as a border and the area surrounded by the borders as a classified area. Thus, we can represent these values with grayscale or color intensity results in comprehensible classified groups; i.e., an SOM.

3.2 Natural Language Processing Studies using SOM

The SOM algorithm has been used in several studies on natural language processing, including the classification of various words and the categorization of large documents (Honkela et al., 1995; Kaski et al., 1998). Several studies have been conducted in the Japanese language from various viewpoints, in-

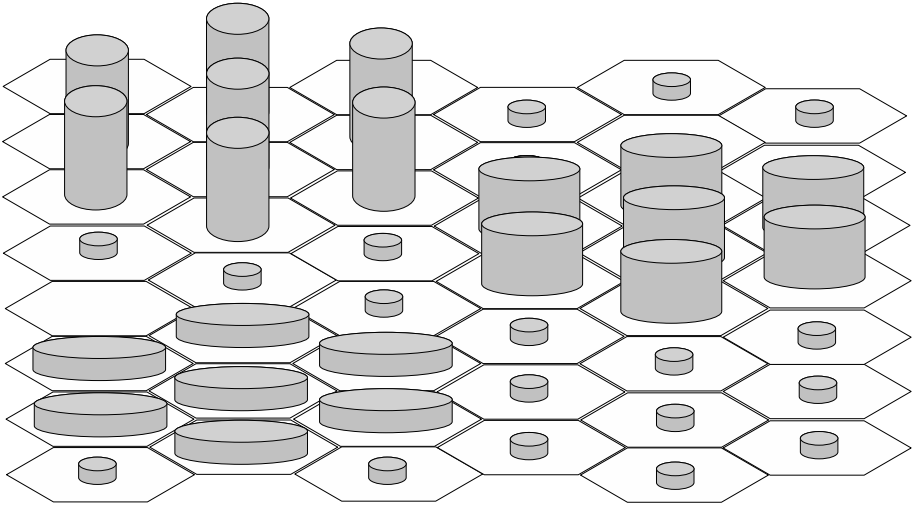


Figure 4. SOM sample results after the learning process is repeated. We can see three areas, i.e., three groups, which seem to be related to containers like a glass (top left), a pan (middle right), and a dish (bottom left).

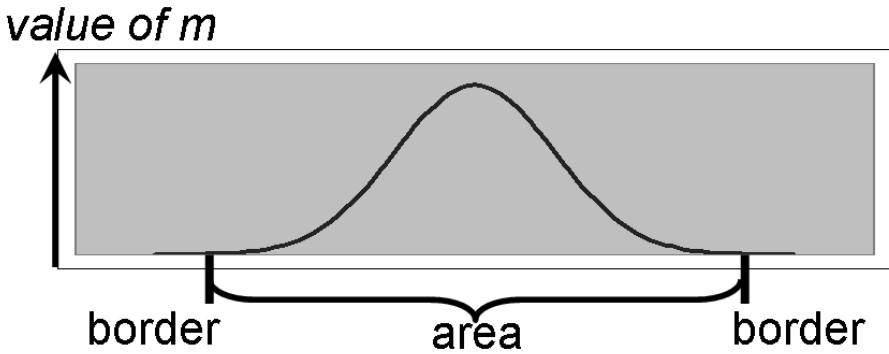


Figure 5. Simulated and virtual distribution of each acquired area. If Figure 4 is viewed from the side, each area should have such distribution. These data are presented as examples, and are not real data.

cluding semantic word maps (Ma et al., 2002), author maps (Jin, 2003), and object-form maps (Kurosawa et al., 2008).

Kurosawa et al. (Kurosawa et al., 2008), in particular, attempted to classify containers such as pans, bowls, and plates according to keywords related to the containers' objects. The keywords they used were heuristics in a sense

and made use of the classification of containers. For example, the keyword “*no soko*” (which means “bottom of”) and its co-frequencies between containers can be used to distinguish them because the keyword rarely appears when expressing a shallow object like a plate. Focusing on these keywords, they confirmed that similar containers were classified into the same group and were continuously arranged beside each other along some invisible axes (e.g., deep-shallow and small-large). It is important for SOM to be capable of arranging items according to such continuous data because bowls and plates do not contain distinct borders. Psychomimes also have continuous characteristics. For example, Akita defined two psychomimes “*pokapoka*” and “*hokahoka*” as ‘feeling pleasantly warm’ for the former and ‘feeling pleasantly hot’ for the latter (Akita, 2006), with their differences depending on their adjectives, i.e., their temperatures did not involve a distinct border. As we previously mentioned, SOM can deal with these types of data. Therefore, we need to ensure that this algorithm can be adapted to our needs, i.e., to classify psychomimes.

4. Experiment

We conducted an experiment to classify psychomimes with SOM. We will explain the procedure, results, and evaluation in this section. Three viewpoints are presented particularly in terms of evaluation:

- Is SOM capable of classifying psychomimes precisely?
- How many vector dimensions does SOM require before it can provide results?
- Do the results become more accurate?

4.1 Psychomimes and their Groupings

We adopted 228 psychomimes determined by Akita (Akita, 2006) and compiled sentences in which the psychomimes occurred as a corpus (as will be explained in the next section). We particularly focused on psychomimes directly preceding verbs because it was highly likely that such psychomimes would modify the verb.

To compare the classification results derived by SOM, we manually defined groups on the basis of Akita’s definitions. Although some may feel this procedure was subjective, the definitions were so simple that there was little room for arbitrary interpretations, e.g., because “*kurakura*” is defined as ‘feeling dizzy being attracted’ and “*furafura*” is defined as simple ‘being dizzy’, we can classify both of them into the same group (“Dizzy”). Thus, we can objectively classify the psychomimes by focusing on specific words, like ‘dizzy’, that occur in definitions. Our defined groups based on this procedure are listed in [Table 1](#).

Table 1. Our manually classified psychomime groups based on certain words in the Akita’s definitions.

Group Name	Ico	Num. of Mimes	Example Focus Words
Dizzy		9	eye, dizzy
Heart Beat		9	throb, heart
Pain		9	sore, pain, pang, gripping
Smell		5	smell
Temperature		14	cool, cold, hot, warm
Stimulus		9	pungent, skin
Others		64	

4.2 Corpus

We used a Google n-gram model (7-gram data) as a corpus (Kudo and Kazawa, 2007). All sentences in the original data were composed of seven morphemes determined by morphological analysis using the part-of-speech tagger in Figure 6a. However, the psychomime “*gangan*” was divided erroneously because it was unregistered in the morpheme database. To avoid these errors, we created a sequential letter string connecting the seven morphemes (Figure 6b), and we used “ChaSen” (Matsumoto et al., 2000) as a part-of-speech tagger to morphologically analyze all the strings. These strings were improved so that we could precisely analyze the psychomimes by registering them. We could thus acquire more precise morphemes (Figure 6c) and psychomime-verb combinations.

4.3 Vector Space

We need to represent data as vectors with n dimensions when performing SOM. Some specific components for the representation, such as frequencies of words, are generally designated as the vector space of the target task. We particularly focused on the co-occurrence frequency between verbs and the psychomimes that preceded them: ‘*gangan*’ and ‘work’ in Figure 6c, ‘*tikutiku*’ and ‘*itamu*’ in sentence (1), and ‘*zukizuki*’ and ‘*itamu*’ in sentence (2).

However, the number of co-occurrence combinations in our corpus was so enormous that the vector dimensions of components also became enormous. Thus, we categorized the verbs using a verb thesaurus dictionary (Takeuchi et al., 2008) to decrease the size. The thesaurus was an example-based one in which verb meanings were hierarchically classified. We particularly used the lowest hierarchy, called a ‘frame’. The verbs *ugoku* (move) and *shifutosuru* (shift) in this classification were included in the same frame (“change of position”). In other words, we could deal with some verbs as the same component

This decrease should certainly enable SOM to learn its map more easily and efficiently. However, such a vector component based on frequency may not work well because SOM is sensitive to differences between component values, as we found from the previous study (Kurosawa et al., 2008). Therefore, we adopted the following p_i , the co-occurrence ratio of the frames per psychomime, in Equation (5.4):

$$p_i = \frac{v_i}{\sum_{i=1}^n v_i}, \quad (5.4)$$

where the value v_i means the co-occurrence frequency that emerged in the i th frame. Table 2 lists example data including two psychomimes with five frames. All values (v_i and p_i) in this figure are the same as those in Equation (5.4).

Table 2. Example table comparing frequency and ratio. The latter one is superior to the former one because of homogeneity.

Psychomime		Frame i					$\sum_{i=1}^5 v_i$
		1	2	3	4	5	
<i>pokapoka</i>	frequency v_i	22	1	5	23,503	2	23,533
	ratio (%) p_i	0.09	0.00	0.02	99.85	0.01	-
<i>hokahoka</i>	frequency v_i	4	0	0	108	0	112
	ratio (%) p_i	3.57	0.00	0.00	96.43	0.00	-

The frequency of “*pokapoka*” is 23,503 and the frequency of “*hokahoka*” is 108 in frame 4. In such cases, the difference between both values is so large that SOM cannot classify them into the same group. However, because these words both have similar ratios (99.85 and 96.43), SOM classifies them into the same group.

4.4 SOM Parameters

We performed a two-stage experiment using a SOM tool called a som_pak (Kohonen, 2001). This subsection describes the SOM parameters we adopted, which were determined through preliminary experiments.

The parameters for the first stage were the number of steps, $T = 50,000$, and the learning rate, $a(0) = 0.05$ in Equations (5.2) and (5.3). The map dimensions were 64 and 48 and the initial neighborhood size was 32. However, the parameters for the second stage were $T = 500,000$ and $a(0) = 0.01$. The map dimensions were also 64 and 48, but the initial neighborhood size was 21.

The neighborhood size of this stage was smaller and the number of steps was higher than in any of the parameters for the first stage. These parameters in the second stage were to reduce the run-time cost of iteration for learning because iteration would take more time. If we used them through all the iterations, the necessary time would have increased although the results might have been better. That is, the first stage was to avoid iteration and aim at enable learning to occur globally and roughly by restricting the parameters. In contrast, the

second stage was carried out to acquire more refined learning results, even though SOM would need more computer resources.

4.5 Results

The map acquired by SOM is shown in [Figure 7](#). We added icons so as to easily identify the psychomime groups in [Table 1](#). Several psychomimes with icons are arranged in the corner at right. This is one of the characteristics of this algorithm, i.e., learning is faster around the corner, while learning is slower at the center of the map. At a glance, this map seems to be classified well to the right of the center because the same icons are closely arranged beside each other (e.g., triangles $\triangle$ and spades $\spadesuit$). That is, the psychomimes defined as “Temperature” are collected at the center of the right side. [Figure 8](#) is a magnification of [Figure 7](#).

The classification results become clearer in [Figure 8](#) because the displayed grayscale nodes look like lines and an area surrounded by the lines seems to come into existence. This might seem like a subjective conclusion; however, the reasoning behind it is that the borders and so-called inner area are arbitrary. This indicates we must clearly determine group areas with distinct borders.

4.5.1 Determination of Group Areas. We needed to clearly determine group areas for an objective evaluation; however, the separate areas did not seem to have distinct borders ([Figures 7](#) and [Figure 8](#)). We focused on significant frames assigned by using a verb thesaurus to divide the areas into appropriate groups. First, we selected three from 165 frames. We will detail the selection process in the next section.

These three frames were then changed so that they ranged from 0.0 to 1.0, according to their distributions. Values over 0.0 were identified as data and given color representations. We also mixed colors where nodes consisted of more than two frames. Examples of these representations are given in [Figures 9](#) and [10](#). The three-color illustration enabled us to determine the separate group areas.

For example, in [Figure 10](#), the psychomimes, surrounded by rectangles, belong to the “Temperature” group previously mentioned. The green (left side) area generally consists of cool psychomimes, while the pink area (center of the right side), which is a mix of blue and red, consists of hot ones. This means that SOM could detect a specific group divided into two sub-groups. Moreover, we can detect slightly different meanings within hot psychomimes because two areas, red (lower area of the right side) and pink (the right side), in the hot-psychomime areas have subtle differences in nuance. This means SOM and its visualization technique can classify the psychomimes hierarchically through overlapping areas as in a Venn diagram. That is, we can accurately express the subtle differences in emotion included in the psychomime group. This is a

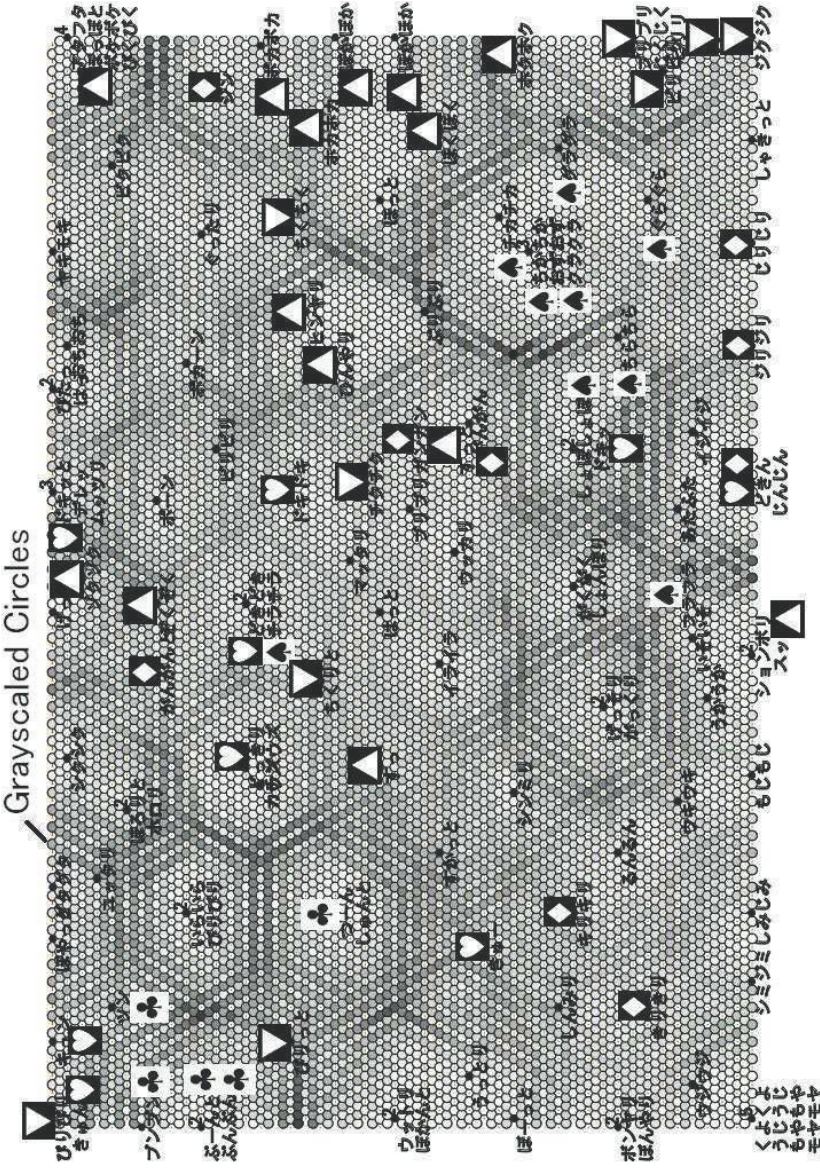


Figure 7. Our acquired map. Each of the smallest circles (labeled 'Grayscaled Circles') represents each reference node. Their grayshades correspond to the mean distance around nodes. Closed circles (black) that are far from neighboring nodes indicate the area that includes similar psychomimes.

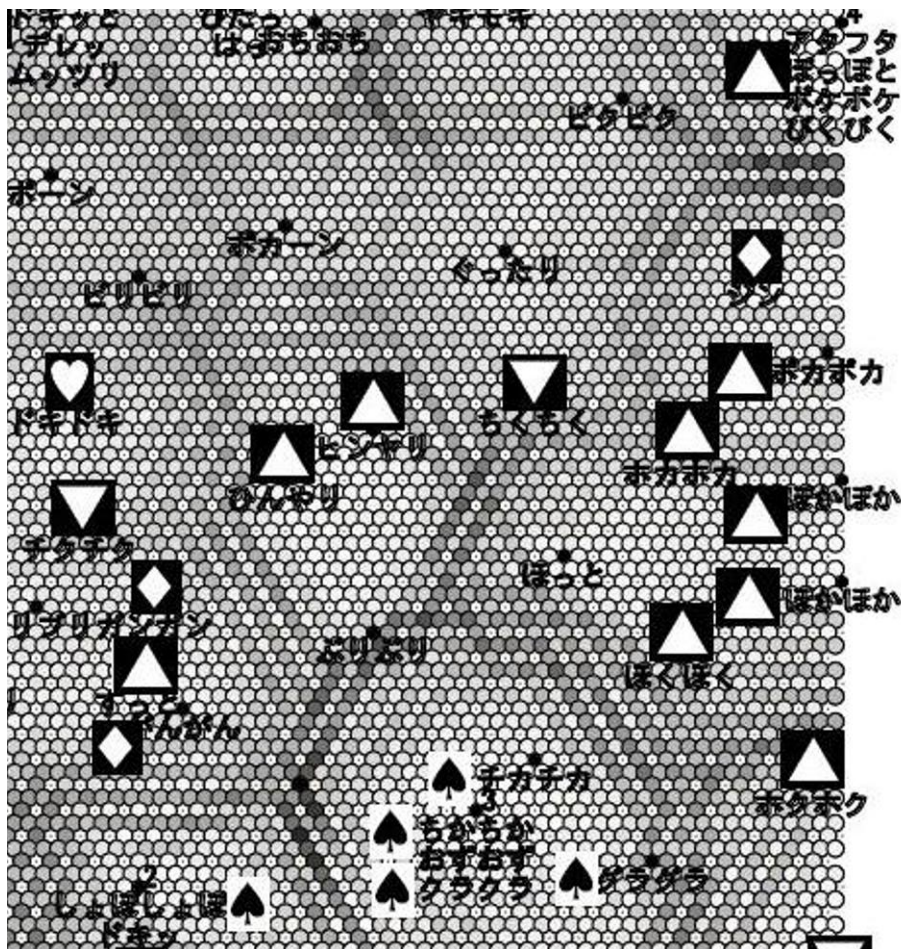


Figure 8. Details of map in [Figure 7](#).

very crucial aspect because psychomimes need to include nuances in emotion. This is why we adopted SOM and selected three frames.

4.5.2 Recall and Precision. Once we determine the group area in the SOM map, as described in the previous section, we can calculate various measures, including precision and recall, using the area for objective testing as follows:

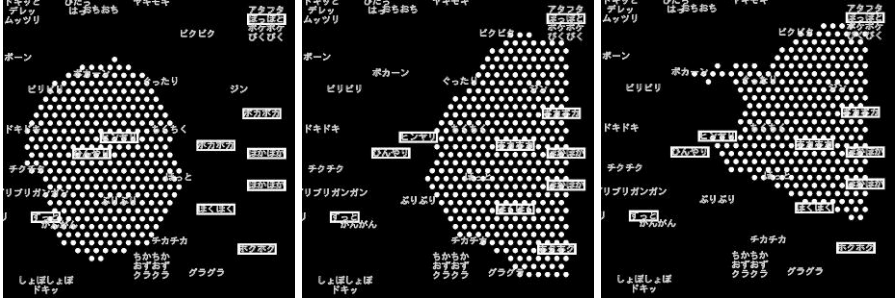


Figure 9. SOM maps to determine each group area using three frames: green, red and blue. They were drawn from the same region used in Figure 8. In fact, these three areas overlapped in original color map and were only represented as one comprehensible map. However, we involuntarily divided it into three sub-maps. Reason we did this was that it was difficult to depict map for publication by using grayscale representation.

$$\begin{aligned}
 \text{precision} &= \frac{N_{\text{appropriate}}}{N_{\text{area}}}, \\
 \text{recall} &= \frac{N_{\text{appropriate}}}{N_{\text{group}}}, \\
 F &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},
 \end{aligned} \tag{5.5}$$

where N_{area} means the number of all psychomimes included in the colored area, N_{group} means the number of psychomimes (with open rectangles) included in the target group, e.g. “Temperature,” and $N_{\text{appropriate}}$ means the number of appropriately classified psychomimes. In this equation, a psychomime is only counted once as unique, even if it appears in the overlapping area. That is, the whole areas are regarded as a union in which two sets overlap although a different weight can be given to the overlapping area, which is not our present concern.

We focused on two maximization values to use our experimental knowledge to detect the group area: the F-measure and precision. First, we calculated the F-measure in Equation (5.5) and selected three frames such that the F-measure was maximized. The calculated results are listed in Table 3. The F-measures in each group were also calculated after some frames had been omitted for two main reasons:

- frames that co-occurred with less than or equal to two psychomimes;
- frames that co-occurred with more than 30 psychomimes.

Frames in the first description were omitted because we may not see the relation between frames without overlapping areas. Moreover, we needed to



Figure 10. This map is not our system output but figure drawn for this section. The mock figure represents Figure 9 as only one map because three maps make it difficult to discuss our results. In this figure, letters ‘R’, ‘G’, and ‘B’ mean original colors ‘Red’, ‘Green’, and ‘Blue’. Letters ‘W’, ‘P’, ‘Y’, and ‘L’ mean overlapping colors of ‘White’, ‘Pink’, ‘Yellow’, and ‘Light-blue’. For example, area labeled ‘P’ is where two colors (‘Red’ and ‘Blue’) overlap.

avoid group classifications that only included one or two psychomimes because the precision of such groups may have been 100%, i.e., 1/1 or 2/2. The reason frames in the second description were omitted is that we also needed to avoid acquiring a huge area because the recall may have been 100%. In this study, we set 30 as a threshold that was twice the maximum number of psychomimes (14). As a result, the number of frames decreased to 120.

Table 3. Precision and recalls when F-measure was maximized in each group. The upper value in a cell is the precision/recall value, while the lower one is the number of psychomimes per calculation.

	Dizzy	Heart	Psychomime Group			
			Pain	Smell	Temp.	Stimu.
precision (%)	46.2	22.2	33.3	25.0	45.8	45.5
	6/13	6/27	7/21	4/16	11/24	5/11
recall (%)	66.7	66.7	77.8	80.0	78.6	55.6
	6/9	6/9	7/9	4/5	11/14	5/9

Maximization of each F-measure

When the maximum F-measure was adopted, the precisions were low while the recalls were relatively higher, except in the “Stimulus” group. Given that the number of psychomimes was greater than 20 in the three groups (“Heart”, “Pain”, and “Stimulus”), this tendency may have resulted from the number in each group, where some incorrect psychomimes were included. That is, better recalls, caused by expanding each group, naturally resulted in worse precision results because of their inverse relationship.

Maximization of each precision

Next, we focused on precision by selecting three frames so that we could maximize it. The calculated results are listed in [Table 4](#) after the frames were omitted in the same way. As a result of focusing on precision, these values were clearly better than those in the previous results, but the tendency was almost the same, i.e., the precision results were low and the recalls were relatively higher.

Table 4. Precision and recalls when maximizing precision in each group.

	Dizzy	Heart	Psychomime Group			
			Pain	Smell	Temp.	Stimu.
precision (%)	46.2	22.7	40.0	25.0	50.0	57.1
	6/13	5/22	4/10	4/16	9/18	4/7
recall (%)	66.7	55.6	44.4	80.0	64.3	44.4
	6/9	5/9	4/9	4/5	9/14	4/9

As we described in the two experimental evaluations, the precision results were lower although the recalls were higher. The reason for the lower preci-

sion was due to manual classification based on Akita's definition. We classified each psychomime into only one group because of the calculations for precision and recall in this study. This manipulation, however, might be too strict, because psychomimes have various ambiguities and may need to be classified into more than one group. We need to reconsider our manually defined psychomime groups.

Because of the lower precision, we feel our proposed method is more suitable to tasks related to extracting psychomimes rather than evaluating them even though they were not typical in these groups. If this is correct, there are those who would naturally assume that three frames are needed to draw certain maps. We will discuss this next, with a focus on the maximization of both F-measures and precision results.

4.5.3 Effects of Selecting Frames and Combinations of Frames.

We adopted three frames to illustrate the classification results, which contained subtle differences in nuance. However, it was not clear whether three was an appropriate value. We can depict the results as a map even if we only use one vector component. Four or more components may alternatively be needed. We discuss the appropriate number of frames in this section.

Maximization of each F-measure

We have listed the calculated values that change the number of needed frames in [Table 5](#). 'The number of frames' in this table means how many are needed to depict the classification results as a map; "1" means only one frame is needed to do so.

Focusing on row "1" for both precision and recall, appropriate psychomimes were found in the four to eleven range due to the use of one frame. This resulted from our adopting the frequency of frames instead of the verbs themselves. In short, our frame representation was effective.

Taking the results into consideration in detail by comparing the precision row "1" with row "3" in [Table 5](#) makes it difficult to conclude that the latter one is superior to the former one. This is because three groups were better ("Dizzy," "Heart," and "Stimulus") and the rest were worse ("Pain," "Smell," and "Temperature"), although the increase in the latter might have been more abrupt than that in the others. In contrast, when these rows were compared in recall, row "3" was superior to the former one, because the values in almost every group increased.

Moreover, row "3" in the precision was comparable to row "5." Only one group was better ("Pain"), and one group was worse ("Smell"), while the others all had the same value. This did not indicate the difference between the two factors related to the number of frames. In contrast, when we focused on the

Table 5. Precision and recalls calculated by changing needed number of frames (only one, three, or five) when maximizing F-measure in each group. Results using three frames are the same as those in Table 3.

	Num.of Frames	Psychomime Group					
		Dizzy	Heart	Pain	Smell	Temp.	Stimu.
precision(%)	1	21.1	20.0	37.5	28.6	50.0	36.4
		4/19	3/15	6/16	4/14	7/14	4/11
	3	46.2	22.2	33.3	25.0	45.8	45.5
		6/13	6/27	7/21	4/16	11/24	5/11
	5	46.2	22.2	35.3	22.7	45.8	45.5
		6/13	6/27	6/17	5/22	11/24	5/11
recall (%)	1	44.4	33.3	66.7	80.0	50.0	44.4
		4/9	3/9	6/9	4/5	7/14	4/9
	3	66.7	66.7	77.8	80.0	78.6	55.6
		6/9	6/9	7/9	4/5	11/14	5/9
	5	66.7	66.7	66.7	100.0	78.6	55.6
		6/9	6/9	6/9	5/5	11/14	5/9

recall, only one group was better (“Smell”), and one group was worse (“Pain”), while the others all had the same value. This was just an inverse relation when comparing precision and therefore did not reveal the difference between the two factors.

However, the numbers of combinations when using five frames was so enormous that they were too costly to calculate. We could not find any differences between both factors from our experimental data with any degree of certainty, but the cost of calculation was a legitimate reason to take them into account and to choose one of them. Consequently, we concluded that three-frame representation should be adopted.

Maximization of each precision

In the same way, we have listed the calculated values changing the number of required frames in Table 6.

Comparing precision row “3” with row “1” seemed to indicate that “1” was relatively worse; three groups were worse (“Dizzy,” “Heart,” and “Stimulus”), two were the same (“Pain” and “Temperature”), while one was better (“Smell”). It might seem difficult to conclude that row “3” was superior to row “1” because there were no significant differences. However, the key point is that this comparison illustrates an unexpected transition. That is, the transition of frames occurred in the “Dizzy” and “Stimulus” groups as it is clear that

Table 6. Precision and recalls calculated by changing the required number of frames when maximizing the precision in each group. Results using three frames were the same as those in Table 4.

	Num.of Frames	Psychomime Group					
		Dizzy	Heart	Pain	Smell	Temp.	Stimu.
precision (%)	1	21.1	20.0	50.0	28.6	50.0	36.4
		4/19	3/15	4/8	4/14	7/14	4/11
	3	46.2	22.7	40.0	25.0	50.0	57.1
		6/13	5/22	4/10	4/16	9/18	4/7
	5	46.2	22.7	40.0	23.5	50.0	57.1
		6/13	5/22	4/10	4/17	9/18	4/7
recall (%)	1	44.4	33.3	44.4	80.0	50.0	44.4
		4/9	3/9	4/9	4/5	7/14	4/9
	3	66.7	55.6	44.4	80.0	64.3	44.4
		6/9	5/9	4/9	4/5	9/14	4/9
	5	66.7	55.6	44.4	80.0	64.3	44.4
		6/9	5/9	4/9	4/5	9/14	4/9

the values of the denominator were decreased from 19 and 11 (in row “1”) to 13 and 7 (in row “3”) , respectively. As previously mentioned, the area is regarded as a union, and for this reason the increase in the number of the frames should generally increase the number of psychomimes. Despite this, the denominators were actually decreased in two groups. This decrease means that the results when selecting one frame may be a local solution rather than an optimal solution because more appropriate values result from the other frames. This indicates that we should avoid local solutions. For this reason, we should adopt three-frame representation.

This type of decrease is also seen in the “Dizzy” group in Table 5. The three-frame representation may result in more accurate results in the same way.

Moreover, it is important to discuss whether we should avoid local solutions, particularly when adapting an evolutionary computing method, for example. An ‘automatically defined groups’ method (ADG), which is a kind of genetic programming (GP), is capable of automatically classifying data into accurate groups in the same way as discussed in this chapter (Hara et al., 2008; Hara et al., 2005). To avoid local solutions, this method needs to make use of ‘mutation’ steps where a completely different combination is often and randomly selected. The probability of ‘mutation’ needs to increase or decrease according to the given data. As we previously mentioned, our data were open to local solutions and needed global searches. Thus, our experimental results indicate

that the increase in mutation probability seems to result in proper classification if evolutionary computing methods like ADG are adapted to our psychomime data.

However, when we focused on the recall, the three-frame representation was superior to one-frame representation because three groups were better (“Dizzy,” “Heart,” and “Temperature”) while the others all had the same value. Therefore, we should adopt three-frame representation.

By comparing precision row “3” with precision row “5”, it is uncertain which factors were useful because all values were the same. However, we should adopt three-frame representation, as previously mentioned in the discussion on maximizing the F-measure.

The details discussed above led us to conclude that three-frame representation could effectively be adapted to our data, given sufficient conditions, and that it was useful for depicting the results as maps using RGB intensity. By using a three-frame representation technique, we will attempt to obtain more precise results.

4.5.4 Effects of Narrowing Area of Groups. As previously mentioned, when the areas were determined, it was possible to calculate measures such as precision: however, one issue remains when detecting the areas. That is, our corpus data might include some peculiar psychomime-verb combinations related to tagging errors (Figure 6). If a morpheme is erroneously regarded as a verb, an incorrect psychomime-verb combination is observed. Some psychomimes areas may possibly cause errors.

If this assumption is correct, psychomimes that include this type of error should be arranged near the border (Figure 5) because these psychomimes and their co-frequencies rarely appear. Figure 10 seems to indicate that there are many incorrect psychomimes around the border. Therefore, to address such incorrect combinatorial effects, we adopted another value, 50% in this study as seen in Figure 11, as a threshold to determine new restricted areas where incorrect psychomimes were pushed aside.

By adopting the threshold value, the area was restricted as a result of omitting nodes around the border and more reliable nodes could be drawn replacing them with more reliable nodes, as shown in Figure 12. The calculated values are listed in Tables 7 and 8. In addition, the number of frames was again set to three.

The precision results in both figures were better; in particular, four of them (“Dizzy,” “Smell,” “Temperature,” and “Stimulus”) were over 70% and two of them (“Temperature,” and “Stimulus”) reached 100% although there were few psychomimes. This suggests that SOM can be a useful tool for classifying psychomimes if given an appropriate threshold and restricted areas. The recall figures, on the other hand, were worse except for the “Smell” group. In terms

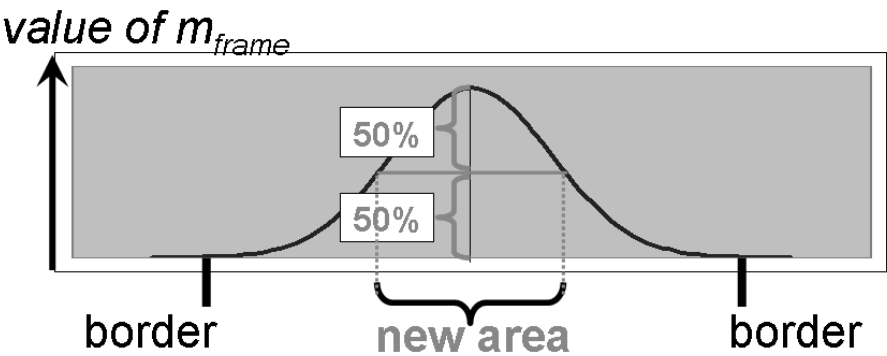


Figure 11. Function of the threshold based on simulated and virtual distributions in Figure 5.

Table 7. Precision and recalls when maximizing the F-measure in each group.

Psychomime Group with threshold 50%						
	Dizzy	Heart	Pain	Smell	Temp.	Stimu.
precision (%)	75.0	42.9	66.7	80.0	70.0	80.0
	3/4	3/7	6/9	4/5	7/10	4/5
recall (%)	33.3	33.3	66.7	80.0	50.0	44.4
	3/9	3/9	6/9	4/5	7/14	4/9

Table 8. Precision and recalls when maximizing the precision in each group.

Psychomime Group with threshold 50%						
	Dizzy	Heart	Pain	Smell	Temp.	Stimu.
precision (%)	75.0	42.9	75.0	80.0	100.0	100.0
	3/4	3/7	3/4	4/5	4/4	3/3
recall (%)	33.3	33.3	33.3	80.0	28.6	33.3
	3/9	3/9	3/9	4/5	4/14	3/9

of this tendency in recalls, the previous procedures without using a threshold were superior to the one related to restriction mentioned here.

Moreover, taking the number of the determined psychomimes between both precision results into consideration, their tendencies were different although the precision results were particularly higher in the “Temperature” group, i.e., it was ‘4/4’ for maximizing precision, while it was ‘7/10’ in maximizing the

traction? This is obviously a trade-off. Although we focused on both maximizing the F-measure and precision to calculate the precision and recalls, we also need to focus on other types of information, e.g., the number of psychomimes, the similarity between the three frames, and the definitions of equations to optimize classification and visualization. Therefore, it is not easy to determine which method is most useful because such determination depends on the target application. We will explain this in greater detail later by focusing on the relation between precision and recall.

4.5.5 How to Take Advantage of Knowledge. This subsection discusses ways of taking advantage of the knowledge we have described here.

In fact, it is difficult to use these data to determine the appropriate threshold because there is a trade-off between precision and recall, as [Tables 3 and 4](#) illustrate, i.e., as precision increases, recall decreases. Moreover, as previously mentioned, there is also a trade-off between selecting whether to maximize the F-measure or precision. We therefore need to consider the purpose this method is being used for. For example, if the intended purpose is to present new psychomimes to native Japanese speakers, lower precision may be acceptable because they will clearly notice incorrect classifications. For Japanese learners who do not know psychomimes, however, lower precision may result in incorrect learning.

We can adapt this notion to an emotional agent in the same way. That is, when the agent needs to issue an utterance including psychomimes, lower precision may be acceptable because users will clearly notice incorrect or atypical usage. When the agent needs to understand what users are saying, on the other hand, higher recall should be adopted, because lower precision would almost certainly result in unnatural communication between the agent and users.

4.5.6 Toward Implementing Emotional Spoken Dialogue System.

We know that SOM can classify psychomimes by focusing on the co-occurrence between psychomimes and verbs. If FVs assigned to verbs are already registered during the process of EGCs, its extension algorithm is easy to produce: for example, by using weighted calculation. Even though verbs are not registered in the case frame database, we can use the verb thesaurus in the same way.

Moreover, intensity or norm calculations are easy to carry out because each psychomime represents itself as a vector. Thus, our proposed method of using SOM can be effective for a broad range of emotional spoken dialogue systems although we have not yet implemented a specific system.

Another important detail is that this method is not speech-based but rather text-based. Text-based systems seem to perform better than speech-based ones,

but even so they frequently result in morphological failure; therefore, we need additional preprocessing. When we expand our method to a speech-based approach, we will also need different types of preprocessing. For instance, we must implement error-detection and correction modules related to automatic speech recognition.

5. Conclusions and Future Work

We proposed a classification method using the SOM algorithm and a visualization method using a three-vector-component representation. Our experimental results indicated that SOM could hierarchically classify psychomimes with a maximum precision of 80% if given an appropriate threshold. We also found three-vector-component representation was a necessary and sufficient condition for depicting classification results as a map, which had subtle nuances after we analyzed our experimental data. We thus concluded that SOM could be a useful tool for classifying psychomimes and our proposed method was effective for doing this.

In future research, we intend to look at ways of classifying psychomimes, which are often ambiguous, into multiple groups. We also need to adapt other vector spaces, such as latent semantic analysis and probabilistic latent semantic analysis to obtain greater precision, and to optimize appropriate vector components for drawing color-maps through experiments and analyses by controlling factors, such as the number of psychomimes in a selected area and the similarity of three frames.

After these investigations are completed, we intend to implement an emotional spoken dialogue system that can process psychomimes as people do in human-human communication.

References

- Akita, K. (2006). Embodied Semantics of Japanese Psychomimes. In *Proceedings of the 30th Annual Meeting of Kansai Linguistic Society*, pages 45–55.
- Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. R. (1995). The TRAINS Project: A Case Study in Building a Conversational Planning Agent. *Journal of Experimental and Theoretical AI*, 7:7–48.
- Gruenstein, A., Seneff, S., and Wang, C. (2006). Scalable and Portable Web-Based Multimodal Dialogues Interaction with Geographical Databases. In Stern, R. M., editor, *Proceedings of Interspeech*, pages 453–456.
- Hara, A., Ichimura, T., and Yoshida, K. (2005). Discovering Multiple Diagnostic Rules from Coronary Heart Disease Database Using Automatically Defined Groups. *Journal of Intelligent Manufacturing*, 16(6):645–661.

- Hara, A., Kurosawa, Y., and Ichimura, T. (2008). Automatically Defined Groups for Knowledge Acquisition from Computer Logs and Its Extension for Adaptive Agent Size. In Castillo, O., Xu, L., and Ao, S.-I., editors, *Trends in Intelligent Systems and Computer Engineering*, pages 15–32. Springer.
- Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual Relations of Words in Grimm Tales Analyzed by Self-Organizing Map. In Dorronsoro, J. R., editor, *Proceedings of International Conference on Artificial Neural Networks*, volume 2, pages 3–7.
- Jin, M. (2003). Authorship Attribution and Feature Analysis Using Frequency of JOSHI with SOM. *Mathematical Linguistics*, pages 369–386. (in Japanese).
- Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM - Self-Organizing Maps of Document Collections. *Neurocomputing*, 21:101–117.
- Kohonen, T. (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin Heidelberg, 3 edition.
- Kudo, T. and Kazawa, H. (2007). *Web Japanese N-gram Version 1*. Gengo Shigen Kyokai. (in Japanese).
- Kurosawa, Y., Hara, A., and Ichimura, T. (2008). Container-Form-Map Generation using Self-Organizing Map for Detecting Container-Content Metonymy. In Kodama, K. and Koyama, T., editors, *Linguistic and Cognitive Mechanisms*, pages 353–374. Hitsuji Syobo, Tokyo. (in Japanese).
- Ma, Q., Kanzaki, K., Zhang, Y., Murata, M., and Isahara, H. (2002). Self-Organizing Chinese and Japanese Semantic Maps. In Tseng, S.-C., Chen, T.-E., and Liu, Y.-F., editors, *Proceedings of the 19th International Conference on Computational Linguistics*, pages 605–611.
- Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., and Asahara, M. (2000). *Morphological Analysis System ChaSen 2.4.0 Users Manual*. Nara Institute of Science and Technology. <http://sourceforge.jp/projects/chasen-legacy/docs/chasen-2.4.0-manual-en.pdf/>.
- Mera, K. (2003). *Emotion Orientated Intelligent Interface*. PhD thesis, Tokyo Metropolitan Institute of Technology, Graduate School of Engineering.
- Mera, K., Ichimura, T., Kurosawa, Y., and Takezawa, T. (2010). Mood Calculating Method for Speech Interface Agent based on Mental State Transition Network. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 22(1):10–24. (in Japanese).
- Nishimura, R., Nishihara Y., Tsurumi R., Lee A., Saruwatari H., and Shikano K. (2003). Takemarukun: Speech-Oriented Information System for Real World Research Platform. In Tanaka, H., Furui, S., Nakajima, M., Shirai, Y., and Tsutiyu, S., editors, *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 77–78.

- Nisimura, R., Omae, S., Kawahara, H., and Irino, T. (2006). Analyzing Dialogue Data for Real-World Emotional Speech Classification. In Stern, R. M., editor, *Proceedings of Interspeech*, pages 1822–1825.
- Price, P. (1990). Evaluation of Spoken Language System: the ATIS Domain. In Stern, R. M., editor, *Proceedings of DARPA Speech and Natural Language Workshop*, pages 91–95.
- Takahashi, K. and Takezawa, T. (2002). An Interaction Mechanism of Multimodal Dialogue Systems. *Systems and Computers in Japan*, 33(11):70–79.
- Takeuchi, K., Inui, K., Takeuchi, N., and Fujita, A. (2008). Fine Grained Classification of Verb Argument Structures Based on Inclusion Relation of Senses. In Association for Natural Language Processing, T., editor, *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 1037–1040. (in Japanese).

Chapter 6

TRENDS, CHALLENGES AND OPPORTUNITIES IN SPOKEN DIALOGUE RESEARCH

Michael McTear

*University of Ulster, Computer Science Research Institute
Ulster, Northern Ireland*

mf.mctear@ulster.ac.uk

Abstract Spoken dialogue technology has developed considerably over the past thirty years both in terms of research activity as well as in terms of commercially deployed applications. This chapter presents an overview of trends in dialogue research based on an analysis of papers that were presented at Eurospeech-Interspeech conferences in 1989, 1999, and 2009. Following this some challenges are identified, in particular, the issues faced by individual researchers and those in small groups who wish to build viable end-to-end dialogue systems and the difficulties that are often encountered by academic researchers when trying to access the many resources and toolkits that are required for dialogue research. The chapter concludes with a discussion of some opportunities for future research, including the integration of dialogue technology into voice search applications and the application of spoken dialogue technology in ambient intelligence environments.

Keywords: Spoken dialogue systems; Voice search; Ambient intelligence.

1. Introduction

The vision of being able to interact with machines using speech goes back a long way but it is only within the past few decades that this dream has become a reality with the emergence of spoken dialogue systems. Progress has been made on two fronts: on the one hand, in terms of research into the computational modelling of spoken dialogue, and on the other, in terms of an increasing number of commercially deployed systems. The aim of this chapter is to examine how spoken dialogue technology has developed over the past thirty years,

to describe some challenges facing dialogue researchers, and to outline some opportunities and avenues for future research.

There has been extensive research activity over the past three decades in terms of funded projects, workshops, special conference sessions, special journal issues, and books. There have also been major developments on the commercial front with the emergence of deployed applications in areas such as call centre automation and voice search on mobile phones. On the research front there has been a healthy competition between those researchers who rely on predominantly hand-crafted, rule-based methods and those who favour data-driven approaches and machine learning. Furthermore, there has been a move away from contrived and restricted applications that aim to demonstrate a particular theory or methodology towards more realistic and more useful systems that can be deployed in everyday environments. These developments in spoken dialogue systems research over the past three decades are substantiated through an analysis of papers on dialogue presented at Eurospeech-Interspeech, the main international conferences for speech research, in 1989, 1999, and 2009.

Following this some of the challenges faced by dialogue researchers are examined, looking first at what sorts of research projects and questions are feasible and useful to investigate and then discussing the problem of how to access appropriate tools and resources to carry out this research. Some new initiatives are described, including a project that aims to make tools and resources more widely available to researchers, and a spoken dialogue challenge that enables researchers to participate in the development and evaluation of a large-scale, ongoing project. Next some opportunities and new areas for research are explored, in particular, the integration of dialogue technology into voice search systems and the role of dialogue systems in ambient intelligence environments. The final section provides a brief summary of the main conclusions.

2. Research in Spoken Dialogue Technology

2.1 The Nature of Dialogue Research

Dialogue has been studied in a wide range of disciplines including sociolinguistics, ethnography, communication science, media studies, social psychology, conversation analysis, and natural language processing. The domains of dialogue research have included global communication between nations, ethnic groups and religions; organisational communication within businesses and in contexts of employment; and various other areas such as cross-cultural, male-female, intergenerational, and professional communication, in contexts such as doctor-patient and teacher-student communication.

Spoken dialogue technology focuses specifically on a computational approach to dialogue, looking at how dialogue can be automated on a computer to support human-machine communication. Early work in this area can be traced

back to the 1970s when researchers first explored ways of interacting with machines using typed input and output (see (McTear, 1987) for a review of this work). It is only within the past three decades or so that researchers began to integrate speech into the dialogue interface, giving rise to a new branch of Human Language Technologies known as Spoken Dialogue Technology (McTear, 2004).

Two main directions can be identified in spoken dialogue research:

- 1 Modelling dialogue as a fundamental aspect of human behaviour.
- 2 Providing tools to enable humans to access data, services, and resources on computers.

Modelling dialogue includes the computational modelling of human conversational competence in order to gain a better understanding of how human dialogue works, usually based on models from linguistics and artificial intelligence. Much of dialogue research from the 1970s and 1980s was of this nature. A different approach aimed to simulate human dialogue but without actually attempting to model human conversational competence. ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975) are early examples of dialogue simulation, and the tradition has continued to the present day with an annual competition, known as the Loebner prize, in which conversational dialogue systems are judged according to the extent to which their performance is "human-like" (<http://www.loebner.net/Prizef/loebner-prize.html>).

2.2 Academic and Commercial Research

Research and development in spoken dialogue technology is carried out both in academic and commercial laboratories, although there are differences in emphasis. Generally speaking, academic researchers focus on topics such as:

- 1 Advancing the technologies involved in spoken dialogue systems — speech recognition, spoken language understanding, response generation, and text-to-speech synthesis.
- 2 The use of technologies from artificial intelligence (AI), and more recently the use of statistical methods, for various aspects of the dialogue management process.
- 3 Dialogue phenomena — aspects of speech and language that are peculiar to dialogue, such as disfluencies in spontaneous speech, the use of prosody and paralanguage, and the production and perception of emotion.

Commercial developers, on the other hand, are guided by business needs and performance factors. Here the focus is on producing usable systems that pro-

vide a return on investment. Thus there are two rather distinct communities interested in interactive speech systems, each with their own conferences and publications (Pieraccini and Huerta, 2005; Jokinen and McTear, 2010). Broadly speaking, academic researchers are motivated to make new contributions to knowledge and to engage in scientific exploration in order to "push back the frontiers of knowledge," whereas commercial developers are driven more by factors such as improving customer experience and generating revenue.

2.3 Three Decades of Research in Spoken Dialogue Systems

There has been active research in spoken dialogue systems since the late 1980s, supported by a number of large-scale research programs including the DARPA Communicator Project, Japan's Fifth Generation program, and the European Union's ESPRIT, Language Engineering, and HLT (Human Language Technology) programs. Dialogue research is published in various speech and language journals, such as *Speech Communication*, *Computer Speech and Language*, *IEEE Transactions on Audio, Speech, and Language Processing*, *Computational Linguistics*, and *Natural Language Engineering*; at conferences such as *Interspeech*, *ICASSP*, *ACL*, *SIGdial* (<http://www.sigdial.org/>), and *Text, Speech and Dialogue* (<http://www.tsdconference.org/tsd2010/>); and at workshops such as *SEM-DIAL* (<http://www.ilc.uva.nl/sem-dial/>), the *Young Researchers Roundtable on Spoken Dialogue Systems (YRRSDS)* (<http://www.yrrsds.org/>), and the *First International Workshop on Spoken Dialogue Systems* (<http://www.uni-ulm.de/en/in/iwds2009/workshop/introduction.html>), first held in December 2009 with plans for future workshops on an annual basis. A new journal, *Dialogue and Discourse*, was launched in 2009 which, in the words of the editors, is "dedicated exclusively to work that deals with language "beyond the single sentence", in discourse (i.e., text, monologue) and dialogue, from a theoretical as well as an experimental and technical perspective" (<http://www.dialogue-and-discourse.org/>). Taken together, these publications, conferences and workshops indicate a healthy interest in dialogue research and a variety of outlets where researchers can publish and present their work.

The remainder of this section reports on an analysis that was made of papers on the topic of spoken dialogue systems that were presented at the *Interspeech* conferences in 1989 (http://www.isca-speech.org/archive/eurospeech_1989/index.html), 1999 (http://www.isca-speech.org/archive/eurospeech_1999/index.html), and 2009 (http://www.isca-speech.org/archive/interspeech_2009/index.html). The main purpose of the analysis was to discover trends in the number of sessions in the conferences

that were devoted specifically to dialogue, the number of papers presented across all sessions that were concerned with dialogue, and the range of topics that were addressed. Table 1 presents the titles of sessions that were devoted to spoken dialogue in each of the conferences along with the number of papers that appeared in each of the sessions, distinguishing between those sessions that involved oral presentations (O) and those that were poster sessions (P).

Table 1. Spoken dialogue sessions at Interspeech conferences.

Eurospeech '89			Eurospeech '99		Interspeech '09		
Language processing: Human Factors, Psychology, and Human-Machine Dialogue (O)		5	Dialogue(O)	5	Spoken Systems (P)	Dialogue	15
			Dialogue 1 (P)	12	Special Session: Machine Learning for Adaptivity in Dialogue Systems (O)		6
			Dialogue 2 (P)	15	User Interactions in Spoken Dialog Systems (O)		6
			Prosody (O)	5			
			Spoken dialogue systems (O)	5			
Total		5		42			27

As can be seen from Table 1, there was no session devoted exclusively to dialogue in 1989, although one session included the term *human-machine dialogue* in its title. In this session 5 out of a total of 10 papers were concerned with dialogue. In 1999 the situation had changed dramatically, with 5 sessions devoted to dialogue, comprising a total of 42 papers. This number was reduced to 3 sessions and 27 papers in 2009. However, there was also a tutorial entitled *Statistical approaches to dialogue systems*, indicating the importance of this new approach to spoken dialogue systems research.

However, papers on dialogue also appeared in other sessions that were not specifically devoted to dialogue. A search was made for all papers in which words such as *dialogue*, *conversation* or *interaction* appeared in the title, revealing that papers on dialogue appeared in a wide range of sessions. Indeed,

as shown in Table 2, there were around 15 additional papers on dialogue that did not appear in sessions that were specifically devoted to the topic of dialogue.

A more detailed analysis of each of the papers identified involved assigning each paper to a specific category, as follows:

- *Spoken dialogue systems* — descriptions of complete systems, sometimes with an emphasis on extension to a new domain or language.
- *Dialogue management* — discussion of strategies for dialogue management, such as the use of forms to represent the current dialogue state and to control the dialogue strategy, or the use of statistical methods for dialogue control.
- *Dialogue modelling* — research involving the modelling of various aspects of dialogue, such as adaptivity, co-operation, user behaviours, user preferences, and turn-taking.
- *Dialogue phenomena* — analysis of a range of phenomena that occur in dialogue, such as paralanguage, gaze behaviours, prosody, and affect.
- *Error handling* — the detection and treatment of errors in human-machine dialogues.
- *Design methods and tools* — the design and development of dialogue systems, including data collection, Wizard of Oz methods, design strategies, and development methods and tools.
- *Evaluation* — methods for the evaluation of dialogue systems.
- *Speech and language processing* — analysis of various aspects of speech and language processing that occur particularly in dialogue data, including confidence measures, perplexity, language modelling, spoken language understanding, semantics, discourse, and pragmatics.
- *Language generation and TTS* — language generation and TTS specifically in the context of generating utterances in dialogue.

Table 3 presents the results of this analysis of dialogue topics in the papers, showing that dialogue research covers a range of topics, some of which are specifically concerned with the design, development and evaluation of spoken dialogue systems, while the focus of others, such as *Speech and Language Processing* and *Language Generation and TTS* is more on the nature of the input and output components of dialogue systems.

Looking at the topics over time, it can be seen that in 1999 a major focus was on descriptions of complete systems and on strategies for dialogue management, and also to a lesser extent on methods for dialogue system design and

Table 2. Number of papers with 'dialogue', 'conversation', or 'interaction' in title.

Eurospeech '89			Eurospeech '99			Interspeech '09		
Speech analysis: prosody (P)	1		Speech recognition: Confidence measures(O)	1		Speech and Audio Segmentation and Classification(P)	1	
Speech synthesis: voice source and prosody (P)	1		Speech analysis and tools (P)	1		ASR: Tonal Language, Cross-Lingual and Multilingual ASR (P)	1	
Speech production (P)	1		Speech Recognition: Language Modelling (P)	1		Special session: Active Listening and Synchrony (O)	3	
Perceptual Aspects of Prosody and Voice Quality (P)	1		Prosody: Prosodic Phrasing and Interruptions (O)	1		Human speech production 2 (P)	1	
Language Processing: Semantic Oriented Language Analysis (P)	6		Speech Communication Education (O)	2		Prosody Production 1 (O)	1	
Language Processing: Speech-Oriented Language Analysis (P)	2		Systems, Architectures (O)	1		ASR: Spoken language understanding (O)	1	
Applications: Designing an Application (P)	2		Speech Understanding: Miscellaneous Topics (P)	4		Topics in spoken language processing (P)	1	
			Speech and the Internet (O)	2		Systems for spoken language translation (O)	1	
			Assessment (O)	1		Applications in learning and other areas (P)	1	
						Assistive speech technology (P)	1	
						Resources, annotation and evaluation (P)	2	
Total	15			14			14	
Total (Tables and 2)	1	20		56			41	

Table 3. Numbers of different dialogue topics.

<i>Dialogue Topic</i>	<i>Eurospeech '89</i>	<i>Eurospeech '99</i>	<i>Interspeech '09</i>
Spoken dialogue systems	1	12	2
Dialogue management	1	12	8
Dialogue modelling	2	4	4
Dialogue phenomena	5	8	8
Error handling	1	1	0
Design	2	6	3
Evaluation	0	2	3
Speech and language processing	7	11	10
Language generation and TTS	1	0	3
Total	20	56	41

development. As might be expected at Interspeech conferences, there was continued interest in dialogue-related issues in speech and language processing, as well as in various dialogue phenomena, especially prosody.

A new development in dialogue research since the 1999 conference has been the increased use of statistical approaches and machine learning, particularly for dialogue management, involving methods such as reinforcement learning as well as example-based and corpus-based methods. Table 4 shows how statistical methods have come to dominate research in dialogue management during the past decade.

Table 4. Statistical methods in dialogue management.

	<i>Eurospeech '89</i>	<i>Eurospeech '99</i>	<i>Interspeech '09</i>
Hand-crafted	1	11	0
Statistical methods	0	1	8

More generally, this analysis has shown how research in spoken dialogue systems was beginning to emerge in 1989, but that the focus of most of the papers at that time was on speech and language issues that occur in dialogue data. By 1999 there was considerable emphasis on the development of dialogue systems, with descriptions of complete systems for tasks such as flight and train timetable information, and discussions of strategies for dialogue management such as the use of frames and mixed initiative. In 2009 there was still interest in speech and language issues as well as dialogue phenomena, but the main change was a shift to machine learning and statistical methods for dialogue management. A new trend was also a greater number of papers on applications

in learning and education, assistive technologies, dialogue with robots and on mobile devices, as well as the integration of speech with other modalities.

While these results are interesting, they should be viewed with some caution. In the first place, the analysis has examined trends in dialogue research by taking snapshots at three points in time. A more detailed year-by-year analysis would provide more fine-grained results. Furthermore, the analysis only looked at papers that were accepted for the Eurospeech-Interspeech conferences and consequently did not include work that was submitted to the other conferences, workshops and journals, such as those listed earlier.

2.4 Application Areas for Dialogue Research

This section provides a brief overview of current work and future trends in dialogue research as reported at the *YRRSDS09 (Young Researchers Roundtable on Spoken Dialogue Systems 2009)*, where young dialogue researchers described their current work, their views on the future of spoken dialogue systems, and the areas that they considered exciting and important for future research. [Table 5](#) presents a summary of these topics.

In terms of the current work of the young researchers, there are several new areas in addition to those that have been discussed earlier, including: *situated interaction*, *incremental dialogue processing*, *listening agents*, and *ontology-driven dialogue management*. New application domains include: *embodied agents*, *voice search*, *mobile applications*, and *troubleshooting*, while areas for future research directions include greater focus on aspects of the user, such as cognitive processes, more natural dialogues incorporating socio-cultural and other relevant information, and the integration of speech with information from sensors in ambient intelligence environments. More generally, we can note a move away from purely spoken language systems focussing on information search and transactions to multi-modal systems for smart homes applications, educational systems, and in-car interactive and entertainment services.

3. Challenges for Researchers in Spoken Dialogue Systems

Looking now at the challenges that face researchers in spoken dialogue systems, there are two main issues to consider:

- 1 How best to conduct research in spoken dialogue systems.
- 2 How to access appropriate resources to carry out this research.

Table 5. Young researchers topics.

<i>Research interests</i>	<i>Future topics</i>	<i>Research areas</i>
Dialogue phenomena: timing, prosody, emotion	Conversational assistants for elderly, tutoring, games	Natural dialogues: human communication processes
Situated interaction	Multi-modal, embodied and situated interactive systems	Combining symbolic and statistical approaches
Incremental dialogue processing	Mobile speech applications	Situation awareness
Modelling socio-cultural aspects of interaction	Automated dialogue building	Shared semantic ontologies
Usability and quality measurement	Voice search	Multi-modal sensor fusion
Stochastic approaches: dialogue management, user simulation, dialogue design	Troubleshooting applications	Incremental dialogue processing
Multi-modal interaction	Toolkits for dialogue design, resources	Miscommunication management using prosody
Dialogue with robots	Dialogue systems for drivers	Integrating speech with sensors
Listening agents		Using knowledge of the user to improve recognition
Dialogue management in pro-active situations		Cognitive load of dialogue systems for drivers
Ontology driven dialogue task management		Cognitive processes of multi-modal interaction
Dynamic integration of heterogeneous knowledge sources		
Adaptivity and learning from interactions		

3.1 Conducting Research in Spoken Dialogue Systems

At the time of Eurospeech '99 there were many papers describing end-to-end spoken dialogue systems, reflecting the fact that there were several international and national projects working in this area. Since then researchers have tended to focus on particular problem areas within spoken dialogue technology rather than designing and developing complete end-to-end systems. There are a number of reasons for this change in research focus:

- 1 It has been amply demonstrated that end-to-end systems can be designed and developed, and indeed there are now a number of tools that facilitate this process. These tools have often been developed to support a particular development methodology or theory of dialogue. For example, there are several toolkits that support the development of VoiceXML applications, while at the more theoretical end there are tools such as *Trindikit* that are used to implement dialogue systems based on *Information State Update Theory* (Larsson and Traum, 2000). For this reason there is little in the way of new research to be achieved in developing another end-to-end system using such toolkits and the only way in which a new research contribution can be realised is by developing a system that demonstrates a new theory or approach. A good example is the *Classic* project (Computational Learning in Adaptive Systems for Spoken Conversation), in which the aim is to develop a data-driven, statistically-based architecture for more robust and more adaptable spoken dialogue systems (<http://www.classic-project.org/>).
- 2 The development of a non-trivial end-to-end system is a major undertaking that is usually best accomplished by a team of researchers rather than by individuals such as graduate students. On the one hand, it would be difficult for one person (or even a small group of researchers) to have the expertise in all of the component technologies required to implement a complete end-to-end system. Furthermore, although various toolkits can make this work easier, it is not necessarily the case that all of the available technical components — which are usually open-source for reasons of costs and availability — are in fact state-of-the-art, so that a lot of research of this type suffers from less than optimal performance, usually in the area of speech recognition, but often too in other areas such as spoken language understanding or speech synthesis.

For these reasons, researchers tend to focus on more narrowly defined topics, such as the following:

- 1 Dialogue-related issues for speech recognition, spoken language understanding, language generation, or speech synthesis.

- 2 Comparison of different approaches to dialogue management, for example, rule-based versus statistical approaches.
- 3 Methods for the representation of contextual information.
- 4 The integration and usage of different modalities to complement and supplement speech.
- 5 Models of human conversational competence.
- 6 The study of incremental processing in dialogue, for example, the time-course of interpretation processes and how information from different strata of dialogue can be incrementally processed.

3.2 The Availability of Resources for the Design and Development of Spoken Dialogue Systems

Developing a spoken dialogue system requires a sound understanding of how the various contributing components function, such as the speech recognition and synthesis engines, and the ability to integrate these components along with the dialogue manager into an efficient and fully functioning system. The resources to perform this task need to be readily available.

In a recent paper (Boves et al., 2009) it was claimed that important lines of speech research are hampered by a lack of suitable resources and tools. There are agencies and catalogues of resources such as *ELDA* (Evaluations and Language resources Distribution Agency) (<http://www.elda.org/>), which is the operational body of *ELRA* (European Language Resources Association) (<http://www.elra.info/>) and *LDC* (Linguistic Data Consortium) (<http://www.ldc.upenn.edu/>). However, in many cases the available tools and resources are fragmented and they require specialised expertise to use them effectively. Moreover, there are often IPR and copyright issues as well as proprietary data formats that are incompatible with other tools and resources. Examples of widely used tools in speech research are *HTK* (<http://htk.eng.cam.ac.uk/>) for speech recognition and *Festival* (Black and Lenzo, 2000) for speech synthesis. However there is a steep learning curve for researchers wishing to integrate these tools into their applications. Similarly, there are toolkits such the *CSLU toolkit* (Cole, 1999), which provides an easy-to-use graphical interface for the development of simple spoken dialogue systems but which is difficult to adapt for the development of more complex systems or to test more advanced dialogue management strategies.

A wide range of resources is required by developers of spoken dialogue systems, including Wizard of Oz toolkits, user simulators, reinforcement learning toolkits, machine learning software, corpora of dialogue examples, and devel-

opment and testing platforms. For this reason in many cases researchers are forced to make the best of whatever tools are available to them and to adapt them as best as possible to their particular research project. As a result researchers are often not able to avail of the best, state-of-the-art components. For example: commercial voice search systems such as *Google App* make use of massive databases and resources such as large, specially designed language models to optimise the performance of the speech recognition component. Resources such as these are generally not available to academic research groups.

The *CLARIN* project, as described in (Boves et al., 2009), is an ambitious attempt to address these issues. CLARIN is a joint effort of over 150 institutions in Europe, with funding from the EC (European Commission) and 23 participating countries. The vision of CLARIN is to make accessible resources for researchers in spoken language processing and by 2015 to have in place a well-funded and sustainable infrastructure. The project is particularly targeted towards researchers in humanities and the social sciences and as such will aim to address obstacles such as incompatible data formats and requirements such as specialised expertise.

In a similar vein the Spoken Dialog Challenge 2010 offers researchers the opportunity to contribute to the development and evaluation of a large scale system developed at the Dialog Research Center at Carnegie Mellon University (CMU) (Black and Eskenazi, 2009). The Spoken Dialog Challenge makes use of the *Let's Go* system which is concerned with bus scheduling queries. The system was developed over a number of years at CMU and was first deployed in 2005 with open source components for ASR (automatic speech recognition), parsing, dialogue management, text generation, and TTS (text-to-speech synthesis) that were developed at CMU (Raux et al., 2006). Researchers can download and install the system and run it from their desktop or they can interact with the system using a Voice Over IP version. The benefit for researchers is that they can contribute to a large-scale, ongoing project without having to first obtain and link together all the required components.

In summary, dialogue research has become more focussed in the past few years. End-to-end systems are still being developed within larger groups such as, for example, the Dialog Research Center at CMU (<http://www.dialrc.org/>) and the Conversational Interaction and the Spoken Dialogue Research Group at University of Rochester (<http://www.cs.rochester.edu/research/cisd/>). Given the large teams at centres such as these, there is scope to develop toolkits and software to support the design, implementation and testing of spoken dialogue systems, to encourage links with other related research areas such as human-robot interaction, and also to facilitate and integrate research on specific, more narrowly defined topics such as those listed earlier. Given that many of the toolkits and resources are openly available for academic purposes, researchers in smaller groups or those working as individuals can avail of opportunities to

make use of resources and also to become involved in larger projects such as the CMU Spoken Dialog Challenge.

4. Opportunities for Future Research in Dialogue

In addition to the research topics mentioned in the preceding section, there are two particular areas that provide exciting opportunities for future research in dialogue:

- 1 Incorporating dialogue technology into voice search.
- 2 Using dialogue systems in ambient intelligence environments.

4.1 Incorporating Dialogue into Voice Search

Voice search is a combination of speech recognition and document search, enabling users to access information in a database or on the World Wide Web (WWW) using a spoken query. In its most basic form the user speaks a query which is processed by an automatic speech recogniser and the result is passed to a search engine to retrieve a ranked list of documents. For example, in a simple experiment the query

little wing pizzeria in belfast

was spoken on an iPhone using the Google App. The speech recognition engine returned the following n-best output:

little wing pizzeria in belfast;
little wayne pizzeria in belfast;
little twin pizzeria in belfast;
digital weighting pizzeria in belfast;
little twin peaks arena in Belfast.

In this case, as in several similar attempts, the 1st-best result was the correct one. Once the user accepted one of the results, the application displayed a list of links that could be clicked to access the web pages.

This is an example of a one-shot search query. In other cases, as described by (Wang et al., 2008), a dialogue manager might employ a disambiguation strategy to determine more accurately the user's intent and to narrow down multiple entries. For example, for residential directory assistance entries for people with the same name may be disambiguated using addresses, while in business directory assistance business categories are often used for disambiguation. A more flexible solution developed in a research prototype system used summarization to narrow down a large ambiguous set of entries (Polifroni and Walker, 2006). For example, in an application for searching for

a restaurant, common attributes such as price range and food type were used to guide users to the most appropriate choice of restaurant, using attributes that were selected automatically from a user model and associated rules in the database subset in the dialogue focus. The following example from (Polifroni and Walker, 2006) illustrates:

User: Tell me about restaurants in London.

System: I know of 596 restaurants in London. All price ranges are represented. Some of the cuisine options are Italian, British, European, and French.

User: I'm interested in Chinese food.

System: I know of 27 restaurants in London that serve Chinese cuisine. All price ranges are represented. Some are near the Leicester Square tube station.

User: How about a cheap one?

System: I know of 14 inexpensive restaurants that serve Chinese cuisine. Some are near the Leicester Square tube station. Some are in Soho.

The requirements for speech recognition, spoken language understanding, and dialogue management are different in voice search compared with traditional dialogue systems (Kawahara, 2009). In a dialogue system the user's spoken language input is typically interpreted as values of concepts elicited in a sequence of pre-determined dialogue states that fill a series of slots — such as destination and departure time of a flight. These slot values are inserted into a database query to obtain a result that is spoken back to the user. In voice search, however, the user can express the query in many different ways (high input space) and there is a large number of targets (large semantic space), often comprising millions of entries, as in a directory assistance system. Thus there are many speech recognition and spoken language understanding challenges, which are currently the main focus of attention in voice search research (Acero et al., 2008; Feng et al., 2009; Zweig, 2009). The performance of current systems has improved dramatically in the past few years, due in part to very large language models that are used to constrain and enhance recognition (Van Heerden et al., 2009). There is also some research into methods for correcting misrecognitions (Vertanen and Kristensson, 2009).

Voice search is currently motivated by what is technically feasible in terms of speech recognition accuracy, search engine technology, and network connectivity. What might be attractive and useful for users would be the ability to conduct a voice search and then continue to engage the system in a spoken dialogue on the results of the search. Taking the query *little wing pizzeria in Belfast* as an example, the following is a possible dialogue:

User: Little wing pizzeria in Belfast.

System: (retrieves some summary information, e.g. address, restaurant

type).

User: When is it open?

System: (retrieves opening times).

User: What is on the menu?

System: (retrieves summary of menu).

User: What is the price range?

System: (retrieves price range).

To be able to perform in this way the system would need, among other things, to be able to:

- construct dynamic language models to focus on the different attributes being queried;
- model the discourse context, in order to be able to handle discourse phenomena such as ellipsis and reference;
- handle information at summary level and then initiate a dialogue to offer more detailed information, if requested.

Each of these requirements presents major challenges for researchers and would involve the integration of other technologies, such as interactive question-answering (Q-A), with voice search. Notwithstanding these problems, systems that could act as virtual agents and engage in spoken dialogue with users of small mobile devices to provide information, services and entertainment, would offer immeasurable advantages compared with the functionalities available on current smart phones. The networking and web mining capabilities are already well advanced. What is now required are approaches that can build on this infrastructure to create intuitive and usable spoken dialogue applications (Gilbert and Feng, 2008).

4.2 Using Dialogue Systems in Ambient Intelligence Environments

Ambient intelligence has been defined as "a digital environment that proactively, but sensibly, supports people in their daily lives" (Augusto and McCullagh, 2007). Current voice search applications are designed for hand-held devices such as smart mobile phones, although voice search technology could (and should) be combined with other technologies to provide user services in ambient intelligence environments. According to (Aghajan et al., 2010):

Ambient intelligence (AmI) is a fast-growing multi-disciplinary field that is ushering in new opportunities for many areas of research to have a significant impact on society. Its foundation is the enrichment of the environment, through sensing and processing technologies, to understand, analyze, anticipate, and adapt to events and to users and their activities, preferences, intentions, and behaviours.

Basically, Aml gathers real-time information from the environment and combines it with historical data accumulated over time, or a knowledge base, to provide user services.

One of the main issues for such technologies is how to interface with the user. Spoken dialogue, which is the most natural mode of communication between humans, is now being applied along with other relevant technologies to provide intuitive and user-friendly interfaces in ambient intelligence environments. Aghajan et al. (2010a) is a recent collection of papers concerned with a range of topics on human-centric interfaces for ambient intelligence. McTear (2010) and Minker et al. (2010) focus specifically on spoken dialogue systems.

A number of recent research projects have also investigated the use of spoken dialogue systems in ambient intelligence environments. A selection of these projects will be reviewed in the following subsection, beginning with the CHAT project, which is concerned with spoken dialogue systems in cars across a number of different domains. This is followed by the SmartKom and SmartWeb projects. SmartKom involves spoken and multi-modal communication in three different environments — a public kiosk information system, a home-based system for device control and information services, and a mobile environment for car drivers and pedestrians. SmartWeb addresses the issue of open-domain information retrieval using spoken language question-answering, while the TALK project was concerned with multi-modal dialogue interaction involving various devices and services in the home and car. The Companions project addresses the issues involved in assisting senior citizens to perform everyday tasks and to access information through the use of an embodied conversational agent. Finally, the Atraco project is concerned with the use of spoken dialogue to control appliances, devices, and services within an ambient intelligence environment. Web pages listing many other spoken dialogue systems can be found at (<http://www.cs.cmu.edu/~dbohus/SDS/>; http://www.ling.gu.se/~sl/dialogue_links.html; <http://www.infj.ulst.ac.uk/~cbdg23/book/projects.htm>).

4.3 CHAT

CHAT (Weng et al., 2007) is a spoken dialogue system for cars that operate in the domains of entertainment (MP3 player), navigation, and web services (restaurant finder). CHAT allows users to switch easily between these domains. The following is an example:

User: Switch to restaurant.

System: OK, switched to the restaurant finder.

User: Can you find me a Chinese restaurant in Palo Alto with good service?

System: There are 11 Chinese restaurants in Palo Alto with good service.

User: Switch to navigation.

System: OK, switched to navigation.

User: Take me to Little Garden Restaurant, but stop by Bosch Page Mill Office.

User: Switch to music player.

System: OK, switched to MP3 music player.

User: What songs do you have by Alicia Keys?

In these examples the user makes use of explicit statements to switch domains. Implicit domain switching is also possible — for example, having selected a restaurant in the restaurant domain and then by saying *find me a fast route to restaurant XYZ* without explicitly saying *switch to navigation*. Keeping track of multiple threads related to different devices and domains is achieved through the use of a dynamically evolving dialogue move tree, which classifies and interprets user utterances in terms of the current dialogue context and selects from a set of possible system responses (Lemon and Gruenstein, 2004).

An important consideration for in-car spoken dialogue systems is that the interaction should not divert drivers from their primary task of driving. The Content Optimizer controls the amount of information to be output to the driver, especially in the case of lengthy lists of results to be returned, by proposing extra criteria to narrow down the results. Additionally, when there is no result returned from the database, a relaxation of the constraints given by the user is employed. These measures have been shown to lead to increased user satisfaction (Pon-Barry et al., 2006). Other measures included an overall task completion rate of 98% for navigation tasks and a high level of dialogue efficiency, with an average of 2.3 turns being required to complete the tasks.

4.4 SmartKom and SmartWeb

SmartKom (Wahlster, 2006; <http://www.smartkom.org/>) is a multi-modal dialogue system that supports face-to-face interaction with an embodied conversational agent (Smartakus) in three different environments:

- *Public* — a kiosk-based help system for tasks such as making telephone calls, sending faxes and emails, retrieving information about movies, making ticket and seat reservations at the cinema, and biometrical authentication (Horndasch et al., 2006).
- *Home* — an infotainment companion for the operation of various TV appliances, such as programming the VCR to record a programme, as well as providing an intelligent interface to the electronic programme guide (EPG) for browsing and creating personalized programme listings (Portele et al., 2006).

- *Mobile* — a mobile travel companion to be used in the car as well as while walking to assist with navigation and point-of-interest information retrieval (Berton et al., 2006; Malaka et al., 2006).

The tasks involving the SmartKom system require a collaborative dialogue between the user and the system to define and elaborate the task as well as to monitor its execution. Given the range of modalities available for input and output — speech, gesture, facial expression — the technical challenges included seamless fusion and mutual disambiguation of the input as well as plan-based methods for multi-modal fission and adaptive output presentation. An important consideration was to integrate the various devices and environments in such a way as to support ubiquitous access to information using any device while providing a seamless interface for the user within a single interaction. This required the ability to manage physical connections when users changed devices and to represent how different modalities were supported on the various devices. For example, the user could transition from a pedestrian scenario to a driving scenario in a car which required a change of device from PDA to in-car devices and a change of permissible modalities from multi-modal input and presentation to speech-based input and output.

While SmartKom works in multiple domains and on various devices, the dialogues that it supports are domain-specific — for example, obtaining cinema information, getting details of TV programmes, or route planning. In a follow-up project — SmartWeb — open-domain question answering is addressed, using the Web as a knowledge base and exploiting the machine-understandable content of semantic web pages for intelligent question-answering (Reithinger et al., 2005; <http://www.smartweb-projekt.de>). Information is extracted from huge volumes of information on the Web using named-entity extraction to obtain information about persons, objects and locations named in the query. The question analysis process is ontology-driven resulting in a paraphrase of the question and a semantic representation in an XML format. The user interface is context-aware in that it supports the user in different roles — as a car driver, motorcyclist, pedestrian, and sports spectator. As with SmartKom, a major emphasis in the SmartWeb project is to develop an infrastructure that can handle multi-modal, open-domain question-answering and requests to services within a Semantic Web framework using standard interfaces and protocols.

4.5 TALK

TALK (Talk and Look: Tools for Ambient Linguistic Knowledge) was concerned with multi-modal dialogue interaction involving various devices and services in the home and car (<http://www.talk-project.org/>). The main aim of the project was to make dialogue systems more conversational, robust and adaptive. The theoretical basis for dialogue management in TALK is the Infor-

mation State Update (ISU) approach (Larsson and Traum, 2000) which provides a rich formalism for the representation of the dialogue context that is essential for the multi-tasking and event-driven dialogues typical of in-car interaction. One aim of the project was to develop the ISU approach to make it reusable for different languages, modalities, and application domains. A second aim was to develop techniques for the automated learning of optimal dialogue strategies from corpora of human interactions that were modelled using the ISU approach (Lemon et al., 2006). Demonstrator systems developed within the project included the SAMMIE in-car dialogue system which was installed in the BMW test car (Becker et al., 2007) and the MIMUS home automation system that allows users to control devices in the home using voice and mouse clicks (Pérez et al., 2006; Amores et al., 2007). Several research issues were addressed in the project, including the role of ontologies and the development of methods to support the automatic learning of dialogue strategies from examples of human interactions (Rieser and Lemon, 2006; Williams and Young, 2007). Other contributions included the collection of corpora that could be used for the automatic training of dialogue systems (Kruijff-Korbayova et al., 2006; Manchón et al., 2006), as well as the development of user simulators to support large-scale empirical studies of user-system interactions without the need for large groups of human users (Georgila et al., 2006; Schatzmann et al., 2006).

4.6 COMPANIONS

The aim of the COMPANIONS project is to assist senior citizens in carrying out everyday tasks and to provide easy access to information using a variety of devices ranging from PCs to handheld devices, and supporting interaction with small robots and embodied conversational agents (Catizone et al., 2008; <http://www.companions-project.org/>). The main technologies involved are speech and natural language processing, dialogue management, emotion processing, and the social use of language.

The COMPANIONS project has two demonstrators: a Health and Fitness Companion and a Senior Companion. The purpose of the Health and Fitness Companion is to monitor information about a user's eating habits and fitness activities to support a healthy lifestyle. The Senior Companion is designed as a conversational companion for the elderly, providing access to information and services, helping them to carry out everyday tasks, and entertaining and chatting with them about past experiences and memories evoked by photographs.

Dialogue management in COMPANIONS requires a different approach from the form-filling method used widely in traditional task-based dialogue systems. In COMPANIONS, where the goal of the interaction is to provide personal assistance and companionship, the interactions are less clearly de-

fined. Moreover, whereas in traditional systems each interaction is normally modeled as a single event, COMPANIONS aims to model the development of a longer term relationship between the user and the conversational agent. There will also be a range of different interfaces available, such as avatars, talking heads, and small robots, compared to the use of the telephone in traditional systems. Finally information gathered by the system will include not only speech but also images — in the form of photos —, as well as information based on GPS and sensor data about the user's location and physical state while taking exercise.

4.7 Atraco

The Atraco project is concerned with the use of a range of interactive appliances, collaborative devices, context-aware artefacts, models, and services, and software components to assist users in a variety of tasks. Adaptive dialogue models are used to communicate the system's state and to interact with users.

The spoken dialogue aspects of the system are described in (Minker et al., 2010; <http://www.uni-ulm.de/in/atrac/>). A number of different spoken dialogue systems technologies are being investigated, including: robust spoken language processing, adaptive and proactive dialogue modelling, multi-modality, intelligent planning, monitoring and plan adaptation, embedded agents, reasoning, inferencing, flexible end-device technology, and networking. Demonstrator systems include:

- a Pedestrian navigation system;
- a Journey planning system;
- a Restaurant advisor.

One interesting feature of the Restaurant advisor is that the dialogue agent monitors an ongoing dialogue between two humans and, when required, takes the initiative and becomes involved in the dialogue. For example, if the discussion is about restaurants, the agent joins the dialogue and suggests some suitable restaurants. In order to support these proactive capabilities, the agent needs to be able to maintain a detailed model of the dialogue context as well as a dialogue history.

4.8 Summary

The research described in the preceding sections involves systems that can interact intelligently and co-operatively across different environments using a range of appropriate modalities to support people in the activities of daily living. Such systems will be useful as aids to the elderly, for people with

disabilities, as educational aids for children and second language learners, and for many other applications.

There are a number of important research issues that need to be addressed, including:

- *Context awareness*: how to integrate information from the environment — internet, sensors, spoken interaction, etc.
- *Contexts for speech and the use of different modalities*: how to decide in which contexts speech is useful and appropriate and whether other modalities are appropriate — for example, when outputting an address or the information in a menu.
- *User modelling*: how to model individual user preferences and needs, distinguishing, for example, between older and younger users, people with disabilities, etc.
- *Technical*: issues such as the optimal usage of microphones, networks, mobile environments, speech recognition accuracy, appropriate dialogue management, acceptable speech output.

5. Concluding Remarks

This chapter has examined trends and developments in spoken dialogue systems research over the past three decades, based primarily on snapshots from an analysis of papers presented at the Eurospeech-Interspeech conferences in 1989, 1999, and 2009. This analysis has shown how dialogue research has matured from its early beginnings in which researchers initially explored spoken language issues associated with dialogue, then moved on to building a range of end-to-end systems, experimenting with different methods of dialogue management and control, up to the present where there is a wide range of different research endeavours, including a new focus on statistical methods and machine learning for dialogue management.

Researchers in dialogue are faced with many challenges, in particular, with accessing resources and tools that can be used to design, implement and test systems. A major barrier to progress is the issue of speech recognition accuracy, especially now that dialogue systems are being developed for deployment in more open-ended and potentially noisier environments, such as cars and smart homes. Recent progress has been made in achieving more acceptable levels of speech recognition accuracy in commercial voice search systems, but the speech recognition engines and language models used in these systems are generally not available for use by academic researchers. Another problem is that it is becoming increasingly more difficult for individuals or researchers in small groups to build end-to-end dialogue systems, since this involves the integration of a wide range of different components. Even where these components

are available, the expertise to develop, modify and integrate them would generally be beyond the capability of an individual researcher. However, projects such as CLARIN, which aims to make tools and resources more readily available for researchers, and the CMU Spoken Dialog Challenge, which provides an opportunity for researchers to participate in a large scale project, are examples of how these challenges and problems are being addressed.

Although considerable progress has been made in dialogue research over the past thirty years, there are still many opportunities for continued fundamental as well as applied research. Two examples that were discussed are voice search using dialogue technology and dialogue systems in ambient intelligence environments. There are many opportunities for research in areas such as these, both as testbeds for theories and technologies as well as in terms of making useful contributions to the development and deployment of applications that provide social and economic benefits.

Web Pages

A List of Spoken Language Interfaces.

<http://www.cs.cmu.edu/~dbohus/SDS/>

Atraco Project Home Page.

<http://www.uni-ulm.de/in/atracoc/>

Classic Project Home Page.

<http://www.classic-project.org/>

Companions Project Home Page.

<http://www.companions-project.org/>

Conversational Interaction and Spoken Dialogue Research Group, Department of Computer Science, University of Rochester.

<http://www.cs.rochester.edu/research/cisd/>

Dialog Research Center at CMU (DialRC).

<http://www.dialrc.org/>

Dialogue and Discourse: An International Journal.

<http://www.dialogue-and-discourse.org/>

Dialogue Systems and Projects.

http://www.ling.gu.se/~sl/dialogue_links.html

ELDA: Evaluations and Language Resources Distribution Agency.

<http://www.elda.org/>

ELRA: European Language Resources Agency.

<http://www.elra.info/>

First European Conference on Speech Communication and Technology (EUROSPEECH '89).

http://www.isca-speech.org/archive/eurospeech_1989/index.html

International Conference Series on Text, Speech, and Dialogue. Brno, Czech Republic.

<http://www.tsdconference.org/tsd2010/>

International Workshop Series on Spoken Dialogue Systems Technology.

<http://www.uni-ulm.de/en/in/iwsds2009/workshop/introduction.html>

Interspeech 2009, Brighton, UK, 6-10 September, 2009.

http://www.isca-speech.org/archive/interspeech_2009/index.html

LDC: Linguistic Data Consortium.

<http://www.ldc.upenn.edu/>

SEMDIAL: Workshop Series on the Semantics and Pragmatics of Dialogue.

<http://www.illc.uva.nl/semdial/>

SIGdial: Special Interest Group on Discourse and Dialogue.

<http://www.sigdial.org/>

Sixth European Conference on Speech Communication and Technology (EU-ROSPEECH'99).

http://www.isca-speech.org/archive/eurospeech_1999/index.html

SmartKom Project Home Page.

<http://www.smartkom.org/>

SmartWeb Project Home Page.

<http://www.smartweb-projekt.de>

Spoken Dialogue Technology: Projects and Links.

<http://www.infj.ulst.ac.uk/~cbdg23/book/projects.htm>

TALK Project Home Page.

<http://www.talk-project.org/>

The HTK (Hidden Markov Model) Toolkit.

<http://htk.eng.cam.ac.uk/>

The Loebner Prize in Artificial Intelligence.

<http://www.loebner.net/Prizef/loebner-prize.html>

Young Researchers' Roundtable on Spoken Dialog Systems.

<http://www.yrrsds.org/>

Notes

1. Originally there were two biennial conferences — *Eurospeech* (*European Conference on Speech Communication and Technology*) and *ICSLP* (*International Conference on Spoken Language Processing*). Since 2000 both conferences have been held under the common label *Interspeech*. The 1989 and 1999 conferences are known as *Eurospeech '89* and *Eurospeech '99*.

References

Acero, A., Bernstein, N., Chambers, R., Ju, Y. C., Li, X., Odell, J., Nguyen, P., Scholz, O., and Zweig, G. (2008). Live Search for Mobile: Web Services by Voice on the Cellphone. In *Proceedings of International Conference on*

- Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 5256–5259, Las Vegas, NV.
- Aghajan, H., López-Cózar Delgado, R., and Augusto, J. C. (2010). Preface. In Aghajan, H., López-Cózar Delgado, R., and Augusto, J. C., editors, *Human-Centric Interfaces for Ambient Intelligence*, pages xix–xxvii. Elsevier, New York.
- Amores, G., Pérez, G., and Manchón, P. (2007). MIMUS: A Multimodal and Multilingual Dialogue System for the Home Domain. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 1–4, Prague.
- Augusto, J. and McCullagh, P. (2007). Ambient Intelligence: Concepts and Applications. *International Journal of Computer Science and Information Systems*, 4(1):1–28.
- Becker, T., Blaylock, N., Gerstenberger, C., Korthauer, A., Perera, N., Pitz, M., Poller, P., Schehl, J., Steffens, F., Stegmann, R., and Steigner, J. (2007). D5.3: In-Car Showcase Based on TALK Libraries. IST-507802 Deliverable 5.3. Technical report.
- Berton, A., Bühler, D., and Minker, W. (2006). SmartKom-Mobile Car: User Interaction with Mobile Services in a Car Environment. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 523–537. Springer, Berlin.
- Black, A. and Eskenazi, M. (2009). The Spoken Dialogue Challenge. In *Proceedings of the SIGDIAL 2009 Conference*, pages 337–340, London, UK. Association for Computational Linguistics.
- Black, A. and Lenzo, K. (2000). Building Voices in the Festival Speech System. <http://festvoc.org/bsv/>.
- Boves, L., Carlson, R., Hinrichs, E., House, D., Krauwer, S., Lemnitzer, L., Vainio, M., and Wittenburg, P. (2009). Resources for Speech Research: Present and Future Infrastructural Needs. In *Proceedings of Interspeech 2009*, pages 1803–1806, Brighton, UK.
- Colby, K. M. (1975). *Artificial Paranoia — A Computer Simulation of Paranoid Processes*. Pergamon Press, New York.
- Cole, R. A. (1999). Tools for Research and Education in Speech Science. In *Proceedings of International Conference of Phonetic Sciences*, San Francisco, CA.
- Feng, J., Banglore, S., and Gilbert, M. (2009). Role of Natural Language Understanding in Voice Search. In *Proceedings of Interspeech 2009*, pages 1859–1862, Brighton, UK.
- Georgila, K., Henderson, J., and Lemon, O. (2006). User Simulation for Spoken Dialogue Systems: Learning and Evaluation. In *Proceedings of Interspeech 2006*, pages 1–4, Pittsburgh, PA.
- Gilbert, M. and Feng, J. (2008). Speech and Language Processing over the Web. *IEEE Signal Processing Magazine*, 25(3):18–28.

- Horndasch, A., Rapp, H., and H. Röttger, H. (2006). SmartKom-Public. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 471–492. Springer, Berlin.
- Jokinen, K. and McTear, M. (2010). *Spoken Dialogue Systems. Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Kawahara, T. (2009). New Perspectives on Spoken Language Understanding: Does Machine Need to Fully Understand Speech? In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2009)*, pages 46–50, Merano, Italy.
- Kruijff-Korabayova, I., Becker, T., Blaylock, N., Gerstenberger, C., Kaiser, M., Poller, P., Rieser, V., and Schehl, J. (2006). The SAMMIE Corpus of Multimodal Dialogues with an MP3 Player. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.
- Larsson, S. and Traum, D. (2000). Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, 6:323–340.
- Lemon, O., Georgila, K., and Stuttle, M. (2006). An ISU Dialogue System Exhibiting Reinforcement Learning of Dialogue Policies: Generic Slot-Filling in the TALK In-Car System. In *EACL (demo session)*, Trento.
- Lemon, O. and Gruenstein, A. (2004). Multithreaded Context for Robust Conversational Interfaces: Context-Sensitive Speech Recognition and Interpretation of Corrective Fragments. *ACM Transactions on Computer-Human Interaction (ACM TOCHI)*, 11(3):241–267.
- Malaka, R., Häußler, J., Aras, H., Merdes, M., Pfisterer, D., Jöst, M., and Porzel, R. (2006). SmartKom-Mobile: Intelligent Interaction with a Mobile Interface. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 505–522. Springer, Berlin.
- Manchón, P., del Solar, C., Amores, G., and Pérez, G. (2006). The MIMUS Corpus. In *Proceedings of LREC 2006: International Workshop on Multimodal Corpora From Multimodal Behaviour Theories to Usable Models*, pages 56–59, Genoa, Italy.
- McTear, M. (1987). *The Articulate Computer*. Blackwell, Oxford.
- McTear, M. (2004). *Spoken Dialogue Technology: toward the Conversational User Interface*. Springer, London.
- Pérez, G., Amores, G., and Manchón, P. (2006). A Multimodal Architecture for Home Control for Disable Users. In *Proceedings of the IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT)*, Aruba.
- Pieraccini, R. and Huerta, J. (2005). Where do we go from here? Research and Commercial Spoken Dialogue Systems. In *Proceedings of 6th SIGdial Workshop on Dialogue and Discourse*, pages 1–10, Lisbon, Portugal.
- Polifroni, J. and Walker, M. (2006). An Analysis of Automatic Content Selection Algorithms for Spoken Dialogue System Summaries. In *Proceedings*

of the IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT), Aruba.

- Pon-Barry, H., Weng, F., and Varges, S. (2006). Evaluation of Content Presentation Strategies for an In-Car Spoken Dialogue System. In *Proceedings of Interspeech 2006*, pages 1930–1933, Pittsburgh, PA.
- Portele, T., Goronzy, S., Emele, M., Kellner, A., Torge, S., and te Vrugt, J. (2006). SmartKom-Home: The Interface to Home Entertainment. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 493–503. Springer, Berlin.
- Raux, A., Bohus, D., Langner, B., Black, A. W., and Eskenazi, M. (2006). Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. In *Proceedings of Interspeech 2006*, Pittsburgh, PA.
- Rieser, V. and Lemon, O. (2006). Using Machine Learning to Explore Human Multimodal Clarification Strategies. In *Proceedings of the IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT)*, Aruba.
- Schatzmann, J., Weillhammer, K., Stuttle, M. N., and Young, S. (2006). A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowledge Engineering Review*, 21(2):97–126.
- Van Heerden, C., Schalkwyk, J., and Strobe, B. (2009). Language Modeling for What-with-Where on GOOG-411. In *Proceedings of Interspeech 2009*, pages 991–994, Brighton, UK.
- Vertanen, K. and Kristensson, P. O. (2009). Recognition and Correction of Voice Web Search Queries. In *Proceedings of Interspeech 2009*, pages 1863–1866, Brighton, UK.
- Wang, Y. Y., Yu, D., Ju, Y. C., and Acero, A. (2008). An Introduction to Voice Search. *IEEE Signal Processing Magazine*, 25(3):29–38.
- Weizenbaum, J. (1966). ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the Association for Computing Machinery*, 9:36–45.
- Weng, F., Yan, B., Feng, Z., Ratiu, F., Raya, M., Lathrop, B., Lien, A., Varges, S., Mishra, R., Lin, F., Purver, M., Bratt, H., Meng, Y., Peters, S., Scheideck, T., Raghunathan, B., and Zhang, Z. (2007). CHAT to your Destination. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 79–86, Antwerp, Belgium.
- Williams, J. and Young, S. (2007). Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language*, 21(2):393–422.
- Zweig, G. (2009). New Methods for the Analysis of Repeated Utterances. In *Proceedings of Interspeech 2009*, pages 2791–2794, Brighton, UK.

Chapter 7

DIALOGUE CONTROL BY POMDP USING DIALOGUE DATA STATISTICS

Yasuhiro Minami, Akira Mori, Toyomi Meguro
NTT Communication Science Laboratories, NTT Corporation
Kyoto, Japan
{minami,akira,meguro}@cslab.kecl.ntt.co.jp

Ryuichiro Higashinaka
NTT Cyber Space Laboratories, NTT Corporation
Yokosuka, Japan
higashinaka.ryuichiro@lab.ntt.co.jp

Kohji Dohsaka, Eisaku Maeda
NTT Communication Science Laboratories, NTT Corporation
Kyoto, Japan
{dohsaka,maeda}@cslab.kecl.ntt.co.jp

Abstract Partially Observable Markov Decision Processes (POMDPs) are applied in action control to manage and support users' natural dialogue communication with conversational agents. Any agent's action must be determined, based on probabilistic methods, from noisy data through sensors in the real world. Agents must flexibly choose their actions to reach a target dialogue sequence with the users while retaining as many statistical characteristics of the data as possible. This issue can be solved by two approaches: automatically acquiring POMDP probabilities using Dynamic Bayesian Networks (DBNs)(DBNs) trained from a large amount of dialogue data and obtaining POMDP rewards from human evaluations and agent action predictive probabilities. Using the probabilities and the rewards, POMDP value iteration calculates a policy that can generate an action sequence that maximizes both the predictive distributions of actions and user evaluations. This chapter focuses on how to make the rewards from predictive

distributions. Introducing rewards lets the policy generate actions whose predictive probabilities are maximized. Experimental results demonstrate that the proposed method can generate actions with high predictive probabilities while generating target action sequences.

Keywords:

Partially Observable Markov Decision Process (POMDP); Dialogue management; Multi-modal interaction; Dynamic Bayesian Network (DBN); Expectation-Maximization (EM) algorithm.

1. Introduction

Our research goal is to automatically acquire a conversation agent's action control strategy for dialogue (we use the term "dialogue" here in its general meaning as an interaction of two entities). Here, we assume that the structure of a dialogue is unknown. Under this situation, the systems must create and establish behavioral strategies based on a large amount of data in their communications. Markov Decision Processes (MDPs) are ordinarily applied to the acquisition of strategies with reinforcement learning (RL) (Sutton and Barto, 1998; Russell and Norvig, 2003). MDPs have also been applied to dialogue control (Levin and Pieraccini, 1997; Levin et al., 1998; Goddeau and Pineau, 2000; Denecke et al., 2004a; Denecke et al., 2004b). However, since MDPs suppose that the states are known, they do not work under unknown-state conditions. In addition, the data generally contain errors and uncertainties that originated from observation problems. Especially for conversations between a user and an agent in the real world, we have to consider diverse errors from recognition of speech, facial expressions, behaviors, and timings of user actions. Under these noisy conditions, learning and behavioral acquisition under MDPs do not always work effectively. In this case, such information includes behaviors, speech, and paralinguistic information that are generated from unobserved internal states. A partially observable Markov decision process (POMDP) (Smallwood and Sondik, 1973; Sutton and Barto, 1998; Russell and Norvig, 2003) can help to formally handle these unobserved internal states.

POMDPs play an effective role in making decisions about selecting the most statistically reliable actions available by observing sensor data having uncertainty. POMDPs have been applied to spoken dialogue management (Roy et al., 2000). RL, which obtains POMDP's action strategies, requires actual interaction between the user and the agent. However, since increasing the hidden states of a POMDP requires a large amount of actual interactions with real users to obtain the dialogue control strategies, RL cannot be applied to this condition. It is necessary to reduce the time consumed in calculating a policy. If the environmental information is known, the optimal policy, which is the action

strategy for selecting the best action by looking at the environmental information, can be obtained by value iteration without user interactions (Smallwood and Sondik, 1973). Recently, Point Based Value Iteration (PBVI), a fast calculation method for value iteration by maintaining only certain state distribution points, was proposed (Pineau et al., 2003). Moreover, introducing Algebraic Decision Diagrams to PBVI improved it (Poupart, 2005). Other fast calculation methods for value iteration have also been proposed (Smith and Simmons, 2004; Smith and Simmons, 2005).

As algorithms for POMDP have been proposed, dialogue methods using these fast value iteration methods have also been proposed (Williams et al., 2005b; Williams et al., 2005a; Williams, 2007; Young et al., 2010). Dialogue support exists for buying train tickets (Williams et al., 2005b; Williams et al., 2005a), weather information dialogues (Kim et al., 2008), dialogues for digital subscriber line troubleshooting (Williams, 2007), and the action control of robots by human speech and gestures (Schmidt-Rohr et al., 2008). The results from these cases demonstrate that POMDPs compensate for the uncertainty of such observed data as speech and gestures in action-determination. As a result, they achieve better performance in terms of the correct accomplishment of given tasks than a conventional Markov Decision Process (MDP). Since these systems are based on task-oriented dialogue management and we know how the agent should work, setting rewards and calculating transition probabilities are easy operations. However, if we do not know how the agent should work, such as in person-to-person communication, we have to estimate this by using a large amount of data. The problem is how to create the POMDP structure from a large amount of data. Although Fujita solved this problem by using dynamic Bayesian networks (DBNs) to model a POMDP structure with a great deal of data, their task was simple and task-oriented (Fujita, 2007).

We extend this method to handle unknown tasks. We suppose that system action controls can be obtained by selecting dialogues that fit our requirements. Then we select the dialogues and use them to train the system so that it is likely to generate the selected dialogues. We call such selected dialogues target dialogues. We conducted some preliminary experiments using DBNs to train POMDP policy by value iteration. However, the system only generated target dialogues. This means that the system ignores the statistics of the rest of the data, while the DBN learns the statistics of all of the training data. We think that for natural dialogues the rest of the data statistics are also important. For example, counseling dialogues contain a variety of conversations other than conversations made for counseling purposes. To construct a good counseling system by training a DBN, we consider using a questionnaire, for example, that asks whether a counseling dialogue is good and then selecting the good counseling dialogues by hand. If only a certain part were selected and used for training a POMDP policy, the system would generate unnatural dialogues. This

is a very reasonable expectation, since the chatting conversation part, which does not directly contribute to good counseling, plays an important role in natural conversation.

To attain natural dialogues, we would like to successfully apply two methods (Minami et al., 2009):

- 1 Automatically obtaining POMDP parameters and action control that generates target dialogues.

Our proposed method automatically obtains the emission probabilities and observation probabilities of hidden states with a dynamic Bayesian network (DBN) based on expectation-maximization (EM) from a large amount of data. Then it sets rewards for POMDP and performs value iteration to train a policy.

- 2 Reflecting action predictive probabilities in action control.

Our method introduces hidden states that match actions with one-to-one correspondence. Then it sets the POMDP rewards to maximize the predictive probabilities of the hidden states using value iteration. This procedure can have the agent perform an action sequence by reflecting the statistical characteristics of the data (Minami et al., 2009).

The latter method is our original approach and different from that of (Fujita, 2007). Although this method might appear very similar to (Hori et al., 2009) in maximizing predictive probabilities of actions, (Hori et al., 2009) requires a look-ahead mechanism to obtain the next action during a decision-making process. Since our method calculates a policy in advance by value iteration, it does not require such a look-ahead mechanism. This chapter shows that our agent can generate a target action sequence as well as an action sequence that reflects the statistics of the training data simultaneously.

The basic POMDP formulation is described in Section 2, our proposed method is presented in Section 3, and the results and evaluations of simulation experiments with our action control algorithm are provided in Section 4. Finally, a discussion, future work, and our conclusion are given.

2. Partially Observable Markov Decision Process

2.1 POMDP Structure

A POMDP is defined as $\{S, O, A, T, Z, R, \gamma, \text{ and } b_0\}$. S is a set of states described by $s \in S$. O is set of observations o described by $o \in O$. A is a set of actions a described by $a \in A$. T is a set of the state transition probabilities from s to s' , given a , $\Pr(s'|s, a)$. Z is a set of the emission probabilities of o' at state s' , given a , $\Pr(o'|s', a)$. R is a set of expected rewards when the agent

performs action a at state s , $r(s, a)$. The basic structure employed is shown in Figure 1. Although it closely resembles HMMs, the difference is that POMDP has actions and rewards to control state transitions. The rhomboids show the fixed values, the dotted circles show the hidden variables, the solid circles show the observed variables, and the solid squares show the agent actions.

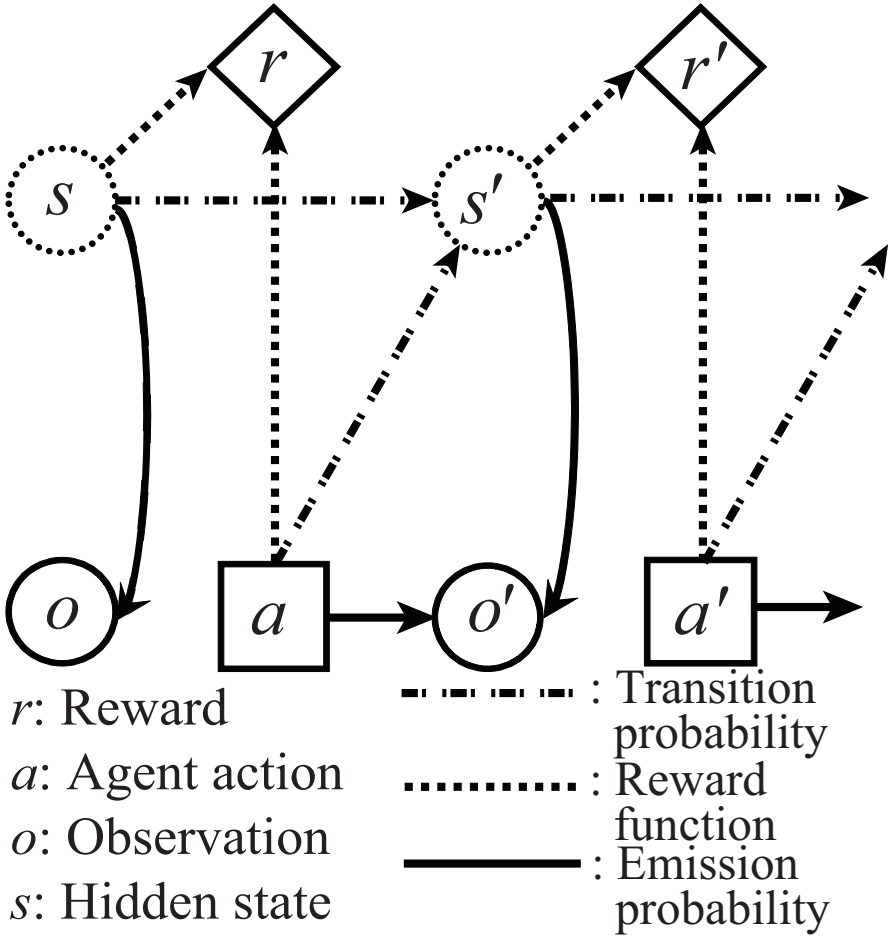


Figure 1. POMDP structure.

Before referring to γ and b_0 , we explain the state transition probability update method. In POMDP, since states are directly unobservable as in HMMs, we can only discuss their distribution. Here, suppose that the distribution of states $b_{t-1}(s)$ is known. Using the transition and emission probabilities, the

distribution update is performed by

$$b_t(s') = \eta \cdot \Pr(o'|s', a) \sum_s \Pr(s'|s, a) b_{t-1}(s), \quad (7.1)$$

where η is a factor so that the distribution summation is one. η is calculated by

$$\eta = \frac{1}{\sum_{s'} \Pr(o'|s', a) \sum_s \Pr(s'|s, a) b_{t-1}(s)}. \quad (7.2)$$

If the initial value of b is set as b_0 , $b_t(s')$ can be obtained iteratively using a recursive equation.

Using this distribution, the average discounted reward at time t , which is the objective function, can be obtained by averaging

$$V_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_s b_{\tau+t}(s) r(s, a_{\tau+t}), \quad (7.3)$$

where γ is a discount factor. POMDP obtains a policy that is a function from $b_t(s)$ to action a by maximizing the average discounted reward in an infinite time.

2.2 Running Cycle and Value Iteration

Figure 2 shows the running phase flow of POMDP. This procedure is as follows.

- 1 Current state $b(s)$ is given. Optimal action is executed by

$$a = \pi^*(b). \quad (7.4)$$

- 2 Accordingly, the user acts. Consequently, observation o is observed.
- 3 Next, $b(s)$ is calculated using Equation (7.1) from current $b(s)$ and the transition and emission probabilities.

π^* , which is called the policy as described above, is obtained by value iteration if the emission and transition probabilities are known. This policy, which is independent of time, maximizes averaged Equation (7.3). In infinite time, the optimal value function tends to reach the equilibrium point in an iterative manner called value iteration. Although this value iteration obtains the optimal policy, it is time consuming. PBVI comprises one approximate solution technique (Pineau et al., 2003).

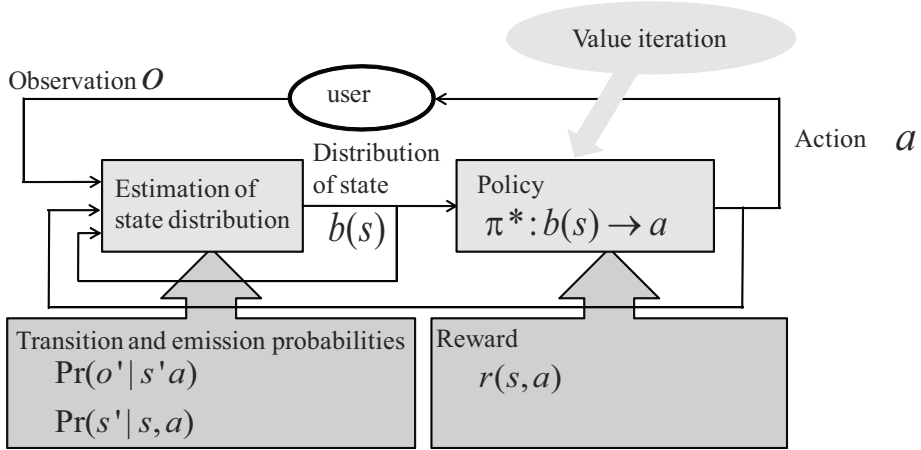


Figure 2. Flow of POMDP.

3. Dialogue Control using POMDP from Large Amounts of Data

3.1 Purpose of Dialogue Control

We assume that the statistics of the data are unknown and have to be estimated. We also assume that the data contain a set of dialogues we would like the agent to achieve. To find this set of dialogues, we evaluate whether a dialogue is a desired target dialogue for the agent by looking at its sequence. To do this, we produced a questionnaire to select the data. For example, we asked in the questionnaire whether the dialogue is good for general greetings. Such general greeting dialogues are comprised of a sequence of shaking hands, greeting each other, and talking. We call such protocols target dialogues. We assume that a set of dialogues contains target dialogues and non-target dialogues. Since some questionnaires are clear, it is easy to select the dialogue part in detail. However, some of them are unclear, so it is difficult to point out a precise part of the dialogue. Nevertheless, we think it is possible to select some sequences of dialogue that match the questionnaire. Under this condition, we would like to achieve two purposes: one is to have the agent perform the target action sequences, and the other is to have the agent perform action sequences that reflect the statistical characteristics of the data. We propose two methods to resolve the above issues:

- 1 Automatically acquiring POMDP parameters and obtaining policies that generate target dialogues.
- 2 Reflecting action predictive probabilities in action control.

3.2 Automatically Acquiring POMDP Parameters and Obtaining a Policy for Target Dialogues

We propose the POMDP training procedure in [Figure 3](#). First, the DBN structure is constructed and trained by the EM algorithm. Second, it is converted into a POMDP structure. Third, POMDP rewards are obtained. Finally, value iteration is performed to obtain a policy.

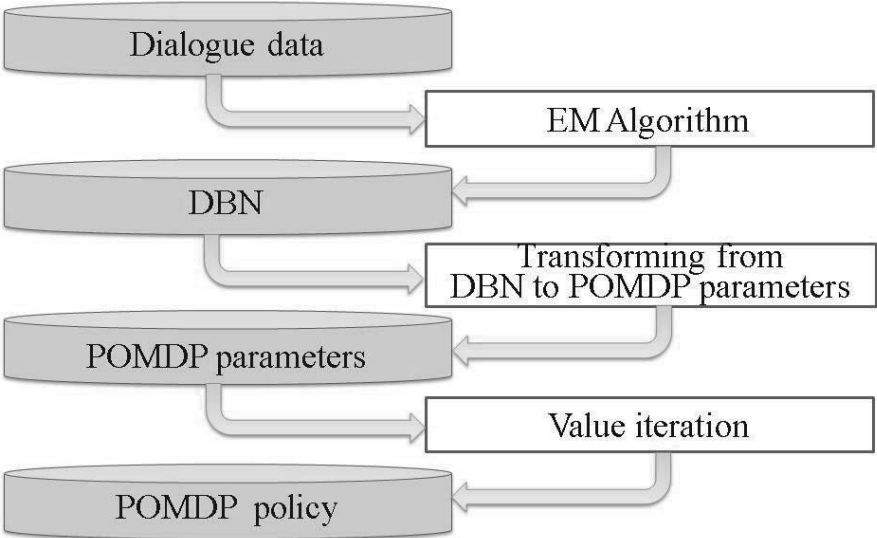


Figure 3. Policy generation procedure by POMDP.

POMDP is required to train the transition probabilities, the emission probabilities, and the rewards described in the previous section. Ordinary dialogue systems assume that the probabilities and the rewards are given. In this paper, these parameters are automatically trained from the data. We reduce the number of parameters using an approximation:

$$\Pr(o'|s', a) \approx \Pr(o'|s'). \quad (7.5)$$

We use this approximation for the following reason. s' is affected by a through $\Pr(s'|s, a)$. Consequently, $\Pr(o'|s')$ is indirectly affected by a through s' . Therefore, a is omitted from $\Pr(o'|s', a)$. Using this approximation, Figure 1 can be converted into Figure 4. The corresponding DBN shown in Figure 5 is used to train the probabilities in POMDP. DBN is trained using the EM algorithm. The dialogue data has some tasks, which we expect to be automatically classified based on their statistics using the EM algorithm.

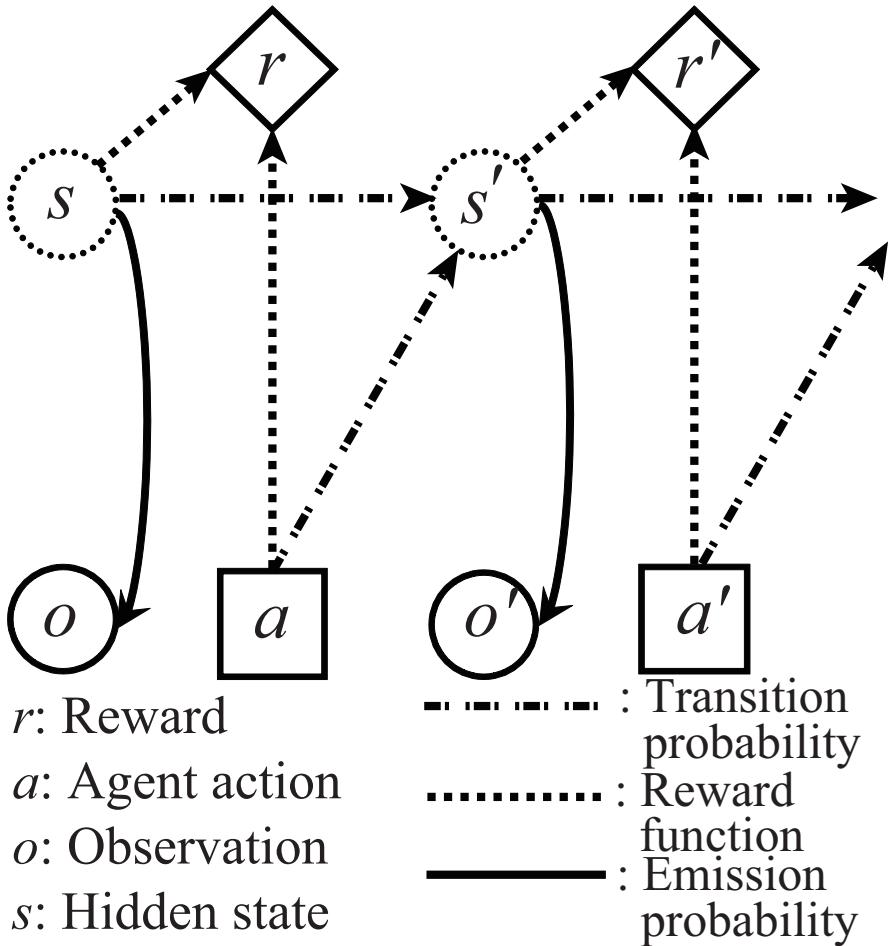


Figure 4. Proposed basic POMDP structure.

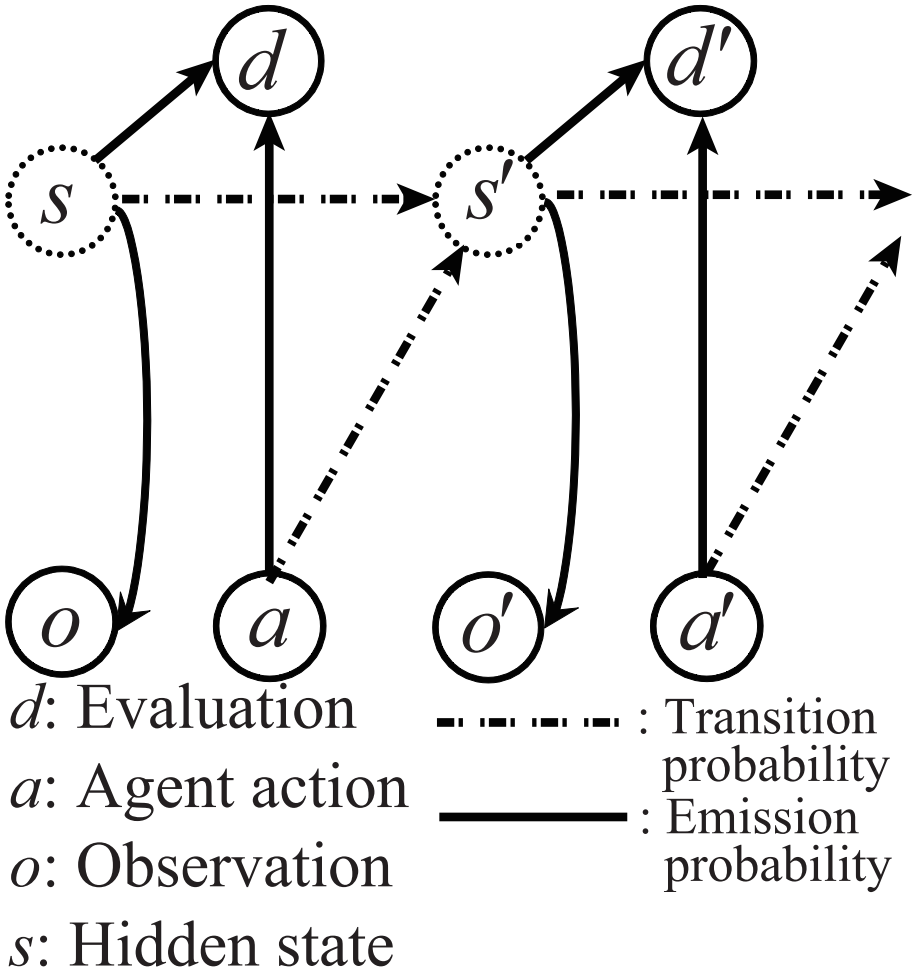


Figure 5. DBN structure corresponding to POMDP.

Examples of data are shown in Table 1. The agent and user perform eight actions: shaking hands, greeting, laughing, moving, speaking, nodding their heads, shaking their heads, and doing nothing. The user and the agent alternately perform an action from among the eight to have a dialogue. After the dialogue, the user evaluates whether the dialogue was a target dialogue by looking at its sequence. Based on this result, the user scores it. In this example, we show the period of a target dialogue by setting one for a certain length

while checking the questionnaire. In this case, the questionnaire for selecting target dialogues is very clear, we can distinguish where the target dialogue is. We used variable d for these scores, as shown in the right end row of [Table 1](#). These values have two purposes: one calculates rewards for POMDPs, and the other separates the DBN structure into two structures for target- and non-target dialogues.

Table 1. Example of dialogue data.

Observation o	Agent action a	User evaluation d
doing nothing	doing nothing	0
nodding	speaking	0
shaking hands	shaking hands	1
greeting	greeting	1
laughing	speaking	1
shaking head	speaking	1
greeting	greeting	1
shaking hands	shaking hands	1
doing nothing	shaking hands	0
greeting	doing nothing	0

The following processes are used in making a POMDP for target dialogue data:

- 1 Positive evaluation score is set to the target data as variable d ([Table 1](#)).
- 2 A DBN is trained. d is also treated as a random variable.
- 3 The DBN is converted to a POMDP, where we convert d values and d 's probabilities into POMDP fixed rewards by Equation (7.6).
- 4 We set the reward into the POMDP structure and perform value iteration.

After training the DBN, rewards are obtained from the d variable by

$$r_1(s, a) = \sum_{d=0}^1 d \times \Pr(d|s, a). \quad (7.6)$$

This means that if a state generates target dialogue data at a higher probability, the state should obtain higher rewards.

3.3 Reflecting Action Predictive Probabilities in Action Control

Our goal is to make an appropriate policy using the interaction data between users. Our target interaction characteristic is that the interaction should be processed based on probabilistic characteristics; however, target dialogues occur,

and the agent should obey them. The problem is obtaining policies that achieve this behavior by policy iteration.

First we introduce extra hidden POMDP and DBN states to the states in Figures 4 and 5 as $s = (s_o, s_a)$ (Figures 6 and 7). s_o is the same as the previous state in Figure 4. s_a is introduced for estimating the predictive probability of action a and for selecting a to maximize the predictive probability. This selection is performed by following the reward settings and the value iterations. This method is an extension of the method described in the previous section.

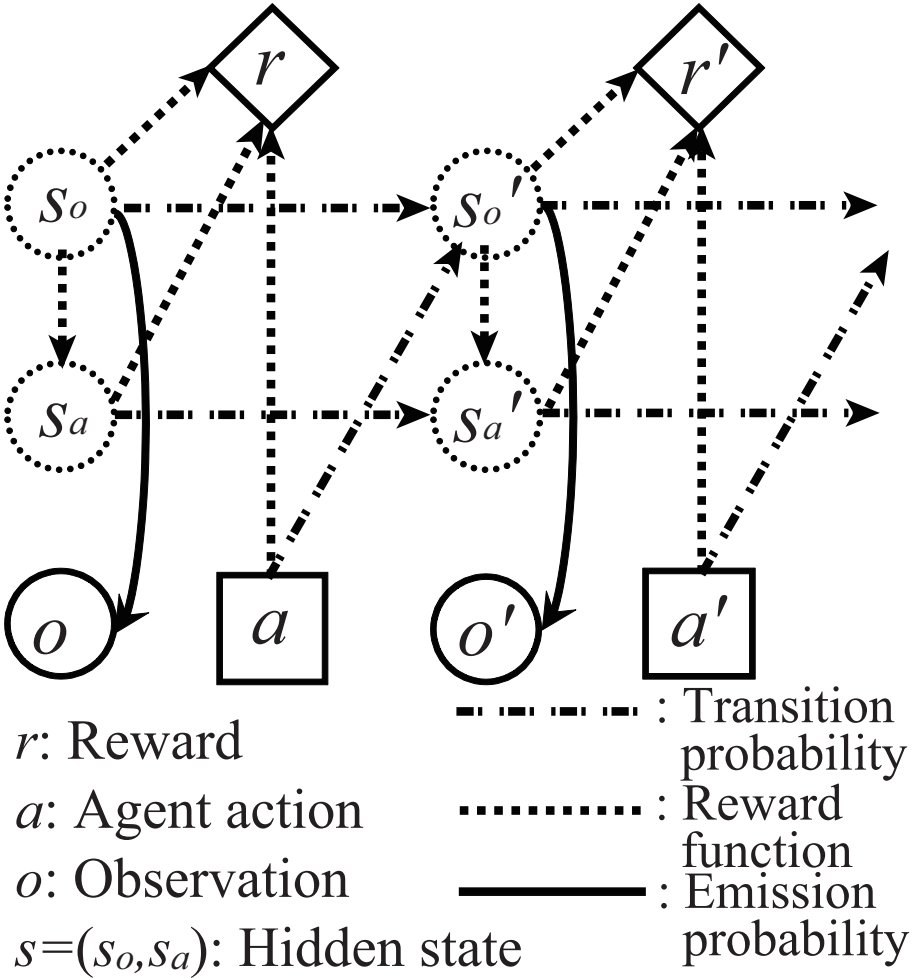


Figure 6. POMDP structure employed in this chapter.

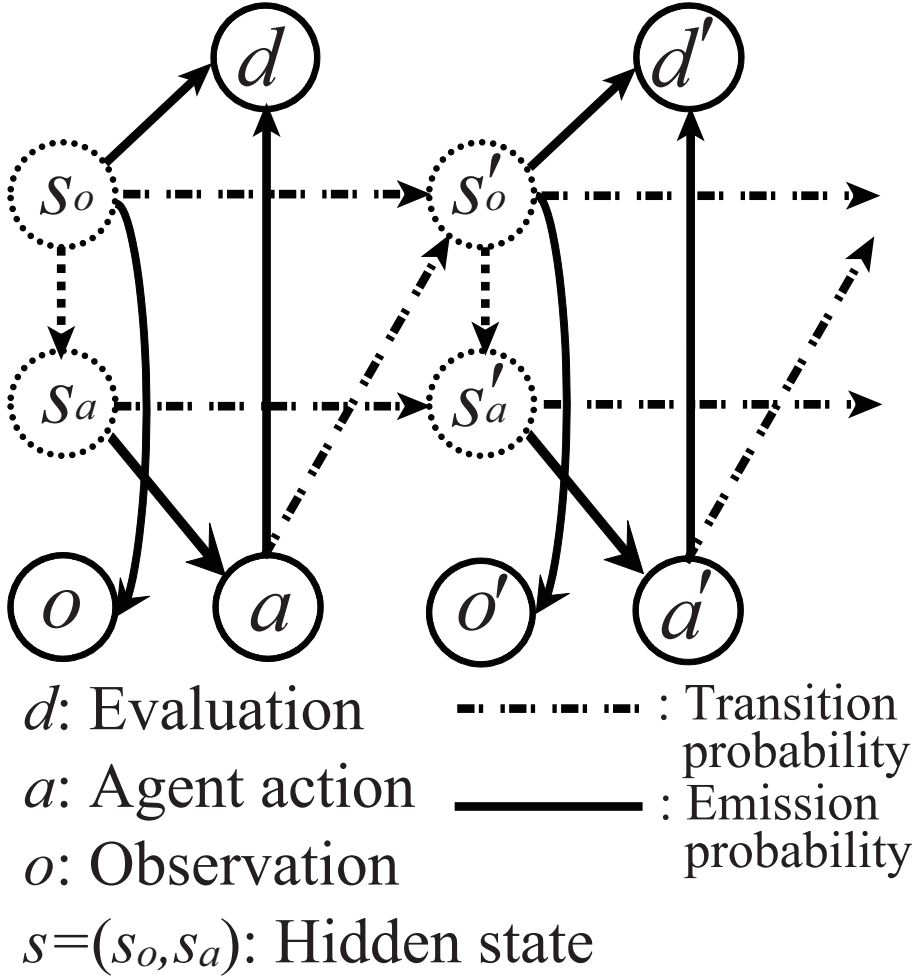


Figure 7. DBN structure employed in this chapter.

Due to increased parameters, probability approximations are used as follows (Figure 6):

$$\Pr(s'|s, a) \approx \Pr(s'_a | s'_o, s_a) \Pr(s'_o | s_o, a), \text{ and} \quad (7.7)$$

$$\Pr(o'|s', a) \approx \Pr(o' | s'_o). \quad (7.8)$$

Using these approximations, $b_t(s')$ can be written as

$$\begin{aligned}
b_t(s') &= b_t(s'_o, s'_a) = \eta \cdot \Pr(o'|s', a) \sum_s \Pr(s'|s, a) b_{t-1}(s) \\
&= \eta \Pr(o'|s'_o) \sum_s \Pr(s'_a | s'_o, s_a) \Pr(s'_o | s_o, a) b_t(s_o, s_a).
\end{aligned} \tag{7.9}$$

The corresponding averaged objective function can be obtained by averaging

$$V_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_s b_{\tau+t}(s_o, s_a) r((s_o, s_a), a_{\tau+t}). \tag{7.10}$$

We introduce $b_{\tau+t}(s_a) r((*, s_a), a_{\tau+t})$ into Equation (7.10) so that if $b_{\tau+t}(s_a)$ is high, POMDP may obtain a higher reward. If $a = s_a$, we set $\Pr(a|s_a) = 1$ in the DBN (Figure 6) so that s_a corresponds one-on-one with a . Based on this, if $a_t = s_a$ is given, we obtain

$$\begin{aligned}
&\Pr(a_t | o_1, a_1, \dots, a_{t-1}, o_t) \\
&= \sum_{s'_a} \Pr(a_t | s'_a) \Pr(s'_a | o_1, a_1, \dots, a_{t-1}, o_t)
\end{aligned} \tag{7.11}$$

$$= \Pr(s_a | o_1, a_1, \dots, o_{t-1}, a_{t-1}, o_t) = b_t(s_a), \tag{7.12}$$

where,

$$b_t(s_a) = \Pr(s_a | o_1, a_1, \dots, o_{t-1}, a_{t-1}, o_t) = \sum_{s_o} b_t(s). \tag{7.13}$$

This is for propagating the predictive probabilities of the actions into the probabilities of the hidden states. Our objective here is to select a_t so that the probability of a_t is maximized when $o_1, a_1, \dots, o_{t-1}, a_{t-1}, o_t$ are given. The reward should be set to satisfy this. This means the rewards should be set by maximizing Equation (7.12). To do this, we set $r_2(s = (*, s_a), a) = 1$ when $s_a = a$, where $*$ is arbitrary s_o . Otherwise, $r_2(s = (*, s_a), a) = 0$. Replacing r in Equation (7.10) into $r_1 + r_2$ as

$$r(s, a) = r_1((s_o, *), a) + r_2((* , s_a), a), \tag{7.14}$$

we obtain new function V_t . We modify reward definition r_1 described in Equation (7.6) using $*$ so that we can handle extra hidden states s_a . Here, the meaning of s_o is the same as s described in Equation (7.6).

The POMDP is then trained by value iteration to generate the policy. Using this formulation, the POMDP can select the action that simultaneously gives higher predictive probability of the action and obeys the target dialogue sequence.

4. Evaluation and Results

We performed a simulated experiment using data generated by the computer and also simulated natural interaction. Here, we assume that natural interaction has two types of interaction sequences: target dialogues and statistical dialogues. The target dialogues have typically long time patterns. The statistical dialogues are short-time reactions yielded by statistics such as bigram and trigram. Two types of patterns were prepared as target dialogues between the agent and the user. The following is one of the two patterns. They shake hands and greet each other. Then they talk randomly, laugh, and nod their heads. Finally, they greet and shake hands again. In the other pattern, first the user moves, and the agent does nothing. Then they greet each other, speak randomly, laugh, and nod their heads. Next they greet each other. Finally, the user moves, and the agent does nothing. Table 2 shows these data sequences. From turn 3 to $3 + \alpha$, we randomly generate the actions listed in the rows for these turns in 2. The length of α is randomly selected from 1 to 3. Consequently, the total length of target patterns varied from 5 to 7. The amount of these data is one-tenth of the total data.

Table 2. Two target dialogue patterns.

Turn	pattern 1		pattern2	
	User	Agent	User	Agent
1	Shaking hands	Shaking hands	Moving	Doing nothing
2	Greeting	Greeting	Greeting	Greeting
	Speaking	Speaking	Speaking	Speaking
3 + α	Nodding	Nodding	Nodding	Nodding
	Laughing	Laughing	Laughing	Laughing
4 + α	Greeting	Greeting	Greeting	Greeting
5 + α	Shaking hands	Shaking hands	Doing nothing	Moving

The rest of the data are randomly generated according to observation and action, so the conditional probabilities of action given the observation (shaking hands/shaking hands, greeting/greeting, laughing/laughing, moving/moving, speaking/speaking, speaking/nodding, speaking/shaking head, doing nothing/doing nothing) have the highest probabilities. The conditional probability $\Pr(a|o)$ of action a given observation o is shown in detail in Table 3

Since these data have no time dependency, the lengths of the sample sequences were set to be identical. When the lengths of the target dialogue samples were shorter than the fixed length, we added these data at the start and the end of the target dialogue samples (i.e., since target dialogues have time dependency, the length of the dialogue is set to be different). In all, 10,000 samples were made for the training data. One-tenth of the samples are for the target di-

alogues, for which we set the reward values to one per frame. In Section 3, we assume that the period of the target dialogue is known. Considering the actual situation, in this experiment, we assumed that the period is unknown. For example, in the case of a counseling dialogue, we do not know which precise part of the dialogue is good. Table 4 shows an example of this data. Although we do not point out an exact part of dialogue as the target dialogue, we can judge whether a dialogue sequence is good or not. Since we assume that we have a large number of such dialogues, DBN can well model these ambiguous data. After the training, a hidden state that likely generates target dialogues obtains a higher probability as the evaluation score. The DBN was trained using all of the data and converted to POMDP by the proposed method. The number of hidden states for s_o , s_a is 16 and 8, respectively. For value iteration, PBVI is used (Pineau et al., 2003).

Table 3. Conditional probabilities of action given observation ($\Pr(a|o)$).

		Action a							
		Hands	Gre.	Laugh.	Mov.	Speak.	Nodd.	Heads	Nothing
Observation o	Hands	0.65	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Gre.	0.05	0.65	0.05	0.05	0.05	0.05	0.05	0.05
	Laugh.	0.05	0.05	0.65	0.05	0.05	0.05	0.05	0.05
	Mov.	0.05	0.05	0.05	0.65	0.05	0.05	0.05	0.05
	Speak.	0.0	0.05	0.0	0.0	0.30	0.30	0.30	0.05
	Nodd.	0.0	0.05	0.05	0.05	0.80	0.0	0.0	0.05
	Heads	0.0	0.05	0.05	0.05	0.80	0.0	0.0	0.05
	Nothing	0.0	0.10	0.05	0.05	0.05	0.5	0.5	0.65

Table 4. Example of dialogue data.

Observation o	Agent action a	User evaluation d
doing nothing	doing nothing	1
nodding	speaking	1
shaking hands	shaking hands	1
greeting	greeting	1
laughing	speaking	1
shaking head	speaking	1
greeting	greeting	1
shaking hands	shaking hands	1
doing nothing	shaking hands	1
greeting	doing nothing	1

Target sequence

2,000 samples were used for the evaluation data. Target dialogue data were generated using the same algorithm to generate the training data. Only the observation data were used. We evaluated the action-generation results of two POMDPs: our proposed POMDP and a POMDP that gives rewards only for the target. The latter POMDP, which extends the POMDP described in (Fujita, 2007), is evaluated as the baseline.

The experimental results show that both POMDPs generated complete sequences for all of the data of the target dialogues. Table 5 shows the results of the conditional probability $\Pr_{proposed}(a|o)$ of action a given observation o in the 2,000 samples generated by the proposed method. For comparison, Table 6 shows the results of the conditional probability $\Pr_{baseline}(a|o)$ of action a given observation o in the 2,000 samples generated by the POMDP that gives rewards only for the target. Table 7 shows the conditional probability $\Pr_{training}(a|o)$ of action a given observation o . Using the probability structure shown in Table 3, the proposed POMDP tries to generate actions that have higher conditional probability in Table 7. In Table 5, some conditional probabilities obtained higher probabilities than those of the baseline, for example, $\Pr_{proposed}(Doingnothing|Doingnothing)$, $\Pr_{proposed}(Speaking|Nodding)$, $\Pr_{proposed}(Speaking|Shakinghead)$. These results confirm that the method simultaneously achieved action control for the target action sequence and maximized the predictive probabilities of actions.

Table 5. Conditional probability of $\Pr_{proposed}(a|o)$.

		Action a							
		Hands	Gre.	Laugh.	Mov.	Speak.	Nodd.	Heads	Nothing
Observation o	Hands	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Gre.	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	Laugh.	0.0	0.11	0.0	0.0	0.87	0.01	0.0	0.00
	Mov.	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.99
	Speak.	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
	Nodd.	0.0	0.12	0.0	0.0	0.33	0.55	0.0	0.0
	Heads	0.0	0.11	0.0	0.0	0.53	0.36	0.0	0.0
	Nothing	0.0	0.0	0.0	0.15	0.0	0.0	0.0	0.85

Although we can tune the weight value for the two rewards, weight tuning was not performed in the proposed method. This remains an issue for future work.

5. Discussion

The remaining points worthy of further consideration are listed as follows.

1 Obtaining state transition probabilities and emission probability.

There are several considerations for this issue. General dialogue agents using POMDPs assume that transition and emission probabilities are given or obtained by RL. In this paper, these probabilities are obtained from a large amount of data and depend on the targets that two users want to discuss. In RL, the probabilities may be considered independently of the target. However, our method trains those dependencies by the EM algorithm.

In addition, we introduced dialogue evaluation by people as an output variable in the DBN. Consequently, the DBN constructed different structures for target dialogue data and other dialogue data. This framework worked well. However, since this experiment was in a sense artificial, we have to evaluate many actual situations to check this method's generality.

Table 6. Conditional probability of $\text{Pr}_{\text{baseline}}(a|o)$.

		Action a							
		Hands	Gre.	Laugh.	Mov.	Speak.	Nodd.	Heads	Nothing
Observation o	Hands	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Gre.	0.0	0.79	0.05	0.15	0.0	0.0	0.0	0.0
	Laugh.	0.0	0.0	0.0	0.0	0.82	0.18	0.0	0.0
	Mov.	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.99
	Speak.	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
	Nodd.	0.0	0.12	0.0	0.0	0.0	0.88	0.0	0.0
	Heads	0.0	0.11	0.0	0.0	0.0	0.89	0.0	0.0
	Nothing	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

Table 7. Conditional probability of $\text{Pr}_{\text{training}}(a|o)$.

		Action a							
		Hands	Gre.	Laugh.	Mov.	Speak.	Nodd.	Heads	Nothing
Observation o	Hands	0.67	0.05	0.05	0.05	0.05	0.04	0.04	0.05
	Gre.	0.04	0.70	0.04	0.04	0.04	0.04	0.04	0.04
	Laugh.	0.04	0.04	0.63	0.04	0.06	0.07	0.04	0.06
	Mov.	0.04	0.04	0.04	0.56	0.04	0.04	0.04	0.20
	Speak.	0.0	0.0	0.07	0.0	0.30	0.31	0.27	0.04
	Nodd.	0.0	0.04	0.06	0.05	0.77	0.01	0.0	0.05
	Heads	0.0	0.04	0.04	0.05	0.08	0.0	0.0	0.05
	Nothing	0.0	0.1	0.05	0.1	0.05	0.04	0.05	0.62

2 Setting evaluation results for calculating rewards.

Setting evaluation results for calculating rewards is also an interesting issue in this study. In the experiment, we set rewards to all states that were assigned for target dialogue data, as shown in Table 4. Here, we supposed that the boundary of a dialogue is ambiguous. In a general task-oriented system, where the final goals of dialogues are important, such as ticket reservation and weather information systems, the rewards should only be set at the final state of task completion. If the final state is known, we can set the reward at the final node. However, we do not always know which is the final node. To find the final node, we should use the training data shown in Table 8. In this data, we set one at the end of the target sequence. Training DBN using such data, we can find the final states that likely output one. However, we believe that in some dialogue cases, the final goal is much less important than the sequences themselves. Since several possibilities exist for giving rewards to POMDPs, we should further investigate this issue.

3 Applying real data.

This issue is strongly related to the first issue. Our method trains dialogue structures from simulated dialogue data. General dialogues may have many kinds of targets concurrently. It would be difficult to capture dialogue structures from such natural dialogues, since we need a large number of states for these targets. However, this is a valuable pursuit: If we can capture the structure, we can generate a dialogue system from a large amount of data without spending excessive time making the system.

4 Interaction timing.

The timing issue is very important. In this paper, interactions between user and agent are performed one after another. We are planning to apply this method to a speech dialogue system. For speech dialogue, we can detect the end of an utterance with high accuracy. Therefore, we can detect the timing of changing a dialogue. However, interactions between user and agent generally occur simultaneously, and thus systems have to handle such interactions. Currently, our POMDP architecture is unable to do this, so there is room for further investigation.

6. Future Work

This work is part of our ambient intelligence project initiated in October 2004 (Minami et al., 2007) that has two main purposes: 1) showing future ideal lifestyles through research and developing communication science and

Table 8. Example of dialogue data for giving rewards.

Observation o	Agent action a	User evaluation d	
doing nothing	doing nothing	0	
nodding	speaking	0	
shaking hands	shaking hands	0	
greeting	greeting	0	
laughing	speaking	0	
shaking head	speaking	0	Target sequence
greeting	greeting	0	
shaking hands	shaking hands	1	
doing nothing	shaking hands	0	
greeting	doing nothing	0	

technologies; and 2) strategically developing transdisciplinary research areas, as symbolized by the term intelligence integration. Such terms as "ubiquitous" and "pervasive" approach a description of the concept. However, since such R&D studies tend to focus on computers and sensors, they are geared toward hardware or devices. That is, we will exploit our areas of expertise, including speech, sound, language, dialogue, vision, data retrieval, and networking to create a new style of ambient intelligence that places human intelligence and intellect at the forefront. This will lead to proposals of future lifestyles and clarify future specific issues. We discussed the possible lifestyles that could be realized by ambient intelligence and suggested specific concrete issues that must be surmounted. To achieve an embodied world of ambient intelligence, we must make prototype systems using state-of-the-art technologies. For this purpose, we focus on thought-evoking dialogues, which are interactions where agents act on the willingness of users to provoke their thinking and encourage their involvement in dialogues. Under this situation, since we have to simultaneously deal with much ambiguous information (e.g., user evaluations), making dialogue controls by hand is difficult. Consequently, our proposed POMDP dialogue control is essential, since it can be trained automatically from a large amount of dialogue data and can treat ambiguous information.

As an example of thought-evoking dialogue systems, we made a quiz dialogue system that transmits knowledge from the system to users while they are playing a quiz (Minami et al., 2007; Higashinaka et al., 2007). This prototype system automatically gives a quiz on a particular person whose name is selected from the Internet. Based on this system, we developed a multi-party dialogue system and experimentally analyzed how conversational agents stimulate human communication in thought-evoking multi-party dialogues between multiple users and multiple agents (Dohsaka et al., 2009). [Figures 8](#) and [9](#) show the multi-party speech dialogue system. We will apply our pro-

posed POMDP method to such multi-party dialogue systems. Currently, we are analyzing listening-oriented dialogues to build listening agents (Meguro et al., 2009a; Meguro et al., 2009b). We will also utilize these results and our proposed dialogue strategies (Higashinaka et al., 2003; Higashinaka et al., 2004) with the proposed method in this paper to develop a thought-evoking multi-party dialogue system.



Figure 8. Multi-agent quiz dialogue system (full view).

7. Conclusions

In this chapter, we presented a POMDP-based dialogue control scheme that can automatically acquire a dialogue strategy with a value iteration algorithm. This algorithm reflects the statistical characteristics automatically acquired with a large amount of dialogue data based on the agent's decision processes in selecting actions. Our simulated experiment's results indicate that the dialogue control algorithm can generate an action sequence whose predictive distribution is maximized while generating the target action sequence. We also described the current issues in a framework in which agents learn dialogue control from the data. Finally, our future plans were discussed.



Figure 9. Multi-agent quiz dialogue system (partial view).

Acknowledgments

This work was partially supported by a Grant-in-Aid for Scientific Research on Innovative Areas, "Formation of robot communication strategies" (21118004), from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

- Denecke, M., Dohsaka, K., and Nakano, M. (2004a). Fast Reinforcement Learning of Dialogue Policies using Stable Function Approximation. *IJC-NLP*, pages 1–11.
- Denecke, M., Dohsaka, K., and Nakano, M. (2004b). Learning Dialogue Policies using State Aggregation in Reinforcement Learning. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 325–328. ISCA.
- Dohsaka, K., Asai, R., Higashinaka, R., Minami, Y., and Maeda, E. (2009). Effects of Conversational Agents on Human Communication in Thought-Evoking Multi-Party Dialogues. In *Proceedings of the SIGdial 2009 Conference*, pages 219–224.
- Fujita, H. (2007). *Learning and Decision-Planning in Partially Observable Environments*. PhD thesis, Nara Institute of Technology.

- Goddeau, D. and Pineau, J. (2000). Fast Reinforcement Learning of Dialog Strategies. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1233–1236. IEEE.
- Higashinaka, R., Dohsaka, K., Amano, S., and Isozaki, H. (2007). Effects of Quiz-Style Information Presentation on User Understanding. In *Proceedings of Interspeech*, pages 2725–2728.
- Higashinaka, R., Miyazaki, N., Nakano, M., and Aikawa, K. (2004). Evaluating Discourse Understanding in Spoken Dialogue Systems. *ACM TSLP*, 1:1–20.
- Higashinaka, R., Nakano, M., and Aikawa, K. (2003). Corpus-based Discourse Understanding in Spoken Dialogue Systems. *ACL-03*, pages 240–247.
- Hori, C., Ohtake, K., Misu, T., Kashioka, H., and Nakamura, S. (2009). Weighted Finite State Transducer based Statistical Dialog Management. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 490–495.
- Keizer, S., Gasic, M., Mairesse, F., Thomson, B., Yu, K., and Young, S. (2008). Human and Robot Behavior Modeling for Probabilistic Cognition of an Autonomous Service Robot. In *RO-MAN*, pages 121–124.
- Kim, K., Lee, C., Jung, S., and Lee, G. G. (2008). A Frame-Based Probabilistic Framework for Spoken Dialog Management Using Dialog Examples. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, pages 120–127.
- Levin, E. and Pieraccini, R. (1997). A Stochastic Model of Computer-Human Inter-action for Learning Dialog Strategies. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 1833–1886.
- Levin, E., Pieraccini, R., and Eckert, W. (1998). Using Markov Decision Process for Learning Dialogue Strategies. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 201–204.
- Meguro, T., Higashinaka, R., Dohsaka, K., Minami, Y., and Isozaki, H. (2009a). Analysis of Listening-Oriented Dialogue for Building Listening Agents. In *Proceedings of the SIGdial 2009 Conference*, pages 124–127.
- Meguro, T., Higashinaka, R., Dohsaka, K., Minami, Y., and Isozaki, H. (2009b). Effects of Personality Traits on Listening-oriented Dialogue. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 104–107.
- Minami, Y., Mori, A., Meguro, T., Higashinaka, R., Dohsaka, K., and Maeda, E. (2009). Dialogue Control Algorithm for Ambient Intelligence based on Partially Observable Markov Decision Processes. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 254–263.

- Minami, Y., Sawaki, M., Dohsaka, K., Higashinaka, R., Ishizuka, K., Isozaki, H., Matsubayashi, T., Miyoshi, M., Nakamura, A., Oba, T., Sawada, H., Yamada, T., and Maeda, E. (2007). The World of Mushrooms: Human-Computer Interaction Prototype Systems for Ambient Intelligence. In *ICMI*, pages 366–373.
- Pineau, J., Gordon, G., and Thrun, S. (2003). Point-Based Value Iteration: an Anytime Algorithm for POMDPs. In *IJCAI*, pages 1025–1032.
- Poupart, P. (2005). *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. PhD thesis, University of Toronto.
- Roy, N., Pineau, J., and Thrun, S. (2000). Spoken Dialogue Management Using Probabilistic Reasoning. In *ACL 2000*, pages 93–100.
- Russell, S. and Norvig, P. (2003). *Artificial Intelligence: a Modern Approach Second Edition*. New Jersey: Prentice Hall.
- Schmidt-Rohr, S. R., Jakel, R., Losch, M., and Dillmann, R. (2008). Compiling POMDP Models for a Multimodal Service Robot from Background Knowledge. In *European Robotics Symposium*, pages 53–62.
- Smallwood, R. D. and Sondik, E. J. (1973). The Optimal Control of Partially Observable Markov Processes over a Finite Horizon. *Operations Research*, pages 1071–1088.
- Smith, T. and Simmons, R. G. (2004). Heuristic Search Value Iteration for POMDPs. In *UAI*, pages 520–527.
- Smith, T. and Simmons, R. G. (2005). Point-Based POMDP Algorithms: Improved Analysis and Implementation. In *UAI*, pages 542–547.
- Sutton, R. S. and Barto, A. G. (1998). *Introduction to Reinforcement Learning*. The MIT Press.
- Williams, J. (2007). Using Particle Filters to Track Dialogue State. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 502–507.
- Williams, J., Poupart, P., and Young, S. (2005a). Factored Partially Observable Markov Decision Processes for Dialogue Management. In *Proceedings of IJCAI Workshop on K&R in Practical Dialogue Systems*, pages 75–82.
- Williams, J., Poupart, P., and Young, S. (2005b). Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, pages 25–34.
- Young, S., Gasic, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K. (2010). The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management. *Computer Speech & Language*, 24(2):150–174.

Chapter 8

PROPOSAL FOR A PRACTICAL SPOKEN DIALOGUE SYSTEM DEVELOPMENT METHOD

A Data-Management Centered Approach

Masahiro Araki

*Department of Information Science, Kyoto Institute of Technology
Kyoto, Japan*

araki@kit.jp

Takashi Hattori

*Department of Information Science, Kyoto Institute of Technology
Kyoto, Japan*

hattokun@nifty.com

Abstract This chapter presents a new rapid prototyping method, Mrails, for a spoken/multi-modal dialogue system. In contrast to previous approaches that define individual dialogue states, our method automatically generates necessary dialogue states from a data model definition. This prototyping method, which has the capability of real data access and basic error management, can be implemented by reconstructing a dialogue transition script of Rails Web application framework. In this chapter, we present an overview of our approach and show rapid development examples of various types of spoken/multi-modal dialogue systems.

Keywords: Prototyping of spoken dialogue system; Multi-modal dialog system; Interactive presentation; User model.

1. Introduction

Over recent years, many spoken/multi-modal dialogue systems have been developed and some telephony-based spoken dialogue systems are now in

practical use. However, an urgent need exists for prototyping tools for spoken/multi-modal dialogue systems, especially those required by industrial companies.

Earlier prototyping tools for spoken dialogue systems (e.g., (McTear, 2004; Katsurada et al., 2005)) were primarily based on the finite-state model (Figure 1), in which a developer defines dialogue states and necessary information by manipulating a GUI (graphical user interface) tool.

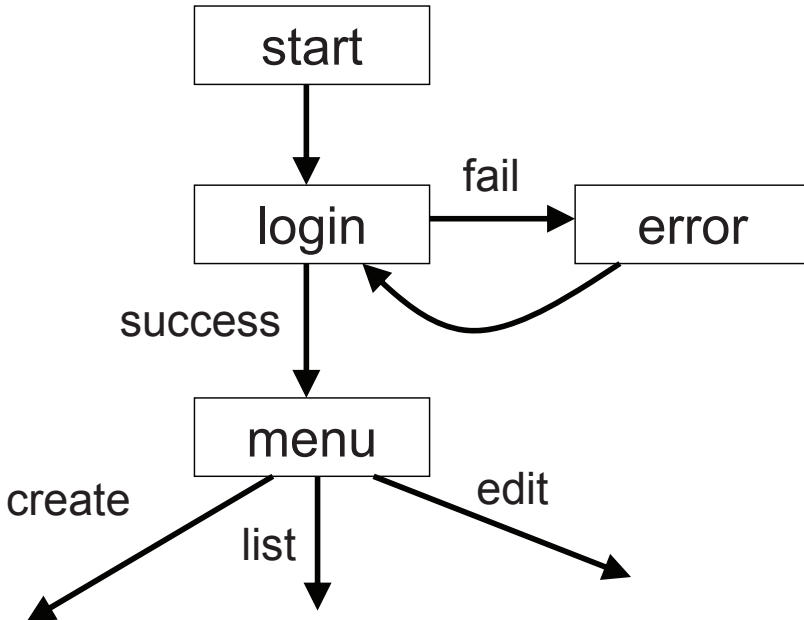


Figure 1. An example finite-state model of dialogue.

In the case of prototyping using such finite-state tools, the system developer defines a set of possible states of the target dialogue as a first step. Next, he/she constructs transition relations with conditions that govern transitions between states. After that, he/she designs other parts of the system for each state, such as a system prompt, a grammar of the user's utterances, and executable code that accesses back-end applications. Although this process seems to be intuitive, it requires many steps to construct even a simple dialogue system that can access databases as a background application. To manage such a flood of scripts and configuration problems, the framework approach is a promising solution.

Bohus et al. (2007) proposed an open-source framework for a spoken dialogue system. The framework was designed to be open, transparent, flexible, reusable and scalable. It is suitable for research purposes but not for rapid and easy prototyping because of its complexity.

We propose a new framework for a rapid prototyping approach to a spoken/multi-modal dialogue system. The core element of this framework is an extension of the Rails Web application framework, which we refer to as Mrails, to allow multi-modal interactions. Mrails enables speech input/output with a GUI using an ordinary Web browser. In contrast to previous prototyping tools for spoken dialogue systems, Mrails starts with a definition of the data structure. Then it automatically generates all the remaining components of a typical MVC (Model-View-Controller) model for a Web application. The controller and the model parts are automatically generated following a "Convention over Configuration" strategy, from which it follows that several variable names (a field name in a database, an object name in the scripting language and a name of a field variable of VoiceXML) can be identified, and a method for accessing the data can be dynamically generated by concatenating the variable name(s) and search condition(s). In addition, view files for basic operations, such as create, read, update, and delete the data, are automatically generated by the template engine. Therefore, the burdensome configuration definition between database records, objects in the scripting language and field variables in VoiceXML can be omitted.

We also present two extensions of the framework. One is a multi-modal interactive presentation system; the other is the inclusion of a user model management component within this framework. Both extensions are realized using our original multi-modal interaction description language MIML (Araki et al., 2005; Araki and Tachibana, 2006; Araki, 2007).

2. Overview of the Data-Management Centered Prototyping Method

Spurred on by the success of Ruby on Rails¹, several developer groups have initiated projects to develop dynamic language-based Web application frameworks, such as Catalyst for Perl, Cake for PHP, and Grails for Groovy. A common feature of these frameworks is to provide a typical code set and configuration files for an MVC-based Web application, which are automatically generated from the description of a domain model. The acronym MVC refers to Model-View-Controller. The main concept of an MVC model (Figure 2) is to separate the application logic (Model), the transitional information

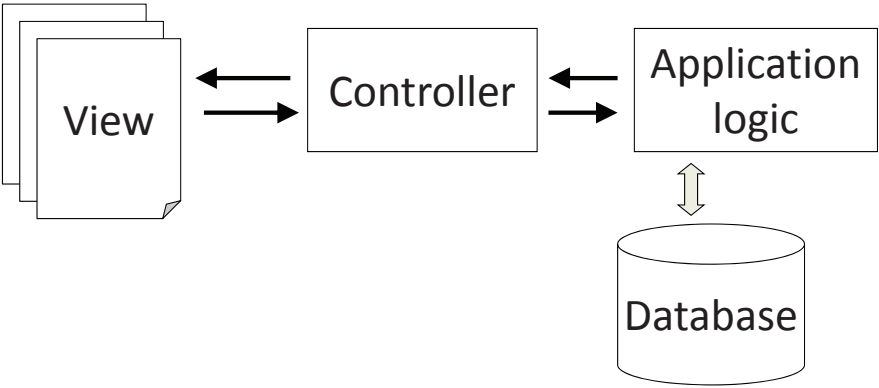


Figure 2. MVC model.

(Controller) and the presentation (View), in order to ease the development and maintenance of programs.

We use Grails² as the base framework for our multi-modal dialogue system prototyping tool. Grails can automatically generate the necessary files for an MVC-based Web application from the definition of a model object expressed using Groovy, which is a Java-based dynamic scripting language (Figure 3).

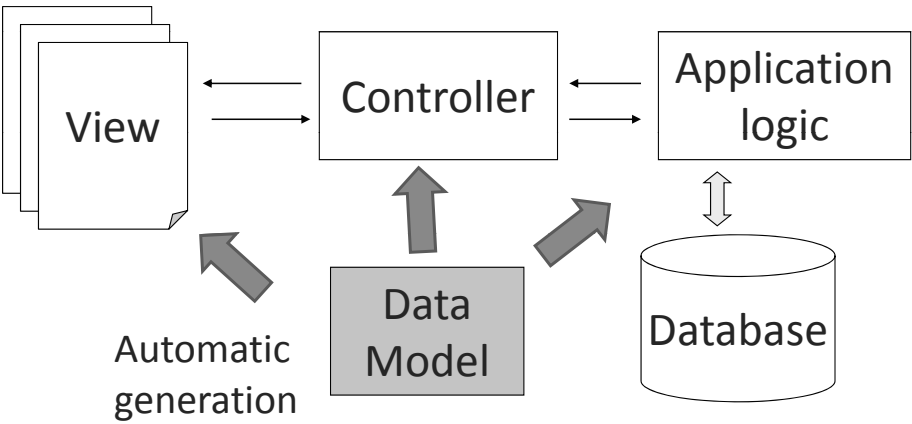


Figure 3. The concept of Grails.

The major steps for creating an MMI application within this framework are as follows.

- 1 Specify an application name
grails create-app banquet
- 2 Construct a data model
grails create-domain-class register
- 3 Edit the data model
- 4 Create the application
grails generate-all register
- 5 Generate the MMI view files and controller
java -jar mrails.jar

3. Prototyping of a Slot-Filling Dialogue System

In this section, we overview a method of prototyping of a slot-filling dialogue system (Araki, 2008) using Mrails. The task of the example slot-filling dialogue system is a banquet registration system for a conference. An attendee can register to attend the banquet by specifying his/her member ID and preferred food.

3.1 Data Model Definition

In order to develop a Web application for a banquet registration system, the developer has to define the model name and attributes (fields) of this model in the Groovy language (Figure 4).

```
class Register {
    Integer memberId
    String food

    static constraints = {
        food(inList: ["meat", "fish", "vegetable"])
    }
}
```

Figure 4. Model description in Grails.

In Figure 4, two attributes are declared. One is the `memberId`, of type integer. The other is the type of `food`, the value of which can be selected from "meat", "fish" and "vegetable", as indicated by the constraints definition. Using only this definition, Grails generates a typical Web application that can create a new entry, list all the entries and edit or delete a specific entry. For

each function, a corresponding view file, actually an HTML file with an embedded Groovy script, is generated.

3.2 Controller Script

Each HTTP request from the Web browser is submitted to the controller. A request is expressed in the following format (Figure 5).

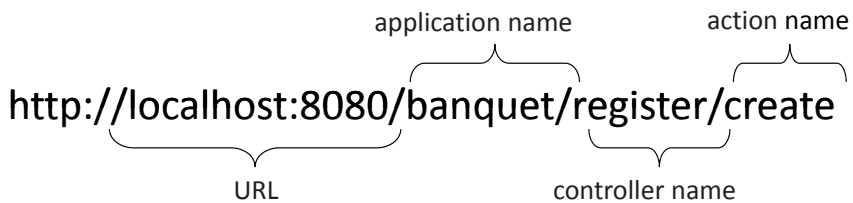


Figure 5. Grails URL format.

The controller, which is also generated automatically by Grails, dispatches a request for one of the functions above to the corresponding view file. This dispatcher information is written as a Groovy action script for each action request (see Figure 6).

```
class RegisterController {
    def index = { redirect(action:list,params:params) }

    def list = {
        script for listing all entries
    }

    def create = {
        def register = new Register()
        register.properties = params
        return ['register':register]
    }

    def delete = {
        script for deleting specified entry
    }

    ...
}
```

Figure 6. An example of a controller script (Part).

Grails generates basic states for data management, i.e., create, list, show and edit, from the model definition as class definitions in the object-oriented scripting language. In addition, Grails provides a basic controller for GUI applications, as shown in [Figure 7](#).

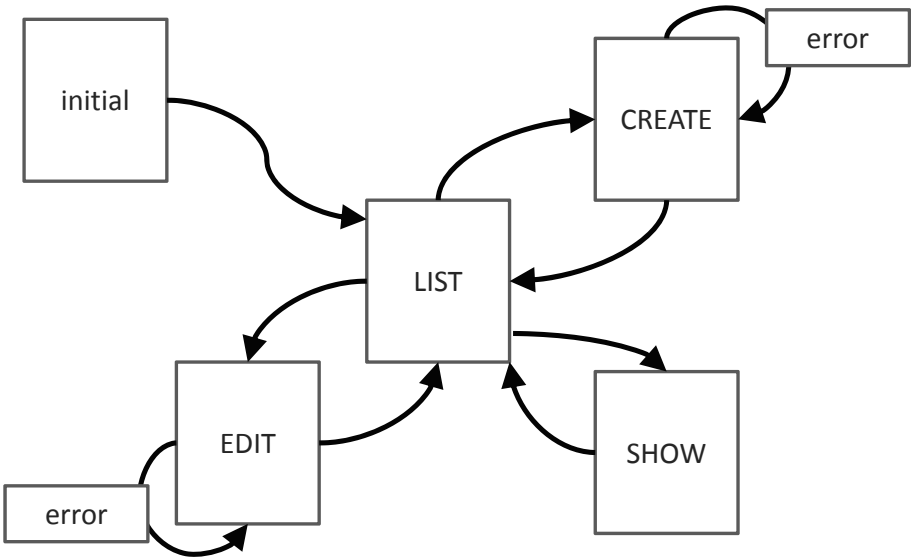


Figure 7. Basic states for GUI application generated by Grails.

3.3 View Files

For each function in the controller script, a corresponding view file (actually an HTML file with an embedded Groovy script), is generated ([Figure 8](#)) using a template for each type of view file.

The stored data can be accessed through a Grails tag library (specified by g:) embedded in HTML files.

By means of these steps, we can construct a GUI Web application for a banquet registration system ([Figure 9](#)), which is a slot-filling type of application.

3.4 Adding a Multi-Modal Interface to a GUI Web Application

To develop a multi-modal accessible Web application, Mrails performs two steps. The first one is to generate speech pages written in VoiceXML. The second step is to add a VoiceXML part to the HTML pages that are automati-

```

<html>
  <head>
    <title>Create Register</title>
  </head>
  <body>
    <div class="body">
      <h1>Create Register</h1>
      <g:form action="save" method="post" >
        <div class="dialog">
          <table> <tbody>
            <tr><td>Member ID:</td>
            <td> <input type='text' name='memberId' > </td>
            <tr><td>Food:</td>
            <td> <g:select name='food' ...> </td>
            ...
            <input type="submit" value="Create">
          </g:form>
        </div>
      </body>
    </html>

```

Figure 8. An example of a view file (Part).

cally generated by Grails. Therefore, the output pages of Mrails are written in XHTML+Voice.

3.5 Generating Speech Interaction

One major application area for adding speech modality to a GUI is to allow speech input functionality for a small device such as a mobile phone or PDA (Personal Data Assistant). Therefore, we focus on the speech input function as a first step towards rich multi-modal dialogue systems.

For example, in creating a new entry, the essence of generation of a VoiceXML file from HTML is to generate system-directive input guidance and to generate a speech grammar for a user's input. The essence of conversion is shown in [Figure 10](#) (unimportant parts are omitted and/or modified). In order to convert child elements of <form> elements in HTML (e.g. <input>, <select>) to VoiceXML <field> elements, we need information about what the system should say (i.e., the content of <prompt> element) and how the user will reply (i.e. <grammar> element). The content of <prompt> element is obtained as label information which is automatically generated by Grails from the attribute name of the model object with slight modifications.

The <grammar> element is generated by following rules:



Figure 9. An example of a view generated by Grails.

- 1 If the attribute has an option list (e.g., food attributes in Figure 4), the `<option>` elements are added to the `<field>` element, equivalent to defining a grammar for each option,

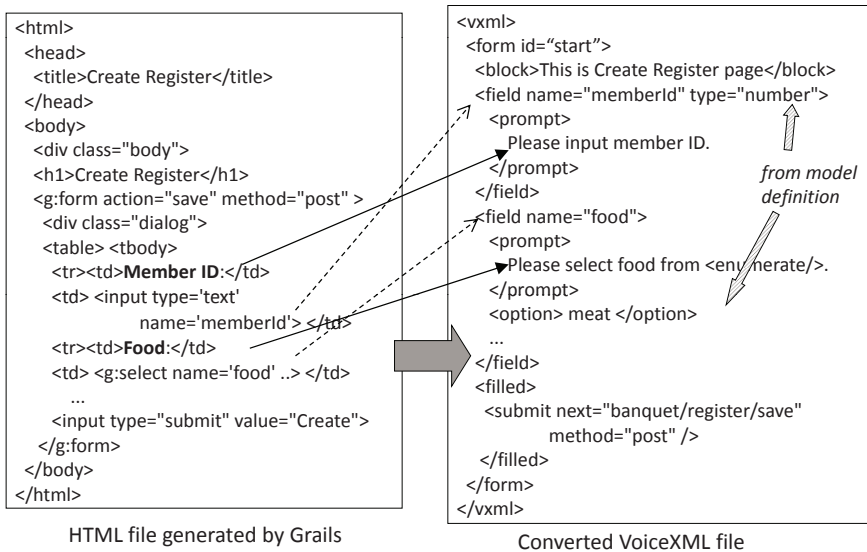


Figure 10. An example conversion from HTML to VoiceXML.

- 2 if the type of attribute has a corresponding built-in grammar in VoiceXML (e.g. date, number, currency), the type attribute of the `<field>` element is specified as the corresponding grammar type,
- 3 otherwise, only the skeleton of a grammar file and its link is generated. The developer has to define the grammar at a later stage.

3.6 Enabling Multi-Modal Interaction

As explained before, the target language for a multi-modal interaction is XHTML+Voice. Mrails assigns a VoiceXML (i.e. speech interaction) script as an event handler of an `<form>` element of HTML. Because a `<form>` element of HTML is itself created by Grails, all Mrails has to do is to assign a VoiceXML part which is generated in the previous step to each child element of `<form>` element of HTML. This process is illustrated in Figure 11.

3.7 Generation of Dialogue Flow

To develop a voice-accessible Web application using Grails, the view component has to be shifted from HTML to VoiceXML. In addition, the simple dialogue flow defined by the controller has to be modified to be suitable for voice interaction. Mrails adds such voice-specific information after generating the GUI-based Web application using Grails.

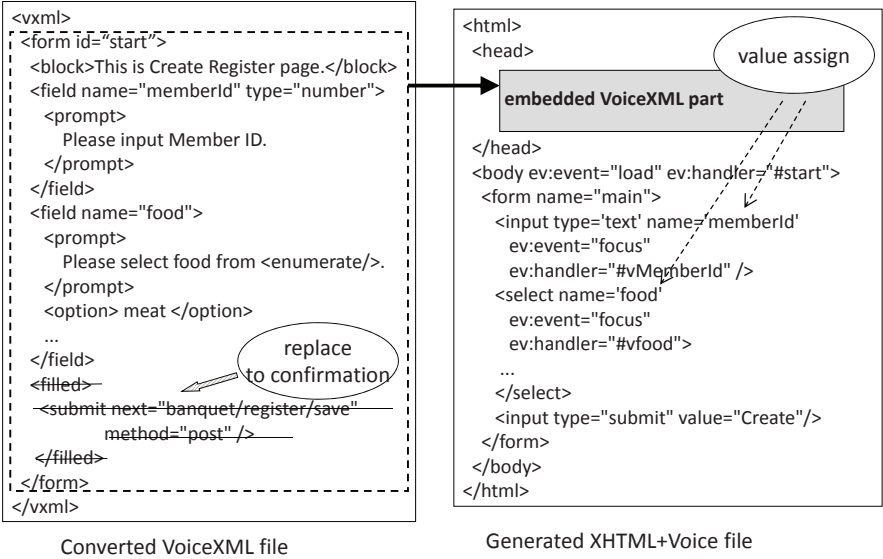


Figure 11. An example of generation of an XHTML+Voice file.

We have prepared two dialogue pattern templates (slot-filling and DB access) according to the direction of information flow (Araki et al., 1999). Figure 12 shows the dialogue pattern for slot-filling.

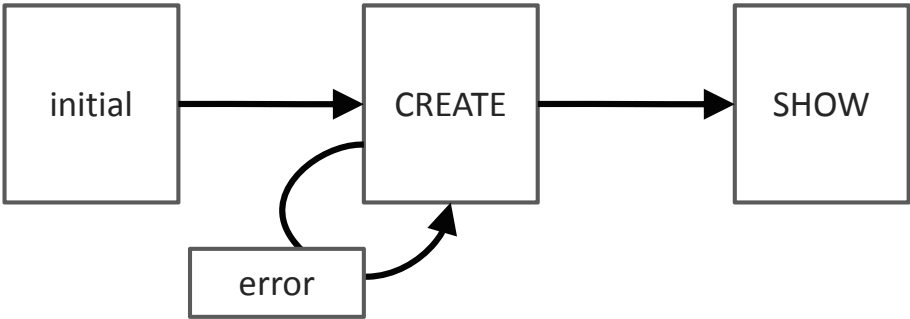


Figure 12. Transformed dialogue flow for spoken dialogue system (slot-filling).

3.8 The Result of the Prototyping

In this banquet registration example, all the steps described above can be performed within 5 minutes. This is because of Grails’ powerful automatic

generation ability and Mrails’ VoiceXML conversion. An example of a realized interaction is shown in [Figure 13](#).

After creating new registration data, the controller dispatches the dialogue to the listing view, in which the user can see all the registered records and can select the action of edit / delete an existing record. This construction is based on Rails’ standard CRUD (Create-Read-Update-Delete) functionality.

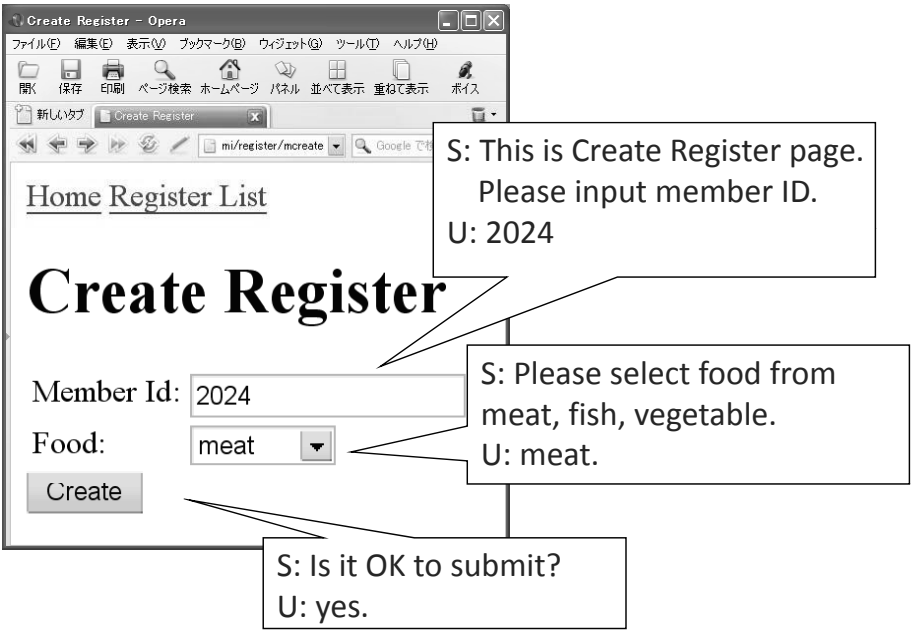


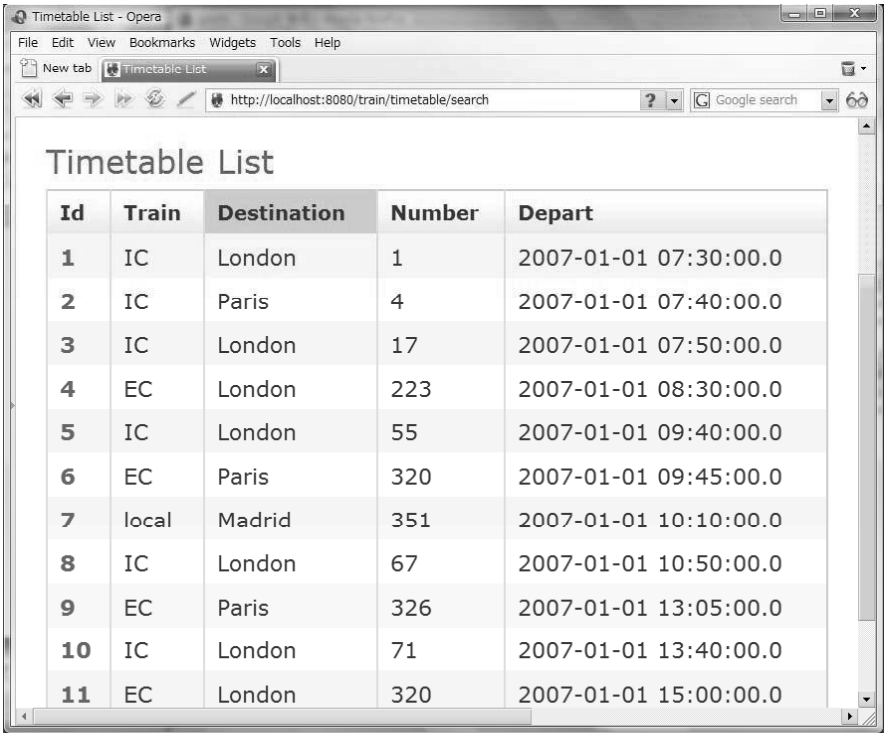
Figure 13. An example of slot-filling multi-modal dialogue.

4. Prototyping of a DB-Search Dialogue System

In this section, we present an example of prototyping a database-search type dialogue system using Mrails. As a task for prototyping, we selected a train timetable search. In this case, we assume we already have a database of train timetables ([Figure 14](#)).

The data model for this task is shown in [Figure 15](#).

Grails generates the basic data management states as described in Section 3. Then, Mrails modifies the view files for voice interaction and the controller file for spoken dialogue patterns, as shown in [Figure 16](#). The view file for data creation is transformed to the query input state because users cannot enter new data in this task. The blank constraints in the data model ([Figure 15](#)) specify which fields are required for making a query. In this case, the query must



The screenshot shows a web browser window titled "Timetable List - Opera". The address bar displays "http://localhost:8080/train/timetable/search". The page content is titled "Timetable List" and contains a table with 11 rows of train schedule data. The table has five columns: Id, Train, Destination, Number, and Depart. The data is as follows:

Id	Train	Destination	Number	Depart
1	IC	London	1	2007-01-01 07:30:00.0
2	IC	Paris	4	2007-01-01 07:40:00.0
3	IC	London	17	2007-01-01 07:50:00.0
4	EC	London	223	2007-01-01 08:30:00.0
5	IC	London	55	2007-01-01 09:40:00.0
6	EC	Paris	320	2007-01-01 09:45:00.0
7	local	Madrid	351	2007-01-01 10:10:00.0
8	IC	London	67	2007-01-01 10:50:00.0
9	EC	Paris	326	2007-01-01 13:05:00.0
10	IC	London	71	2007-01-01 13:40:00.0
11	EC	London	320	2007-01-01 15:00:00.0

Figure 14. An example of a train timetable.

```
class Timetable {
  String train
  String destination
  Integer number
  Date depart

  static constraints = {
    train(inList:['EC', 'IC', 'local'])
    destination(inList:['London','Paris','Madrid'] blank:false)
    number()
    depart(blank:false)
  }
}
```

Figure 15. Data model definition for the train timetable search.

include destination and departure times and optionally can include a train type and train number.

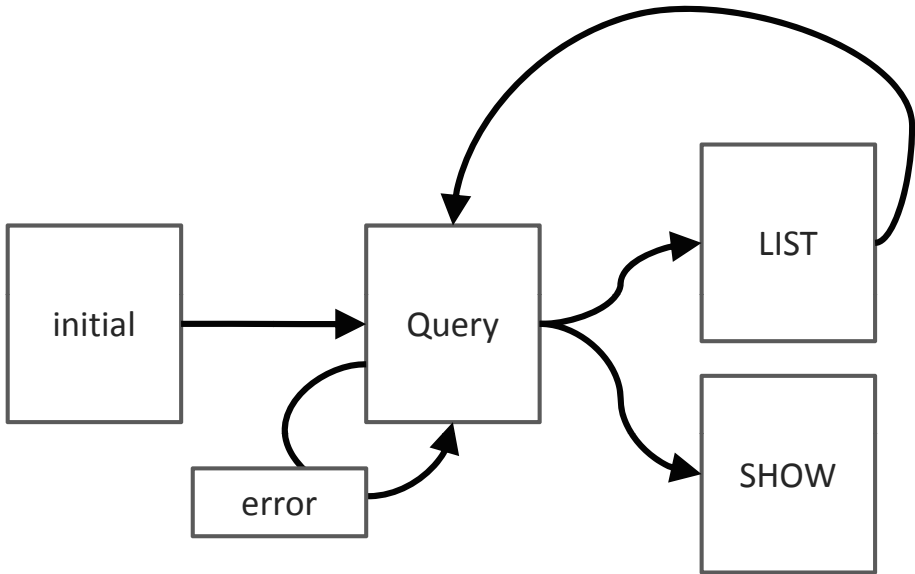


Figure 16. Transformed dialogue flow for a spoken dialogue system (DB access).

In the DB-search task, some kind of mixed-initiative behavior is required because sometimes a user's query does not contain adequate information for a DB search. The mixed-initiative dialogue is realized by the form interpretation algorithm (FIA) of VoiceXML. In the first step, the system asks the user an open question (e.g., "Please input search condition."). In ordinal FIA, fields that are not filled by the user's first utterance are one of the next targets of the dialogue.

However, this action is not suitable for the search condition input because in this situation, the user does not have to mention all the field variables. Therefore, in our implementation of mixed-initiative dialogues, all field variables are guarded by the "cond" attribute of the <field> element in order to construct the search query only from non-empty (i.e., mentioned by the user) variables. If there are necessary fields in the question, the developer can modify this constraint simply by deleting the "cond" attribute of the necessary <field> element.

5. Prototyping of a Multi-Modal Interactive Presentation System

In this section, we explain another prototyping method for a multi-modal interactive presentation system based on a data-management centered approach.

In this method, a scenario for multi-modal interaction is automatically generated from metadata of video content. In addition, a QA database and language model of speech recognition for a user’s question are generated from text annotation of video content.

Fine-grained video control and statistical speech recognition are beyond XHTML+Voice specifications. Therefore, we use MIML (Araki et al., 2005) as a multi-modal interaction description language.

The outline of prototyping is shown in Figure 17.

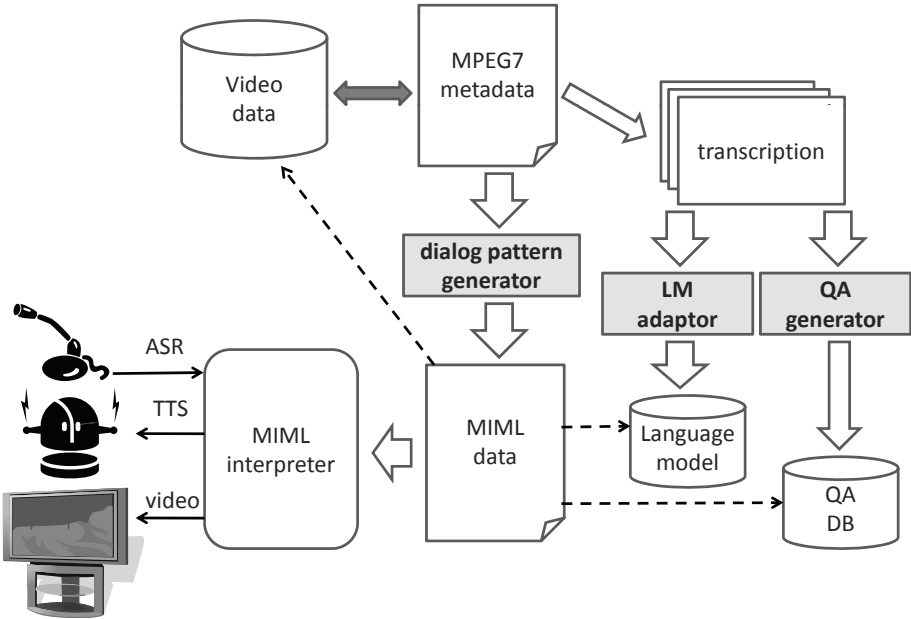


Figure 17. Overview of the MMI prototyping method.

5.1 Dialogue Pattern Generation from Metadata

Although there are several metadata formats for video content, MPEG-7 is an internationally standardized metadata format for multimedia content. An example of MPEG-7 metadata is shown in Figure 18. The target video content is an explanation of the Japanese traditional splash pattern “Kasuri” for “Kimono” garments. Video content is divided into segments indicated as <AudioVisualSegment> element in Figure 18. Each segment contains a scene name (as the value of “id” attribute), information about the time point and duration in the video, and hand-transcribed narrative (as the content of <FreeTextAnnotation> element).

```

<Mpeg7 type="complete"
  xmlns="http://www.mpeg7.org/2001/MPEG-7_Schema"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<CreationInformation id="001">
  <Creation>
    <Title>
      Kasuri: Japanese traditional splash pattern
    </Title>
  </Creation>
</CreationInformation>
<ContentDescription xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="AudioVisualType">
    <AudioVisual id="kasuri">
      <MediaLocator>http://ex.org/kasuri.mpg</MediaLocator>
      <MediaTime>
        <MediaRelTimePoint>PT0S</MediaRelTimePoint>
        <MediaDuration>PT10M44S</MediaDuration>
      </MediaTime>
      <TemporalDecomposition>
        <AudioVisualSegment id="introduction">
          <MediaTime>
            <MediaRelTimePoint>PT0S</MediaRelTimePoint>
            <MediaDuration>PT2M42S</MediaDuration>
          </MediaTime>
          <TemporalDecomposition>
            <TextAnnotation type="description">
              <FreeTextAnnotation>
                Kyoto is a city of Japanese traditional garments;
                Kasuri is one of ...
              </FreeTextAnnotation>
            </TextAnnotation>
          </AudioVisualSegment>
        </AudioVisualSegment>
      </TemporalDecomposition>
    </AudioVisual>
  </MultimediaContent>
</ContentDescription>
...

```

Figure 18. Example of MPEG-7 annotation.

In order to generate a dialogue pattern, MPEG-7 data is converted to MIML (Multi-Modal Interaction Markup Language) data. MIML is one of the markup languages for multi-modal interaction developed by our group.

MIML consists of three layers: the task, interaction and device layers. The task layer describes the general flow of dialogue for the task and defines background information, such as the data model, user model and device model. The interaction layer describes a series of interactions, which correspond to <form> elements in HTML or VoiceXML. The output information and input pattern is described in a device independent way. Filling the gap between these general descriptions and specific modality components, such as ASR, TTS and

robot, is the role of the device layer in MIML. An example of an MIML interaction layer document is shown in Figure 19.

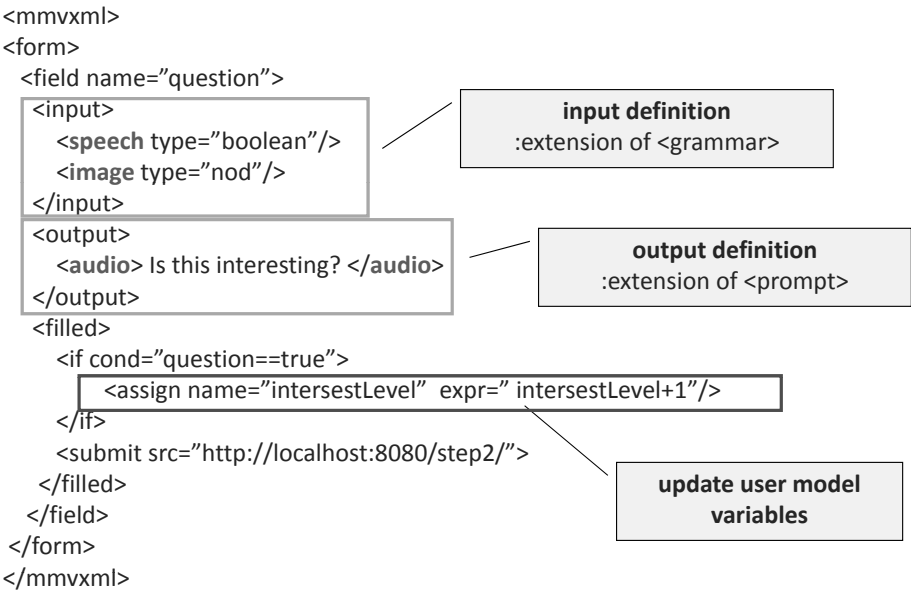


Figure 19. Example of an MIML document (confirmation for continuing a video presentation).

The conversion is done as follows:

- 1 Each video segment is separated into a <form> element that plays a video and interacts with the user.
- 2 In each <form> element, at a time just before presenting a video segment, a robot makes a comment about the segment. The comment is constructed using the scene name (e.g. "Let's watch an introduction.").
- 3 After presenting the video segment, the robot looks for confirmation of continuation of this presentation (e.g. "Is this interesting?") or asks a question (e.g. "Do you have any questions?").
- 4 A question-answering part is inserted in each segment. The detail of the QA part is explained in the next subsection.

5.2 Generation of QA Database

The target interactive presentation system has the capability of answering questions from the user. The mechanism for answering questions is a database-based method as in (Nisimura et al., 2005).

In a DB-based QA method, a set of pairs of questions and corresponding answers are stored in a DB. When the user's question is input, the most similar question entry in the DB is matched and the corresponding answer is output.

With our method (see [Figure 20](#)), a QADB is generated from the narrated text in the metadata. For each sentence in the narrated text, a new question is generated by replacing the content word by a Wh-question word. For example, if the sentence "Kasuri was developed in the Heian period." exists in the narrated text, a pair consisting of the question "When was Kasuri developed?" and the original sentence (as the answer) is stored in the QADB.

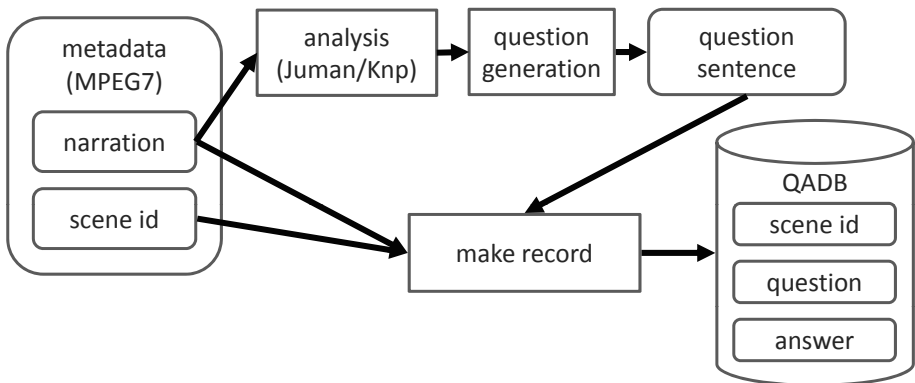


Figure 20. Generation of QADB.

5.3 Adaptation of the Language Model

In order to recognize a user's question, it is necessary for the language model in ASR to adapt the target domain. However, it is not practical to prepare a large in-domain dialogue corpus in such a rapid prototyping situation. Therefore we decide to adapt LM to the domain using Web resources. We used an LM adaptation tool (Misu and Kawahara, 2006) and the CIAIR In-car speech database (Japanese) (Kawaguchi et al., 2004). The adaptation procedure is shown in [Figure 21](#).

- 1 A baseline LM is constructed by merging the QADB generated by the procedure explained in the previous subsection and a general spontaneous speech corpus (in this case, we use CIAIR In-car speech database).
- 2 A set of Web search queries is constructed from the QADB by extracting keywords in the sentence. Using these queries, Web pages related to the target contents are gathered using a Web search engine.

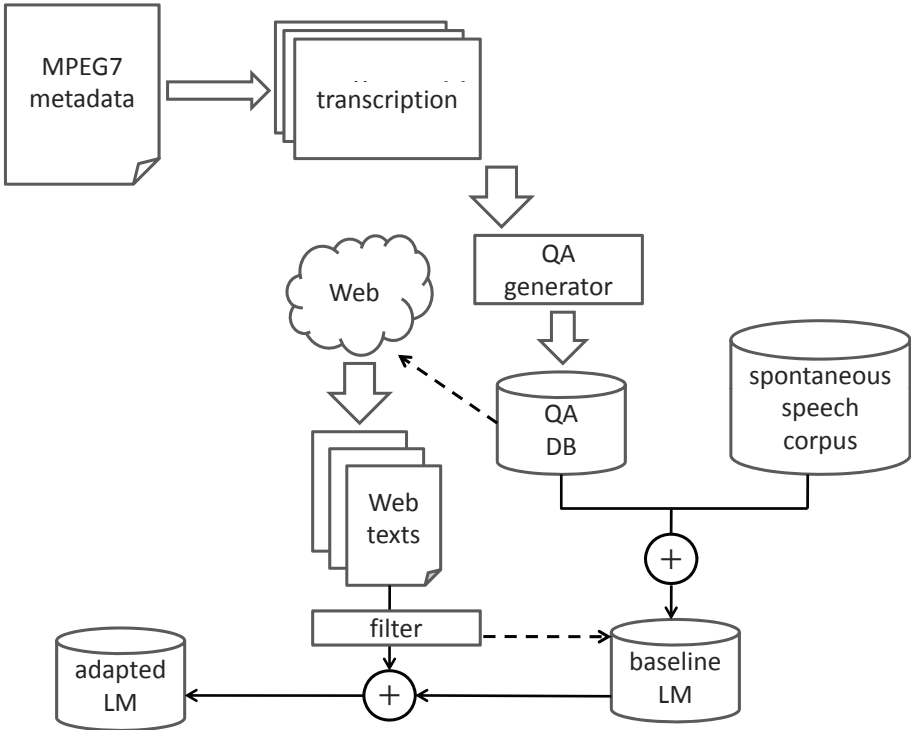


Figure 21. Adaptation procedure of LM.

- 3 These Web pages are filtered measuring the test set perplexity compared to the baseline LM and well-adapted pages are used to adapt the baseline LM to the target domain LM.

5.4 Implementation and Evaluation

Using the proposed prototyping method, we implemented two interactive presentation systems. The first is an explanation of the Japanese traditional splash pattern "Kasuri" for "Kimono" garments, and the second is an explanation of how to cook fried noodles.

The Kasuri video has about 11 minutes of content and 8 segments. The fried noodle cooking video has about 9 minutes of content and 4 segments. The narration of each video is transcribed. These contents are manually annotated following the MPEG-7 standard.

Although the two target contents are quite different in domain, the dialogue pattern generation, QADB generation and LM adaptation are done in the same way.

Table 1. Word Accuracy for Kasuri content (%).

	correct	phonetically correct	incorrect
Web60K	49.9	6.6	43.4
Adapted LM	67.3	11.6	21.1

Table 2. Word Accuracy for fried noodle cooking content (%).

	correct	phonetically correct	incorrect
Web60K	65.0	4.1	30.9
Adapted LM	77.8	4.1	18.2

In order to evaluate the effect of language model adaptation, we collected a set of questions from 10 persons that were supposed to have appeared in an interactive situation. We obtained 62 questions on the Kasuri content and 31 questions on the fried noodle cooking content. Some example questions are shown in Figure 22 (original Japanese questions are translated into English).

Kasuri:
When was Kasuri developed?
How much does Kasuri clothing cost?
Where can I get Kasuri clothing?
Fried noodle cooking:
What kind of ingredients are needed?
How many noodle blocks are needed?
When do I add the noodles?

Figure 22. Example questions.

Each question was read by 3 persons. As a result, we obtained 186 speech data items for the Kasuri content and 93 speech data items for the fried noodle cooking content. We used Julius (Lee et al., 2001) as a decoder. As an acoustic model, the PTM triphone model that is distributed with the Julius dictation kit was used. We compared two language models: the Web60K model also distributed with the Julius dictation kit and our adaptation model.

The results of the speech recognition experiment are shown in Tables 1 and 2. Phonetically correct means that there was a substitute error with a word that had the same phonetic sequence. Typically, an original Chinese character word was recognized as a Hiragana character. This caused a mismatch error in calculating similarity in the QADB search in our method.

We observed a considerable increase of accuracy for both contents. Therefore we can conclude that our proposed method is effective for various types of content.

6. Incorporation of the User Model

Some multi-modal interactive platforms are assumed to be used by a small number of people for each system. For example, a personal robot can be used in a household (ordinarily, one to five people), and a mobile phone can be used by only one person.

In such situations, adaptability of the system to each user is a very important characteristic. In order to realize user adaptability, a representation of the user model and an update mechanism of the attribute values in the user model are necessary.

Generally, in earlier user adaptable interaction systems, the user model representation and its update procedure have been embedded in system code. Therefore, it is not easy to port one user-modeling module to another system. In order to deal with the portability problem of multi-modal interaction systems, a multi-modal interaction markup language MIML is useful for handling user modeling variables.

In this section, an advanced user modeling component and user adaptation mechanism achieved through interactions are proposed. The proposed user modeling description enables the developer to define a set of user model attributes and possible causal relations between them. In addition, these user model attributes can easily be used for control information for multi-modal interactions.

6.1 User Model in Multi-Modal Interaction Systems

Several studies have examined user modeling methodologies and adaptation techniques to a specific user or group of users (Ardissono et al., 2005). However, few studies have examined the development methodologies of multi-modal dialogue systems with adaptable user modeling functions.

Heckman et al. proposed the UserML user model markup language and the GUMO user modeling ontology for realizing decentralized user modeling in ubiquitous computing (Heckmann and Krueger, 2003; Heckmann et al., 2005). The goal of UserML and GUMO is not to examine a specific aspect of user modeling but to develop a general architecture (or framework) for realizing various user adaptable interaction systems in a ubiquitous computing environment.

The proposed user modeling component and user adaptation mechanism deal with a more inclusive aspect of multi-modal interaction system development, including UserML, as a user modeling markup language.

6.2 User Model Component of MIML

In the previous specification of MIML (Araki et al., 2005), user model variables were treated as global variables of an interaction session. In the present chapter, an extension of the `<userModel>` element of MIML is proposed in order to deal with the ontology-based user model representation and the persistence mechanism using a markup language other than MIML.

As a user modeling markup language, userML (Heckmann and Krueger, 2003) is considered. An example of a user model definition is shown in [Figure 23](#).

```
<UserModel>
  <UserData id="123">
    <category> preference.news.society </category>
    <range> low-medium-high </range>
    <value> medium </value>
    <ontology> http://kit.ac.jp/UserOL/ </ontology>
  </UserData>
</UserModel>
```

Figure 23. Example of User modeling markup.

For each user, who is identified by an `id` attribute of the `<UserData>` element, the user modeling variable (indicated by the `<category>` element) and its value (specified by the `<value>` element among the range of the `<range>` element) are represented by XML. The user modeling variables are located at the task ontology specified by the `<ontology>` element.

Using this ontology-based user modeling, each system can develop its own task ontology and can be connected to other existing general ontologies

6.3 Functions for User Adaptation

In an RSS-based news reader application, for example, if the user listens to sports topics (or reads the headlines in a graphical display), the user model variable of `preference.news.sports` can be assigned to be "high" at the end of this interaction.

This user model variable is saved as an XML representation. Therefore, the system can recall this value at the next interaction with the same user. This is one example for implementing of adaptability to a specific user. If the developer wants to use a more sophisticated user model update mechanism,

he/she can implement the mechanism in Groovy Script as a service component of Grails.

Using this persistence of user model variables, a data mining method can help to determine the value of an unobserved user model variable from the behavior of other users. For example, if such a user model representation is converted to the AIFF format and WEKA Bayesian network structure learning (Witten and Frank, 2005) is used, the Bayesian network shown in Figure 24 can be obtained.

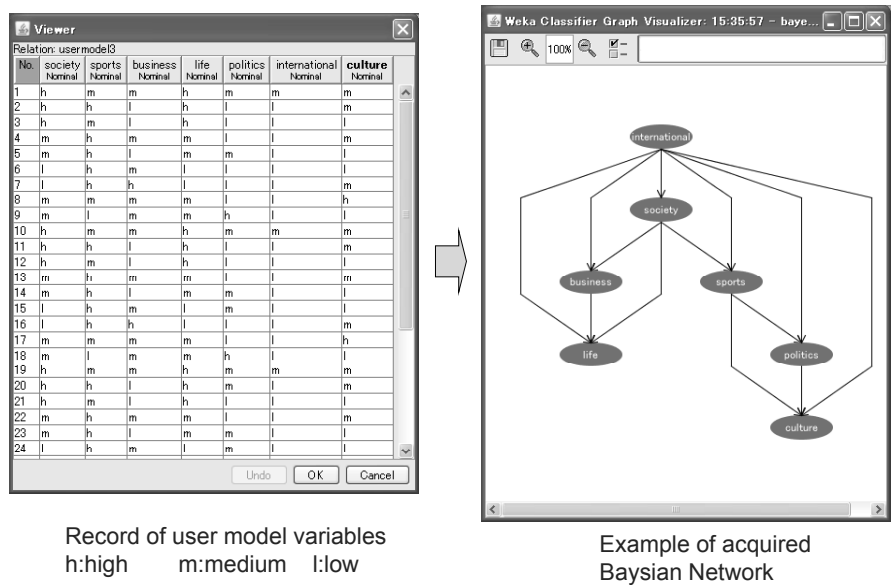


Figure 24. Acquired Bayesian network from simulated log data.

7. Conclusions

We have proposed a rapid prototyping system for a multi-modal dialogue system Mrails. It enables speech input/output with a GUI using an ordinary Web browser. Our contribution to the Rails framework is to (1) to add the voice interaction part to view files (as XHTML+Voice) automatically, (2) to generate a grammar definition following a data definition and (3) to add a mixed-initiative interaction pattern to apply to speech-based interactive systems. In addition, we introduce another framework for developing multimodal interactive presentation systems and user adaptable systems.

Acknowledgments

We thank the members of our group, in particular Mr. Yoshinori Minami and Mr. Yoshihiro Moritoki, for their help in implementing the system.

Notes

1. <http://rubyonrails.org/>
2. <http://www.grails.org/>

References

- Araki, M. (2007). Proposal of a Markup Language for Multimodal Semantic Interaction. In *Proceedings of Workshop on Multimodal Interfaces in Semantic Interaction*, pages 58–62.
- Araki, M. (2008). Filling the Gap between a Large-Scale Database and Multimodal Interactions. In Tokunaga, T. and Ortega, A., editors, *Third International Conference on Large-Scale Knowledge Resources, LKR 2008, LNCS*, volume 4938, pages 179–185. Springer.
- Araki, M., Komatani, K., Hirata, T., and Doshita, S. (1999). A Dialogue Library for Task-Oriented Spoken Dialogue Systems. In *Proceedings of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 1–7.
- Araki, M., Kouzawa, A., and Tachibana, K. (2005). Proposal of a Multimodal Interaction Description Language for Various Interactive Agents. *IEICE Trans. INF. & SYST*, E88-D(11):2469–2476.
- Araki, M. and Tachibana, K. (2006). Multimodal Dialog Description Language for Rapid System Development. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 109–116.
- Ardissano, L., Brna, P., and Mitrovic, A., editors (2005). *User Modeling 2005: 10th International Conference, UM 2005*. Springer.
- Bohus, D., Raux, A. A., Harris, T. K., Eskenazi, M., and Rudnicky, A. I. (2007). Olympus: an Open-Source Framework for Conversational Spoken Language Interface Research. In *Proceedings of HLT-NAACL 2007 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, pages 32–39.
- Heckmann, D. and Krueger, A. (2003). A User Modeling Markup Language (UserML) for Ubiquitous Computing. In *Proceedings of the Ninth International Conference on User Modeling*, pages 393–397.
- Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., von Wilamowitz-Moellendorff and, M. Ardisson, L., Brna, P., and Mitrovic, A. (2005). GUMO - the General User Model Ontology. In *User Modeling 2005: 10th International Conference, UM 2005*, pages 428–432. Springer.

- Katsurada, K., Sato, K., Adachi, H., Yamada, H., and Nitta, T. (2005). Rapid Prototyping Tool for Constructing Web-based MMI Applications. In *Proceedings of Interspeech*, pages 1861–1864.
- Kawaguchi, N., Matsubara, S., Yamaguchi, Y., Takeda, K., and Itakura, F. (2004). CIAIR In-Car Speech Database. In *Proceedings of Interspeech*, pages 2789–2792.
- Lee, A., Kawahara, T., and Shikano, K. (2001). Julius — an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 1691–1694.
- McTear, M. F. (2004). *Spoken Dialogue Technology*. Springer.
- Misu, T. and Kawahara, T. (2006). A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proceedings of Interspeech*, pages 9–13.
- Nisimura, R., Lee, A., Yamada, M., and Shikano, K. (2005). Operating a Public Spoken Guidance System in Real Environment. In *Proceedings of Interspeech*, pages 845–848.
- Witten, I. H. and Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques, second edition*. Morgan Kaufmann.

Chapter 9

QUALITY OF EXPERIENCING MULTI-MODAL INTERACTION

Benjamin Weiss, Sebastian Möller, Ina Wechsung and Christine Kühnel

Quality & Usability Lab, Deutsche Telekom Laboratories

TU Berlin, Germany

{ BWeiss, Sebastian.Moeller, Ina.Wechsung, Christine.Kuehnel }@telekom.de

Abstract In this chapter, we discuss the contributions of different modalities to the overall quality of multi-modal interaction. After reviewing some common systematics and findings concerning multi-modality, we present experimental results from several multi-modal scenarios, involving different (human-to-human and human-to-machine) interaction paradigms, different degrees of interactivity, and different (speech, audio, video, touch, gesture) modalities. The results show that the impact of each modality on overall quality in interaction depends heavily on the scenario and degree of interactivity. Complementary modalities are not considered in this paper, but the models presented allow predicting overall system quality on the basis of individual modality ratings with an appropriate accuracy. These models still have to be validated in order to be used as tools for system developers estimating whether adding modalities will have an impact on the quality experienced by the user.

Keywords: Usability; User experience; Perceived quality; Multi-modal integration.

1. Introduction

Multi-modal dialog systems appear to offer better interaction experience, as multi-modality seems to have fundamental advantages over unimodal interaction. However, there are few matching examples beyond the standard “put-that-there” scenario. Much more often, simply providing alternative input or output modalities resulting in sequential multi-modality seems to be the state-of-the-art. The question is what constitutes a “good” interaction, i.e. what aspects contribute to the user having a good or bad impression of the system she has been using. This is commonly understood by the term “Quality of Experience”, QoE.

In this chapter, we will summarize major results concerning multi-modality at first, and then provide a common ground on what Quality of Experience really means. We will then present experimental results from different interaction scenarios: Audio-visual transmission systems (like IP-based television or audio-visual telephony), interactions with Embodied Conversational Agents (ECAs), as well as interactions with different non-embodied multi-modal dialogue systems providing speech, touch and motion input capabilities. For each scenario, algorithmic models are presented which quantify the impact of each modality on the overall system quality, as it is perceived subjectively by the user. The goodness of the models are described in term of Pearson's correlation R between the models' estimates and the real data obtained, as well as the root mean squared prediction error (RMSE). We conclude by identifying some research questions which should be answered in order to fully support the design and evaluation of multi-modal dialog applications.

2. Advantages of Systems Providing Multi-Modal Interaction

One major assumption concerning human-computer interfaces is that the interaction is significantly facilitated by providing multiple input modalities and by presenting information over different output channels. From a usability point of view – i.e. discounting hedonic aspects like appearance and style of the interface or the possibility to express the user's identity with a given product – a multiple of possible input modalities can increase the recognition rate by fusing different input modalities (e.g. on the signal level) and it allows people to use those modalities most adequate in their specific situation, mood and capability (López-Cózar Delgado and Araki, 2005; Oviatt, 2004). For example, touch may be favoured in noisy or public environments, speech for the task of selecting objects in longer lists, typing for editing text and pointing gestures to refer to spatial information. Concerning the system output, multiple modalities allow for selecting the most appropriate way to present a specific piece of information (e.g. Graphical User Interfaces for lists, Embodied Conversational Agents for emotions, auditory icons for alarms, short vibrations for positive feedback). Another benefit is the possibility to present information redundantly to increase salience.

Furthermore, there seem to be cognitive advantages for multi-modal interfaces. Redundant and complementary information may distribute the use of cognitive resources and thus make processing faster and less demanding. With the theory of multiple resources (Wickens, 1999) for example, the tasks of speaking and gesturing or hearing and watching use different resources that in principle should not interfere with each other. As a result, users seem to prefer multi-modal interaction, especially, when the cognitive load increases

due to time pressure or task difficulty (cf. (Oviatt et al., 2004)). However, there are also examples which show that this benefit is not always observable, and the theoretical basis of a strict separation of the unimodal signals is questioned (cf. (Sarter, 1995)). Instead, multi-modality may even increase cognitive load (Schomaker et al., 1995) compared to single-modality usage.

As humans naturally interact with each other multi-modally – i.e. face-to-face communication with speech, non-speech sounds, gestures, expressions – an Embodied Conversational Agent used adequately as an interface to computers can increase user's experience of a system (cf. (Benoît et al., 2000)). Of course, this also holds for communication services enabling human-to-human interaction with more than one modality, commonly by providing audio-visual (AV) communication.

Certainly, multi-modal interfaces enable a new quality of human-to-human and human-to-machine interaction. To achieve this expected benefit in user experience, we have to know how users experience the interaction with such systems and services. In the following, relevant mechanisms are explained showing how users come to their judgments of system quality and how different modalities contribute to this. For three different scenarios, namely multi-modal signal transmission, Embodied Conversational Agents (ECAs), and non-embodied dialog systems, experimental results are summarized to derive simple algorithmic models of the integration processes for overall quality ratings of multi-modal systems. It will be pointed out which problems have to be dealt with and what steps have to be taken in order to really predict the quality users experience when interacting with multi-modal dialog systems.

2.1 Modality Relations

There are different approaches to formalize the relationship between different modalities during an interaction. Typically, there are two dimensions addressed: The *temporal assignment* (parallel vs. sequential multi-modality) and the *amount of information conveyed* with each modality (complementary vs. redundant). One of the most common systematics is described by the CARE properties (Coutaz et al., 1995). Apart from formal definitions of the name-giving four properties, the relationship between the multi-modal behavior of the user and the one of the system is discussed:

- **Complementarity:** Different modalities have to be used in order to reach the target.
- **Assignment:** Only one modality is selected, either by the system or the user.
- **Redundancy:** Different modalities are used, bearing comparable information, either in parallel or sequentially.

- **Equivalence:** Any available modality can be used. There are no restrictions on the temporal order.

Such formal descriptions of modality relations can be used to specify how a multi-modal system outputs information generically, or dependent on the user input. For specified tasks and user groups this formalization can also be used for evaluating the system's appropriateness in modality choices. But also face-to-face and thus human-to-ECA communication might be formalized by this account. It is not trivial to simulate human behavior with ECAs, as linguistic and non-linguistic information might be naturally redundant (e.g. mood is expressed with voice as well as facial expressions and posture), but information often conveyed complementarily (the famous "put-that-there" scenario).

3. Quality of Experience

Developers of multi-modal systems tend to highlight the performance of their system and the individual input and output modules in order to justify how good their system is. In this context, we can define "performance" as follows:

Performance: *The ability of a unit to provide the function it has been designed for (Möller, 2005).*

Easy-to-calculate performance figures are e.g. the recognition rates for speech or gesture recognizers, the intelligibility of TTS modules, or the conveyability of intended emotions by an ECA. A pre-defined set of performance figures can be used to characterize the so-called "Quality of Service", QoS. This term which is commonly used for media transmission services has been defined as follows:

Quality of Service (QoS): *The collective effect of service performance which determines the degree of satisfaction of the user of the service (ITU-T Rec. E.800, 1994). This includes service support, service operability, serveability, and service security.*

Although system performance (and thus QoS) will have a severe impact on user satisfaction, there is no one-to-one relationship between the two. User satisfaction is just one aspect of quality, i.e.:

Quality: *Result of appraisal of the perceived composition of the service with respect to its desired composition ((ITU-T Rec. P.851, 2003), following (Jekosch, 2004; Jekosch, 2005)).*

Apparently, quality requires a perception and a judgment process to take place inside the human user. Obviously, the result of this process is severely impacted by the system characteristics (and so system performance), but there are

other characteristics of the usage situation and context as well as user-internal factors (memory, expectation, etc.) which will decide on which level of quality the user finally attributes to the interaction with the system. As a corresponding concept to Quality of Service, the term “Quality of Experience” (QoE) is now in use to summarize the user perceptions resulting from the interaction with the system. Unfortunately, QoE is still ill-defined in the international bodies:

Quality of Experience (QoE): *The overall acceptability of an application or service, as perceived subjectively by the end user. Quality of Experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.) (ITU-T Rec. P.10, 2007).*

However, overall acceptability may be influenced by user expectations and context. A better definition emerged from discussions by the participants of the Dagstuhl Seminar 09192 “From Quality of Service to Quality of Experience” which was held in May 2009 in Dagstuhl, Germany:

Quality of Experience (QoE): *Degree of delight of the user of a service. In the context of communication services, it is influenced by content, network, device, application, user expectations and goals, and context of use.*

Service: *An event in which an entity takes the responsibility that something desirable happens on the behalf of another entity.*

Acceptability: *Characteristic of a service describing how readily a person will use the service. Acceptability is the outcome of a decision which is partially based on the Quality of Experience.*

In order to assess Quality of Experience, perception and judgment processes have to take place inside a human user. As a consequence, subjective evaluation methods are necessary in order to quantify the QoE which can be achieved with a particular multi-modal system. In (Möller et al., 2009), we have shown that QoE is a multidimensional construct, the components of which can be quantified with the help of dedicated questionnaires. [Table 1](#) is taken from (Möller et al., 2010) and summarizes some commonly used questionnaires which have been proved adequate to quantify sub-aspects of QoE, and which will be used in some of the studies cited and summarized hereafter.

4. Audio-Video Quality Integration in AV-Transmission Services

In the case of network services providing audio-visual signals like television, video-clips and especially AV-telephony, the perceived quality of the signals is one of the main factors to be assessed. Evaluating the visual and audio

Table 1. Comparison of questionnaires and captured QoE aspects^a. ●: completely captured; ◐: partially captured; ○: not captured.

Sub-scales	Questionnaire			
	SUS	AttrakDiff ¹	SUMI ²	SASSI ³
<i>Learnability</i>	●	◐(PQ)	●(LEA)	●(LIK, HAB)
<i>Effectiveness</i>	●	●(PQ)	◐(CON, HEL)	◐(ACC, HAB)
<i>Efficiency</i>	●	●(PQ)	●(EFF)	◐(SPE, CD)
<i>Intuitivity</i>	○	○	○	○
<i>Aesthetics</i>	○	●(HQ-S, ATT)	◐(AFF)	○
<i>System Personality</i>	○	◐(HQ-S)	○	◐(ANN, LIK)
<i>Appeal</i>	○	●(HQ-S, HQ-I)	◐(AFF, LIK)	◐(ANN, LIK)

^a Cf. (Möller et al., 2010) ©Elsevier 2010.

channel independently does not necessarily provide an insight into the quality experienced by the user. Instead, the mechanism of integrating both modalities during perception and appraisal has to be known in order to monitor and adjust the service.

4.1 Videotelephony

With a straight-forward approach, the perceived multi-modal quality (MOS_{AV}) is evaluated and modelled as a combination of the separate uni-modal quality ratings (MOS_A and MOS_V). Here, each rating is obtained on a 11-point Absolute Category Rating (ACR) scale as it is specified in (ITU-T Rec. P.920, 2000); then, Mean Option Scores (MOS) are derived for the audio, video and audio-visual quality of the transmission, by averaging the individual ratings over all users of the different test conditions. In this experiment, 24 subjects (aged 18–30 years) had to do the building block task (ITU-T Rec. P.920, 2000) and the short conversation test (Möller, 2000) via AV dialog. Three different simple relationships were tested:

$$MOS_{AV} = c_1 \cdot MOS_A + c_2 \cdot MOS_V + c_3, \quad (9.1)$$

$$MOS_{AV} = c_1 \cdot MOS_A \cdot MOS_V + c_2, \quad (9.2)$$

$$MOS_{AV} = c_1 \cdot MOS_A + c_2 \cdot MOS_V + c_3 \cdot MOS_A \cdot MOS_V + c_4. \quad (9.3)$$

With the second model (9.2) correlations between estimated and measured MOS_{AV} between $R = 0.93$ and $R = 0.99$ could be obtained.⁴ As shown in Table 2, models with an interaction term describe the perceptual integration better than simple linear models.

As expected, the visual channel contributes stronger to the multi-modal quality ratings than the auditory channel. Therefore, the correlation between

Table 2. Modelling audio-visual integration for videotelephony^a.

Model	$MOS_{AV} =$	Pearson's R	RMSE
Linear:	$0.677 + 0.217 \cdot MOS_A + 0.888 \cdot MOS_V$	0.96	0.53
Interaction:	$1.3 + 1.1 \cdot MOS_A \cdot MOS_V$	0.99	0.95
Complete:	$0.517 + 0.0058 \cdot MOS_A + 0.654 \cdot MOS_V +$ $0.042 \cdot MOS_A \cdot MOS_V$	0.97	0.57

^a Cf. (Belmudez et al., 2009) ©IEEE 2009.

MOS_V and MOS_{AV} is higher than between MOS_A and MOS_{AV} (see Figure 1). Within these models, the variance of MOS_{AV} is basically determined by the variance of MOS_V alone. Notably, the impact of audio quality increases with that of the video. Apparently, MOS_V comes first, but with better video quality, there is a perceptual saturation effect, and audio quality gets more important for the test participants. However, the exact weighting depends on the type of task and the degree of interactivity (passive test from literature vs. dialog). In the short conversation test, the audio plays a crucial role to fulfill the task: Quality in conditions with bad audio quality is rated significantly worse than in the building block task with comparable conditions. These are two very important context effects, that are of strong influence in all multi-modal interaction scenarios, as shown in the next sections.

With all these models presented, there is always the problem of collecting valid data from the test participants: Ideally, unimodal ratings should be assessed separately from the multi-modal condition. However, in some of the scenarios presented, this is not practical or even impossible (e.g. rating visual quality for the short conversation test or articulating ECAs). Most importantly, rating scales are not used in a linear way: For example, there are saturation effects of the scale itself, and categories of the ACR-scale used in the experiments presented do not match the idea of continuous quality ascription. With a more linear scale the models might benefit from a better description of the ratings obtained.

4.2 IP-Television

In the case of quality of IP-based TV quality assessment, results were first transformed to a so-called “perceptual scale” (R-scale, cf. (ITU-T Rec. G.107, 2005)) which is used as a basis for transmission planning models by the International Telecommunication Union, ITU-T. This scale is thought to avoid some of the non-linearities of the ACR scales used in the experiment. In a study by Garcia and Raake (2009), two different modelling approaches were evaluated. Both models estimate AV-quality (Q_{AV}) on the R-scale: On the one hand using single modality quality ratings as in the data of Belmudez et al. (2009) (quality

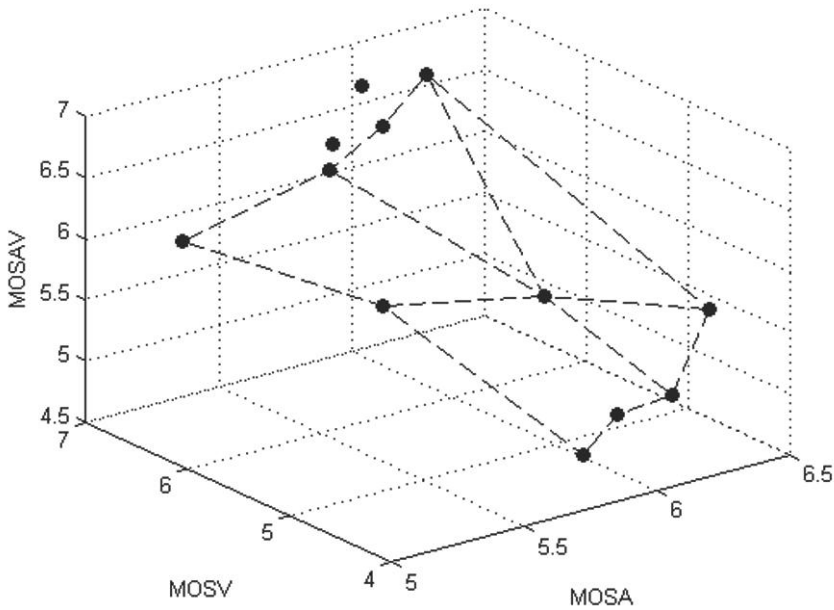


Figure 1. AV quality integration in video transmission services (from (Belmudez et al., 2009)), ©IEEE 2009.

based approach with the complete model, see Equation 9.4), and on the other hand an estimation of audio-visual quality based on impairments factors (impairment factor based approach, see Equation 9.5). For their data, impairment factor have been estimated from the subjective ratings, not from parametric descriptions of the transmission (e.g. packet loss). MOS are obtained from 24 different subjects (aged 21–44) for each of the three conditions: Audio-only, video-only and audio-visual. Both approaches show comparable results (see Figure 2).

$$Q_{AV} = 27.805 + 0 \cdot Q_A + 0.129 \cdot Q_V + 0.006 \cdot Q_A \cdot Q_V \quad (9.4)$$

The quality-based model correlates with the subjects' ratings with $R = 0.96$ ($RMSE = 3.38$):

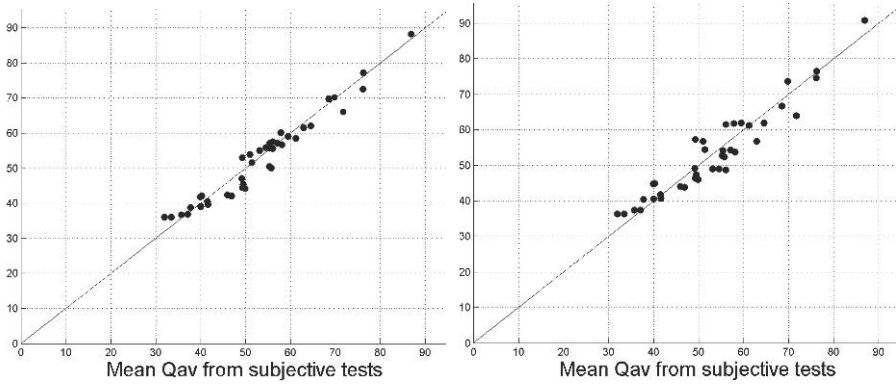


Figure 2. AV quality integration in IPTV. left: Quality-based approach, right: Impairment factor based approach (from (Garcia and Raake, 2009)), ©IEEE 2009.

$$\begin{aligned}
 Q_{AV} = & 88.195 - 0.379 \cdot Icod_A - 0.588 \cdot Icod_V \\
 & - 0.625 \cdot Itra_A - 0.625 \cdot Itra_V \\
 & + 0.005 \cdot Icod_A \cdot Icod_V \\
 & + 0.007 \cdot Itra_A \cdot Itra_V \\
 & + 0.011 \cdot Icod_V \cdot Itra_A \\
 & + 0.007 \cdot Icod_A \cdot Itra_V.
 \end{aligned} \tag{9.5}$$

The impairment-based model performs slightly better ($R = 0.98$, $RMSE = 2.57$).

5. Quality of Embodied Conversational Agents

Dialogue systems with Embodied Conversational Agents are frequently represented by 3D modelled animated humans heads. Other realizations span from abstract icons (e.g. “smiley” faces) to animals, cartoons or fictional creatures. Concerning realistic human appearances, there are also visual models of the full body and upper part of the body in use. Apart from application in virtual realities, such an ECA can offer a number of benefits to a dialog system, including:

- intuitively display emotions and feedback (e.g. system state is idle, concentration on one of several user, system is busy);
- display of facial expressions or gestures for paralinguistic and linguistic usage;
- supporting the user to concentrate on the human-computer interface;

- increased robustness in speech perception (with lip-synchronous ECAs, e.g. in noise);
- general *Persona Effect* of better subjective ratings (cf. (Dehn and Van Mulken, 2000) for a meta-analysis and summary).

In a series of experiments, audio-visual quality of different talking heads was evaluated and modelled from single modality ratings of speech quality and visual quality (Weiss et al., 2010). In this case, different text-to-speech and head modules were used (see Figure 3 for pictures of the three talking heads tested). Transmission quality is not in scope of this research. Therefore AV quality of the heads presented on a display were comparable concerning codec and frame-rate. Instead, the perceived user experience of the talking head component was assessed as basis for the usability in their specific application. Subjects in the experiments rated the ECAs on several scales to assess various quality aspects.

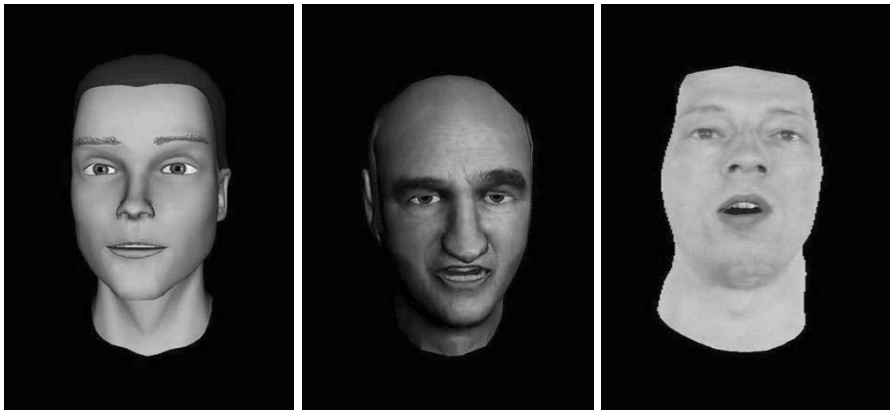


Figure 3. Three facial models tested (cf. (Kühnel et al., 2008)).

As a result of the experiments, talking heads overall quality (MOS_{heads}) could be described as a linear combination of visual MOS_V and speech quality rating MOS_A . However, the models do not perform comparable to AV quality assessed in IPTV or IP-videotelephony scenarios. The data sets obtained and described by the models are not truly comparable due to differences in stimuli. However, the most important results can be extracted from the models presented in Table 3. There were four different conditions: the passive rating test (14 subjects, aged 20–32), a simulated interaction (23 subjects, aged 21–60), a simulated interaction with an information screen in addition to the ECA screen (23 subjects, aged 20–57), and a real interaction with the system (49 subjects,

aged 20–61): With an increase in interactivity and an increase in distraction, module differences get blurred and the models' fit decreases. The distraction of the rating process was introduced by a second screen (the first displays the ECA). This additional screen presented information from the system as lists, whereas the degree of interactivity merely refers to the difference between a passive rating test versus a simulated interaction experiment. The last model is from a real interaction experiment which also included the second screen (cf. (Kühnel et al., 2009)).

Table 3. Modelling audio-visual quality of talking heads for passive, simulated (one and two screens), and real interaction scenarios (2 screens)^a.

<i>type of experiment</i>	<i>overallquality_{heads} =</i>	<i>Pearson's R</i>	<i>RMSE</i>
Passive:	$0.47 + 0.51 \cdot MOS_A + 0.33 \cdot MOS_V$	0.83	0.49
Simulated (1 screen):	$0.16 + 0.42 \cdot MOS_A + 0.30 \cdot MOS_V$	0.71	0.59
Simulated (2 screens):	$0.35 + 0.36 \cdot MOS_A + 0.23 \cdot MOS_V$	0.57	0.65
Interaction:	$0.30 + 0.26 \cdot MOS_A + 0.40 \cdot MOS_V$	0.57	0.62

^a Cf. (Weiss et al., 2010) and (Kühnel et al., 2009).

From the questionnaires used it became clear, that at least for the data obtained during interaction, ECA quality cannot be sufficiently equated with auditory and visual quality. Other factors have also an significant impact on overall quality of talking heads: I.e. overall system quality, how entertaining the embodiment is due to non-linguistic movements, naturalness of the ECA, as well as perceived goodness of synchronization (cf. (Weiss et al., 2009)).

6. Quality of Systems with Multiple Input Modalities

The last scenario presented here are multi-modal dialog systems which can be controlled by different input modalities. Like current commercial systems, the experimental setup does allow to change the input modality sequentially, the appropriate CARE property is *Equivalence* (see Chapter 2.1). The questions addressed here are two-fold:

- 1 *Which modality is preferred by the user? How consistent is the individual modality usage?*
- 2 *Is a multi-modal interface better than an unimodal one? For systems providing Equivalence, can the multi-modal systems quality be modelled by ratings of the unimodal interfaces?*

Two experiments studying these issues are presented in the following: One interface is attached in an office area and one is a mobile device.

6.1 Smart Office

The so called *Attentive Display* is a room information system installed at T-Labs, Berlin. It provides information on the colleagues currently present and their desk and room bookings. Additionally, you can be informed about events (lectures, meetings). The interface is a big screen, fixed in the entrance area. It can be operated by touch and/or speech (that is enabled automatically when the camera tracks a face, thus the name “attentive display”). The output is always visual.

In the first experiment, there were three blocks: Touch only, speech only, and the multi-modal session always at the end of the test (cf. (Wechsung et al., 2009b) for the full description and results). User experience ratings were assessed with the AttrakDiff questionnaire (Hassenzahl et al., 2003), that covers hedonic and pragmatic aspects of the users perception (36 subjects, aged 21–39). For the three conditions, differences in the ratings on the hedonic and pragmatic scales were observed: The ranking concerning the pragmatic quality was touch over multi-modal over speech. This means, there was no benefit of providing speech in addition to touch. As subjects were explicitly asked to use the system multi-modally in the last session, the speech usage lead to lower ratings. However, the results are different for the hedonic scales: Here, multi-modal interaction was rated best. Please note, that the *Pragmatic* scale can be interpret as indicator of functionality and usability, whereas the global scale *Attractiveness* is related to user experience! Additionally, there is overall quality, which is the mean of all items used.

Concerning the integration of the quality ascribed with different input modalities, the multi-modal quality can be described as linear combination of the single modality ratings (see Table 4). As you can see, the fit of the models is far better for the overall quality and the *Attractiveness* scale. Mostly, touch is more important than speech – especially for *Attractiveness*, except for the *Identity* scale.

Table 4. Integration of perceived quality aspects of speech (Q_S) and touch (Q_T) to multi-modal ratings (Q_{MM} , ordered last)^a.

Scale	$Q_{MM} =$	Pearson's R	RMSE
Overall:	$0.14 + 0.81 \cdot Q_T + 0.68 \cdot Q_S$	0.91	0.35
Attractiveness:	$-0.20 + 0.85 \cdot Q_T + 0.48 \cdot Q_S$	0.92	0.41
Pragmatic:	$0.22 + 0.80 \cdot Q_T + 0.47 \cdot Q_S$	0.79	0.67
Stimulation:	$0.11 + 0.69 \cdot Q_T + 0.63 \cdot Q_S$	0.83	0.51
Identity:	$0.38 + 0.28 \cdot Q_T + 0.66 \cdot Q_S$	0.78	0.53

^a Cf. (Wechsung et al., 2009b).

The multi-modal condition was always presented last in order to have the subjects become familiar with both modalities before using them together. A possible explanation for the observed strong correlation between the linear combination of both single modality ratings and the multi-modal condition could be that the subjects tried to rate consistently. To verify the results obtained, a second study was conducted with the multi-modal condition always at the first position (cf. (Wechsung et al., 2009a)). For this experiment the models extracted are significantly lower in power and stability (leave on out cross-validation, 18 subjects, aged 22–30). See Table 5 for the results. The *Pragmatic* and *Stimulation* scales are not included, as an estimation on basis of the single modality ratings was not possible.

Table 5. Integration of perceived quality aspects of speech (Q_S) and touch (Q_T) to multi-modal ratings (Q_{MM} , ordered first)^a.

Scale	$Q_{MM} =$	Pearson's R	RMSE
Overall:	$0.18 + 0.679 \cdot Q_T + 0.553 \cdot Q_S$	0.76	0.55
Attractiveness:	$0.29 + 0.653 \cdot Q_T + 0.545 \cdot Q_S$	0.77	0.73
Identity:	$0.06 + 0.664 \cdot Q_T + 0.485 \cdot Q_S$	0.87	0.41

^a Cf. (Wechsung et al., 2009a).

6.2 Mobile

The application tested is a multi-modal information-box (e-mail, SMS, fax), that runs on a smart-phone (cf. (Wechsung et al., 2009a)). In addition to touch and speech, there is a motion input modality to navigate and select with tilting the whole device. The system's output is generally *assigned* to visual output. Additionally, there is context dependent *Redundancy* (cf. Section 2.1 for the CARE properties) for speech input (audio and visual): This is vibration as positive feedback for the motion modality and audio feedback to signal *match* and *nomatch* for voice input. The procedure is similar to the first experiment presented in the last section (30 subjects, two age groups: 25–29, 55–66): The multi-modal condition is always last. Findings include the relevance of the frequency of each modality used in the multi-modal condition: Motion is not included in the regression models (7% usage), and speech (19%) only for *Stimulation* (see Table 6). Combinations of modalities were only used infrequently (6%).

With a leave one out cross-validation the scale *Pragmatic* was identified as being unstable.

Table 6. Integration of perceived quality aspects of speech (Q_S), touch (Q_T) and motion (not significant) to multi-modal ratings (Q_{MM} , ordered last)^a.

Scale	$Q_{MM} =$	Pearson's R	RMSE
Overall:	$0.16 + 0.69 \cdot Q_T$	0.69	0.48
Attractiveness:	$0.04 + 0.79 \cdot Q_T$	0.56	0.68
Stimulation:	$0.31 + 0.60 \cdot Q_T + 0.35 \cdot Q_S$	0.86	0.40
Identity:	$0.22 + 0.75 \cdot Q_T$	0.69	0.45
Pragmatic:	$0.41 + 0.49 \cdot Q_T$	0.36	0.77

^a Cf. (Wechsung et al., 2009a).

6.3 Summary

With the three experiments presented here, it could be shown that perceived quality aspects – including pragmatic and hedonic aspects – of multi-modal interaction could be described as linear combination of ratings for single modalities. Results are satisfying for overall quality and *Attractiveness*. Apparently participants are better in mentally “adding” than in “subtracting” modality ratings during evaluation, as subtracting one’s own ratings from memory is more demanding (Kamii et al., 2001). This interpretation is supported by the finding, that older users – who often have decreased working memory capacity – multi-modal ratings are less good predicted. Interestingly, this became especially obvious for the *Pragmatic* scale.

While multi-modal conditions did not perform better than the best unimodal condition for the *Pragmatic* scale, on hedonic scales the quality did benefit from multi-modality. The amount of modality usage affects the weights of the single modalities.

7. Conclusions

Multi-modal communication systems can be found in a great variety of application scenarios. We presented evaluation experiments from fields of IP based audio-video transmission for TV and videotelephony, Embodied Conversational Agents for smart-home environments and stationary and mobile non-embodied multi-modal user interfaces. We showed how to assess perceived quality and user experience of such systems. Our results show that quality of multi-modal systems comprises a multitude of aspects – depending on the application – and is influenced by the measurement process.

For the case of audio-visual integration in video transmission applications, visual quality mostly has a much stronger influence on overall quality than audio/speech quality. Stable models with sufficient power can be derived for AV quality on the basis of single modalities’ quality. However, the exact weightings depend on interactivity.

For audio-visual integration in ECA applications, interactivity also plays an important role: The degree of interactivity determines the impact of animation and speech on overall quality of the animated agent, but definitely other factors affect the ECAs overall quality as well, like the smoothness of interaction and other representations of the system. For example, additional information nicely presented by the system improved the ratings of the ECA.

Multi-modal quality and attractiveness of multi-modal interactive systems can be estimated on the basis of judgments for unimodal conditions. Complementary multi-modality (“put-that-there” scenario) was not tested, but are considered not common in commercial interactive systems. Weightings for overall quality reflect modality usage to a certain extent. Interestingly, weightings for hedonic qualities are also influenced by less-used modalities.

In all cases, the quality user are experiencing was assessed by questionnaires. To find models predicting perceived quality is difficult indeed: What kind of constructs (quality aspects) are relevant in the specific case and how are they assessed best? The AttrakDiff has shown great potential to cover many important aspects in a valid and reliable way for interactive systems. In the case of IP based transmission applications a continuously scale is recommended. But currently only some important factors have been identified to be included into the models or at least to be controlled in the experiments. The order of presentation of modalities and degree of interactivity are stated in this text, but of course the progress of interaction and topic of transmitted signals are relevant, too. There is a bunch of open questions, regarding this topic:

- If modality weightings are influenced by modality usage, what does influence actual modality usage?
- What is the impact of modality effectiveness and efficiency?
- For interactive systems, what is the impact of output modalities for the usage of input modalities and the multi-modal quality judgment?
- What type of model (linear, multiplicative, other nonlinear) is most adequate for multi-modal quality prediction?
- For which (input and output) modalities does such modeling work well?

In all scenarios presented, weighted combinations of ratings for single modalities (either in unimodal conditions or as separate ratings for multi-modal conditions) could be used to describe multi-modal quality of experience. Apparently, estimating multi-modal quality works best for transmission quality and can be used for prediction already. For this case, it seems, there are not as many possibly influencing factors as in dialog system, as those include the interactive part when rated. In Human Computer Interaction scenarios, findings are obtained revealing fundamental mechanisms influencing the judgment

process for multi-modal interaction. In both cases, however, we have far to go to model any real cognitive processes.

Acknowledgments

We would like to thank Benjamin Belmudez, Marie-Neige Garcia and Alexander Raake for providing their data and figures on IPTV and videotelephony.

This work was partly supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant MO 1038/6-1.

Notes

1. Sub-scales: *Attractiveness* (ATT), *Hedonic Qualities – Identity* (HQ-I), *Hedonic Qualities – Stimulation* (HQ-S), *Pragmatic Qualities* (PQ).
2. Sub-scales: *Affect* (AFF), *Control* (CON), *Efficiency* (EFF), *Learnability* (LEA), *Helpfulness* (HEL). SUMI is generally not recommended for evaluating multi-modal systems.
3. Sub-scales: *System Response Accuracy* (ACC), *Annoyance* (ANN), *Cognitive Demand* (CD), *Habitability* (HAB), *Likeability* (LIK), *Speed* (SPE).
4. With constant factors of $c_1 = 0.107 \dots 0.121$ and $c_2 = 1.1 \dots 1.5$

References

- Belmudez, B., Möller, S., Lewcio, B., Raake, A., and Mehmood, A. (2009). Audio and Video Channel Impact on Perceived Audio-Visual Quality in Different Interactive Contexts. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP'09)*.
- Benoît, C., Martin, J.-C., Pelachaud, C., Schomaker, L., and Suhm, B. (2000). Audio-Visual and Multimodal Speech-Based Systems. In Gibbon, D., Mertins, I., and Moore, R. K., editors, *Handbook of Multimodal and Spoken Dialogue Systems*, pages 102–203. Kluwer Academic Publ., Boston MA.
- Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., and Young, R. (1995). Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties. In Nordby, K., Helmersen, P., Gilmore, D., and Arnesen, S., editors, *Human-Computer Interaction, Interact '95*, pages 115–120. Chapman & Hall, London.
- Dehn, D. M. and Van Mulken, S. (2000). The Impact of Animated Interface Agents: a Review of Empirical Research. *International Journal of Human-Computer Studies*, 52(1):1–22.
- Garcia, M. and Raake, A. (2009). Impairment-Factor-based Audio-Visual Quality Model for IPTV. In *Proceedings of the 1st International Workshop on Quality of Multimedia Experience (QoMEX'09)*.
- Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer

- Qualität. In Ziegler, J. and Szwillus, G., editors, *Interaktion in Bewegung. Proc. Mensch & Computer '03*, pages 187–196, Stuttgart. B.G. Teubner.
- ITU-T Rec. E.800 (1994). Terms and Definitions Related to Quality of Service and Network Performance Including Dependability. International Telecommunication Union, Geneva.
- ITU-T Rec. G.107 (2005). *The E-model, a Computational Model for Use in Transmission Planning*. International Telecommunication Union, Geneva.
- ITU-T Rec. P.10 (2007). Vocabulary for Performance and Quality of Service. International Telecommunication Union, Geneva.
- ITU-T Rec. P.851 (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.
- ITU-T Rec. P.920 (2000). *Interactive Test Methods for Audiovisual Communication*. International Telecommunication Union, Geneva.
- Jekosch, U. (2004). Basic Concepts and Terms of "Quality", Reconsidered in the Context of Product-Sound Quality. *Acta Acustica united with Acustica*, 90(6):999–1006.
- Jekosch, U. (2005). *Voice and Speech Quality Perception. Assessment and Evaluation*. Springer, Berlin.
- Kamii, C., Lewis, B. A., and Kirkland, L. D. (2001). Subtraction Compared with Addition. *Mathematical Behavior*, 20:33–42.
- Kühnel, C., Weiss, B., and Möller, S. (2009). Talking Heads for Interacting with Spoken Dialog Smart-Home Systems. In *Proceedings of the 10th Ann. Conference of the Int. Speech Communication Assoc. (Interspeech '09)*, pages 304–307.
- Kühnel, C., Weiss, B., Wechsung, I., Fagel, S., and Möller, S. (2008). Evaluating Talking Heads for Smart Home Systems. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'08)*.
- López-Cózar Delgado, R. and Araki, M. (2005). *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. John Wiley & Sons, Chichester.
- Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, Boston.
- Möller, S. (2005). *Quality of Telephone-based Spoken Dialogue Systems*. Springer, New York.
- Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., and Weiss, B. (2009). A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction. In *Proceedings of the 1st International Workshop on Quality of Multimedia Experience (QoMEX'09)*.
- Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., and Weiss, B. (2010). Evaluation of Multimodal Interfaces for Ambient Intelligence. In Aghajan,

- H., López-Cózar Delgado, R., and Augusto, J. C., editors, *Human-Centric Interfaces for Ambient Intelligence*, pages 347–370. Elsevier, Amsterdam.
- Oviatt, S. (2004). Multimodal Interfaces. In Sears, A. and Jacko, J., editors, *The Human Computer Interaction Handbook*, pages 413–432. Lawrence Erlbaum, New York, 2 edition.
- Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 129–136.
- Sarter, N. (1995). Multiple-Resource Theory as a Basis for Multimodal Interface Design: Success Stories, Qualifications, and Research Needs. In Kramer, A., Wiegmann, D., and Kirlik, A., editors, *Attention: From Theory to Practice*, pages 187–195. Oxford University Press.
- Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., and Blauert, J. (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System*. NICI, Nijmegen.
- Wechsung, I., Engelbrecht, K.-P., Nauman, A., Schaffer, S., Seebode, J., Metze, F., and Möller, S. (2009a). Predicting the Quality of Multimodal Systems Based on Judgements of Single Modalities. In *Proceedings of the 10th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech '09)*, pages 1827–1830.
- Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F., and Möller, S. (2009b). Usability-Evaluation multimodaler Schnittstellen: Ist das Ganze die Summe seiner Teile? In Kain, S. and Struve, D., editors, *Grenzenlos frei. Proc. Mensch & Computer '09*, pages 495–498. Oldenbourg Wissenschaftsverlag.
- Weiss, B., Kühnel, C., Wechsung, I., Möller, S., and Fagel, S. (2009). Comparison of Different Talking Heads in Non-Interactive Settings. In *Proceedings of Human Computer Interaction International (HCII), San Diego*, pages 349–357.
- Weiss, B. and Kühnel, C., Wechsung, I., Fagel, S., and Möller, S. (2010). Quality of Talking Heads in Different Interaction and Media Contexts. *Speech Communication*, 52(6):481–492.
- Wickens, C. (1999). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science*, 3:159–177.

Chapter 10

DIALOGUE ACTS ANNOTATION TO CONSTRUCT DIALOGUE SYSTEMS FOR CONSULTING

Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, and
Satoshi Nakamura

Kyoto, Japan

{ kiyonori.ohtake, teruhisa.misu, chiori.hori, hideki.kashioka, satoshi.nakamura }@nict.go.jp

Abstract This chapter introduces a new corpus of consulting dialogues designed for training a dialogue manager that can handle consulting dialogues through spontaneous interactions from the tagged dialogue corpus. We have collected more than 150 hours of consulting dialogues in the tourist guidance domain. This chapter outlines our taxonomy of dialogue act (DA) annotation that can describe two aspects of an utterance: the communicative function (speech act (SA)), and the semantic content of the utterance. We provide an overview of the Kyoto tour guide dialogue corpus and a preliminary analysis using the DA tags. We also show a result of a preliminary experiment for SA tagging via Support Vector Machines (SVMs). In addition, we mention the usage of our corpus for the spoken dialogue system that is being developed.

Keywords: Corpus; Dialogue act tagging.

1. Introduction

This chapter introduces a new dialogue corpus for consulting in the tourist guidance domain. The corpus consists of speech, transcripts, speech act tags, morphological analysis results, dependency analysis results, and semantic content tags. In this chapter, we describe the current status of a dialogue corpus that is being developed by our research group, focusing on two types of tags: speech act (SA) tags and semantic content tags. These SA and semantic content tags have been designed to express the dialogue act (DA) of each utterance.

Many studies have focused on developing spoken dialogue systems. Their typical task domains included the retrieval of information from databases or

making reservations, such as airline information, e.g. Defense Advanced Research Projects Agency (DARPA) Communicator (Walker et al., 2001), and train information, e.g. Automatic Railway Information Systems for Europe (ARISE) shown by Bouwman et al. (1999) and Multi-modal-Multimedia Automated Service Kiosk (MASK) by Lamel et al. (2002). Most studies assumed a definite and consistent user objective, and the dialogue strategy was usually designed to minimize the cost of information access. Other target tasks include tutoring and trouble-shooting dialogues (Boye, 2007). In such tasks, dialogue scenarios or agendas are usually described using a (dynamic) tree structure, and the objective is to satisfy all requirements.

In this chapter, we introduce our corpus, which is being developed as part of a project to construct consulting dialogue systems, that helps the user in making a decision. Thus far, several projects have been organized to construct speech corpora such as the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000). The size of CSJ is very big, and a large part of the corpus consists of monologues. Although, CSJ includes some dialogues, the size of the dialogues is not enough to construct a dialogue system via recent statistical techniques. In addition, as compared to consulting dialogues, the existing large dialogue corpora covered very clear tasks in limited domains.

However, consulting is a frequently used and very natural form of human interaction. We often consult with a sales clerk while shopping or with staff at a concierge desk in a hotel. Such dialogues usually form part of a series of information retrieval dialogues that have been investigated in many previous studies. They also contain various exchanges, such as clarifications and explanations. The user may explain his/her preferences vaguely by listing examples. The server would then sense the user's preferences from his/her utterances, provide some information, and then request a decision.

It is almost impossible to handcraft a scenario that can handle such spontaneous consulting dialogues; thus, the dialogue strategy should be bootstrapped from a dialogue corpus. If an extensive dialogue corpus is available, we can model the dialogue using machine learning techniques such as partially observable Markov decision processes (POMDPs) (Thomson et al., 2008). Hori et al. (2008) have also proposed an efficient approach to organize a dialogue system using weighted finite-state transducers (WFSTs); the system obtains the structure of the transducers and the weight for each state transition from an annotated corpus. Thus, the corpus must be sufficiently rich in information to describe the consulting dialogue to construct the statistical dialogue manager via such techniques.

In addition, a detailed description would be preferable when developing modules that focus on spoken language understanding and generation modules. In this study, we adopt DAs (Bunt, 2000; Shriberg et al., 2004; Bangalore

et al., 2006; Rodriguez et al., 2007; Levin et al., 2002) for this information and annotate DAs in the corpus.

In this chapter, we describe the design of the Kyoto tour guide dialogue corpus in Section 2. Our design of the DA annotation is described in Section 3. Sections 4 and 5 respectively describe two types of tag sets, namely, the SA tag and the semantic content tag. Section 6 describe the usage of the Kyoto tour guide dialogue corpus to construct our spoken dialogue system.

2. Kyoto Tour Guide Dialogue Corpus

We are currently developing a dialogue corpus with tourist guidance for Kyoto City as the target domain. Thus far, we have collected itinerary planning dialogues in Japanese, in which users plan a one-day visit to Kyoto City. There are three types of dialogues in the corpus: face-to-face (F2F), Wizard of OZ (WOZ), and telephonic (TEL) dialogues. The corpus consists of 114 face-to-face dialogues, 80 dialogues using the WOZ system, and 102 dialogues obtained from telephone conversations with the interface of the WOZ system. [Figures 1](#) and [2](#) show the snapshots of the recordings for F2F and TEL dialogues.



Figure 1. Recording of F2F dialogue.

The overview of these three types of dialogues is shown in [Table 1](#). Each dialogue lasts for almost 30 minutes. Almost all the dialogues have been man-



Figure 2. Recording of TEL dialogue.

ually transcribed. Table 1 also shows the average number of utterances per dialogue.

Table 1. Overview of Kyoto tour guide dialogue corpus.

Dialogue type	F2F (ja)	WOZ (ja)	TEL (ja)	F2F (en)
Number of dialogues	114	80	102	48
Number of guides	3	2	2	1
Average number of utterances per dialogue (guide)	365.4	165.2	—	—
Average number of utterances per dialogue (tourists)	301.7	112.9	—	—

Each face-to-face dialogue involved a professional tour guide and a tourist. Three guides, one male and two females, were employed to collect the dialogues. All three guides were involved in almost the same number of dialogues. The guides used maps, guidebooks, and a PC connected to the internet.

In the WOZ dialogues, two female guides were employed. Each of them participated in 40 dialogues. The WOZ system consists of two Internet browsers, a speech synthesis program, and an integration program for the collaborative work. Collaboration was required because in addition to the guide, operators were employed to operate the WOZ system and support the guide. The guide and the operators had their own individual computers that were connected to each other; further, they collaboratively operated the WOZ system to serve the user (tourist). Figure 3 shows the interface of the WOZ system.

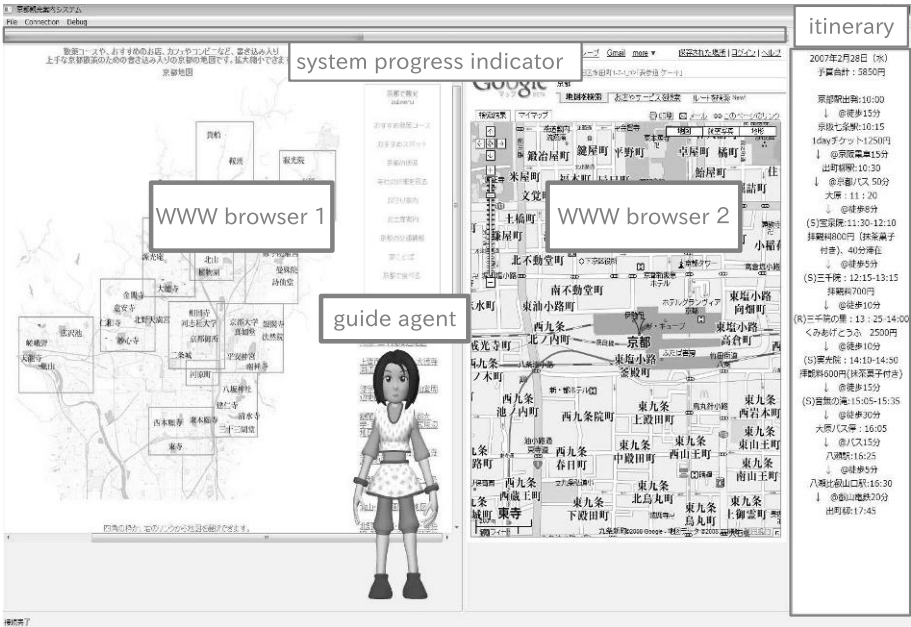


Figure 3. WOZ system interface.

In the telephone dialogues, the same two female guides as for the WOZ dialogues were employed. In these dialogues, we used the WOZ system, but we did not need the speech synthesis program. The guide and a tourist shared the same interface in different rooms, and they could talk to each other through the hands-free headset.

Dialogues to plan a one-day visit consist of several conversations for choosing the places to visit. The conversations usually included sequences of requests from the users and provision of information by the guides as well as consultation in the form of explanation and evaluation. It should be noted that in this study, unlike information kiosk systems such as those developed in (Lamel et al., 2002) or (Thomson et al., 2008), enabling the user to access

information is not an objective in itself. The objective is similar to the problem-solving dialogue of the study by Ferguson and Allen (1998); in other words, accessing information is just an aspect of consulting dialogues.

An example of dialogue via face-to-face communication is shown in [Table 2](#). This dialogue is part of a consultation to decide on a sightseeing spot to visit. The user asks the location of a spot, and the guide answers it. Then, the user provides a follow-up by evaluating the answer. The task is challenging because there are many utterances that affect the flow of the dialogue during a consultation. The utterances are listed in the order of their start times with the utterance ids (UID). From the column ‘Time’ in the table, it is easy to see that there are many overlaps.

3. Annotation of Communicative Function and Semantic Content in DA

We annotate DAs in the corpus to describe a user’s intention and a system’s (or the tour guide’s) action. Recently, several studies have addressed multilevel annotation of dialogues (Bangalore et al., 2006; Rodriguez et al., 2007; Levin et al., 2002); in our study, we focus on the two aspects of a DA indicated by Bunt (2000). One is the communicative function that corresponds to how the content should be used to update the context, and the other is a semantic content that corresponds to what the act is about. We consider both as important information to handle the consulting dialogue.

We designed two different tag sets to annotate DAs in the corpus. The SA tag is used to capture the communicative functions of an utterance using domain-independent multiple function layers. The semantic content tag is used to describe the semantic content of an utterance using domain-specific hierarchical semantic classes.

4. SA Tags

In this section, we introduce the SA tag set that describes the communicative functions of the utterances.

4.1 Annotation Unit

There have been numerous discussions on the base unit of an SA annotation. As the simplest base unit, we can use a sentence or an utterance. However, sentence boundaries are not necessarily obvious in human-human dialogues. In addition, a long sentence tends to contain multiple dialogue functions. Thus, it is desirable to define a short unit so that the tags can elaborate the utterance. In addition, if the SA tag is used as an input of a dialogue system, the unit should be detected automatically (not manually). Therefore, we apply

Table 2. Example dialogue from the Kyoto tour guide dialogue corpus.

UID	Time (ms)	Speaker	Transcript
56	76669–78819	User	<i>Ato</i> (and) <i>Ohara ga</i> (Ohara) <i>dono heN ni</i> (whereabouts) <i>narimasuka</i> (be?) (Where is Ohara?)
57	80788–81358	Guide	<i>kono</i> (here) <i>heN desune</i> (around be) (Around here.)
58	81358–81841	Guide	<i>Ohara wa</i> (Ohara)
59	81386–82736	User	<i>Chotto</i> (a bit) <i>hanaresugite masune</i> (be too far) (Ohara seems to be too far from Kyoto st.)
60	83116–83316	Guide	<i>A</i> (ah)
61	83136–85023	User	<i>kore demo</i> (it) <i>ichinichi dewa</i> (one day) <i>doudeshou</i> (how about?) (Can I do Ohara in a day?)
62	83386–84396	Guide	<i>Soudesune</i> (let me see)
63	85206–87076	Guide	<i>Ichinichi</i> (one day) <i>areba</i> (if be) <i>jubuN</i> (enough) <i>ikemasu</i> (can go) (One day is enough to visit Ohara.)
64	88392–90072	Guide	<i>Oharamo</i> (Ohara) <i>sugoku</i> (very) <i>kireidesuyo</i> (be a beautiful) (Ohara is a very beautiful place.)
65	89889–90759	User	<i>Iidesune</i> (sounds nice)

the clause boundary annotation program (CBAP) (Kashioka and Maruyama, 2004) to the transcript of the dialogue session, and adopt a clause as the base unit of tag annotation. Thus, in the following discussions, ‘utterance’ denotes a clause. We have already tagged more than 55 dialogues with SA tags. Roughly speaking, one dialogue consists of one thousand utterances.

4.2 Tag Specifications

There are two major policies in SA annotation. One is to select exactly one label from the tag set (e.g., the Augmented Multi-party Interaction (AMI) corpus¹). The other is to annotate with as many labels as required. Meeting Recorder Dialog Act (MRDA) (Shriberg et al., 2004) and Dynamic Interpretation Theory (DIT) and DIT++ (Bunt, 2000) are defined on the basis of the second policy. We believe that the utterances are generally multifunctional and this multifunctionality is an important aspect for managing consulting dialogues through spontaneous interactions. Therefore, we have adopted the latter policy.

By extending the MRDA tag set and DIT++, we defined our SA tag set that consists of six layers to describe six groups of function: *General*, *Response*, *Check*, *Constrain*, *ActionDiscussion*, and *Others*. A list of the tag sets excluding *Others* layer is shown in Table 3. The *General* layer has two sublayers under the labels *Pause* and *WH-Question*, respectively. The two sublayers are used to elaborate on the two labels, respectively. A tag of the *General* layer must be labelled to an utterance, but the other layer's tags are optional; in other words, layers other than *General* can take null values when there is no tag that is appropriate to the utterance. In practical annotation, the most appropriate tag is selected from each layer, without taking into account any of the other layers.

The descriptions of the layers are as follows:

4.2.1 General Layer. Each tag of this layer represents the basic form of the unit. Most of the tags in this layer are used to describe forward-looking functions. The tags are classified into three large groups: 'Question,' 'Fragment,' and 'Statement.' The tag 'Statement===' denotes the continuation of the utterance. The following are the tags of the General layer.

Statement, Pause, Backchannel, Y/N-Question, WH-Question, OR-Question, OR-segment-after-Y/N, Open-Question

In the *General* layer, there are two sublayers for the labels: *Pause* and *WH-Question*. The *Pause* sublayer consists of Hold, Grabber, Holder, and Releaser. The *WH* sublayer labels the WH-Question type.

4.2.2 Response Layer. The tags of this layer denote the responses directed to a specific previous utterance made by the addressee. The following are the tags of the Response layer.

Answer, Acknowledgment, Accept, PartialAccept, AffirmativeAnswer, Reject, PartialReject, NegativeAnswer

Table 3. List of SA tags and their occurrence in the experiment.

Tag	Percentage (%)		Tag	Percentage (%)	
	User	Guide		User	Guide
(General layer)			(Response layer)		
Statement	45.25	44.53	Acknowledgment	19.13	5.45
Pause	12.99	15.05	Accept	4.68	6.25
Backchannel	26.05	9.09	PartialAccept	0.02	0.10
Y/N-Question	3.61	2.19	AffirmativeAnswer	0.08	0.20
WH-Question	1.13	0.40	Reject	0.25	0.11
Open-Question	0.32	0.32	PartialReject	0.04	0.03
OR-after-Y/N	0.05	0.02	NegativeAnswer	0.10	0.10
OR-Question	0.05	0.03	Answer	1.16	2.57
Statement==	9.91	27.79			
(ActionDiscussion layer)			(Check layer)		
Opinion	0.52	2.12	RepetitionRequest	0.07	0.03
Wish	1.23	0.05	UnderstandingCheck	0.19	0.20
Request	0.22	0.19	DoubleCheck	0.36	0.15
Suggestion	0.16	1.12	ApprovalRequest	2.01	1.07
Commitment	1.15	0.29			
(Constrain layer)					
Reason	0.64	2.52			
Condition	0.61	3.09			
Elaboration	0.28	4.00			
Evaluation	1.35	2.01			

4.2.3 Check Layer. The tags of this layer denote the confirmation of a certain expected response. The following are the tags of the Check layer.

RepetitionRequest, DoubleCheck, UnderstandingCheck, ApprovalRequest

4.2.4 Constrain Layer. The tags of this layer denote the functions to restrict or complement the target of the utterance. The following are the tags of the Constrain layer.

Reason, Condition, Elaboration, Evaluation

4.2.5 Action Discussion Layer. The tags of this layer mark the functions of the utterances that pertain to a future action. The following are the tags of the Action Discussion layer.

Wish, Opinion, Suggestion, Request, Commitment

4.2.6 Others Layer. The tags of this layer describe various functions of the utterance, e.g. Greeting, SelfTalk, Welcome, Apology, etc. The following are the tags of the Others layer.

Greeting, Introduction, Thank, Apology, Welcome, SelfRepair, Correct, CollaborativeComplementation, SelfTalk, Repeat, Mimic, Maybe, Inversion

It should be noted that this taxonomy is intended to be used for training spoken dialogue systems. Consequently, it contains detailed descriptions to elaborate on the decision-making process. For example, checks are classified into four categories because they should be treated in various ways in a dialogue system. *UnderstandingCheck* is often used to describe clarifications; thus, it should be taken into account when creating a dialogue scenario. In contrast, *RepetitionRequest*, which is used to request that the missed portions of the previous utterance be repeated, is not concerned with the overall dialogue flow.

An example of an annotation is shown in Table 4. Since the *Response* and *Constrain* layers are not necessarily directed to the immediately preceding utterance, the target utterance ID is specified. The interface for the annotation of SA tags is shown in Figure 4.

Table 4. Example of SA annotation for the data shown in Table 2.

UID	SA tag
56	WH-Question_Where
57	State_Answer→56
58	State_Inversion
59	State_Evaluation→57
60	Pause_Grabber
61	Y/N-Question
62	State_Acknowledgment→59
63	State_AffirmativeAnswer→61
64	State_Opinion
65	State_Acknowledgment→64_Evaluation→64

Tags are concatenated by a delimiter ‘_’ and omitting the null values. The number following the ‘→’ denotes the target utterance of the function.

4.3 Evaluation of the Annotation

We performed a preliminary annotation of the SA tags in the F2F corpus. Thirty dialogues (900 minutes; 23,169 utterances) were annotated by three labellers. When annotating the dialogues, we took into account textual infor-

	A	B	C	D	G	H	I	J	K
I	start	end	rol	TEXT	UID	General	Response	Target(res)	Check
53	72645	76039	u	なんとなくは嵯峨野とか。	520	State			
54	73124	73461	g	はい。	530	Backchannel			
55	75759	76165	g	はい。	540	State	Acknowledg	520	
56	76615	78725	u	あと、大原がどの辺になりますか。	550	WH^Where			
57	78725	79815	u	地図があんまり言うほど	OR				
58	79815	80895	u	頭に入っていない。	OR(after Y/N)				
59	80788	81368	g	この辺ですね。	Open				
60	81368	81841	g	大原は。	WH^Who		answer	550	
61	81386	82736	u	ちょっと離れすぎてますね。	WH^When				
62	83116	83316	g	あ、	WH^Where				
63	83136	83356	u	これ。	WH^What/Which				
64	83356	85041	u	でも、一日ではどうでしょう？	WH^Why				
65	83386	84396	g	そうですね。	610	Pause	Grabber		
66	84556	84976	g	ええ。	620	State			
67	85206	87076	g	一日あれば充分行けます。	630	WH^How			
68	86527	87393	u	行けますか。	640	Pause^Hold			
69	87296	87781	g	はい。	650	Pause^Hold			
70	87812	88822	u	うーん。	660	State	Answer	630	
71	88822	89822	u	うーん。	670	Y/N			
72	89822	90822	u	うーん。	680	State	Accept	670	
73	90822	91822	u	うーん。	690	State			

Figure 4. Example of an interface for the annotation of SA tags.

mation, audio information, and contextual information. The result was cross-checked by another labeller.

4.3.1 Distributional Statistics. The frequencies of the tags, expressed in percentages, are shown in Table 3. In the General layer, nearly half of the utterances were *Statement*. This bias is acceptable because 66% of the utterances that are tagged as *Statement* had tag(s) of other layers.

The percentages of the tags in the *Constrain* layer are relatively higher than those of the tags in the *ActionDiscussion* and *Check* layers. They are also higher than the corresponding percentage figures for MRDA (Shriberg et al., 2004) and SWBD-DAMSL (Jurafsky et al., 1997).

These statistics characterize the consulting dialogue of sightseeing planning, where elaborations and evaluations play an important role during the decision process.

4.3.2 Inter-Annotator Agreement. We investigated the inter-annotator agreement for SA tags. Three labellers were employed to make six annotated dialogues from two dialogues (2,087 utterances). Each dialogue was annotated by the three labellers and the agreement among them was examined. These results are listed in Table 5. The agreement ratio is the average of all the combinations of the three individual agreements. In the same way, we also computed the average Kappa statistic, which is often used to measure the agreement by considering the chance rate.

A high concordance rate was obtained for the *General* layer. When the specific layers and sublayers are taken into account, The Kappa statistic was

Table 5. Agreement among the labellers.

	General layer	All layers
Agreement ratio	86.7%	74.2%
Kappa statistic	0.74	0.68

0.68, which is considered a good result for this type of task (e.g. Shriberg et al. (2004)).

4.3.3 Analysis of the Occurrence Tendency during the Progress of the Episode. We then investigated the tendencies of tag occurrence through a dialogue to clarify how consulting is conducted in the corpus. We annotated the boundaries of the episodes that determined the spots to visit to carefully investigate the structure of the decision-making processes. In our corpus, users were asked to write down their itinerary for a practical one-day tour. Thus, the beginning and ending of an episode can be determined on the basis of this itinerary.

As a result, we found 192 episodes. We selected 122 episodes that had more than 50 utterances, and analyzed the tendency of tag occurrence. The episodes were divided into five segments so that each segment had an equal number of utterances. An example of the tendencies of tag occurrence is shown in Figure 5. The relative occurrence rate is obtained by dividing the number of times the tags appeared in each segment by the total number of occurrences throughout the dialogues.

We found three patterns in the tendencies of occurrence. The tags corresponding to the first pattern frequently appear in the early part of an episode; this typically applies to Open-Question, WH-Question, and Wish. Figure 6 shows the result of this pattern.

The tags of the second pattern frequently appear in the later part; this typically applies to Evaluation, Commitment, and Opinion. Figure 7 shows the result of this pattern.

The tags of the third pattern appear uniformly over an episode; this typically applies to Y/N-Question, Accept, and Elaboration. Figure 8 shows the result of this pattern.

These statistics characterize the dialogue flow of sightseeing planning, where the guide and the user first clarify the latter's interests (Open, WH-Questions) and then list and evaluate candidates (Evaluation), following which the user takes a decision (Commitment).

This progression indicates that the management of a session or a dialogue phase requires wide contextual information within an episode to manage the consulting dialogue, even though the test-set perplexity², which was calculated

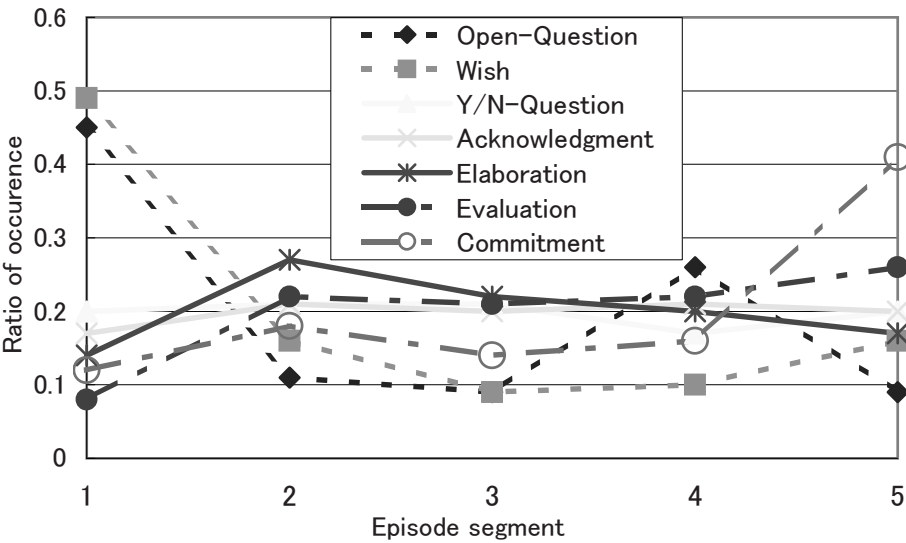


Figure 5. Progress of episodes vs. the occurrence of SA tags.

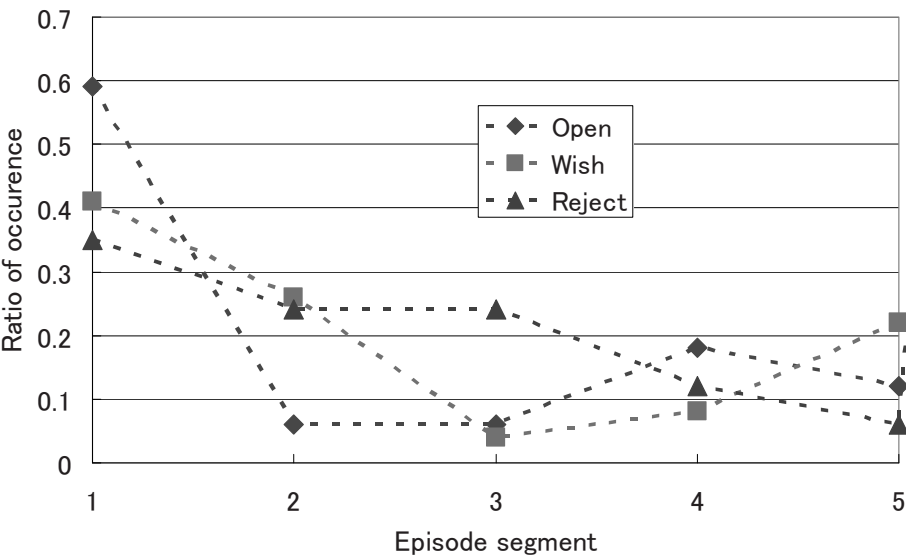


Figure 6. Tendency of SA tags: peak in the beginning.

by a 3-gram language model trained with the SA tags, was not high (4.25 using the general layer and 14.75 using all layers).

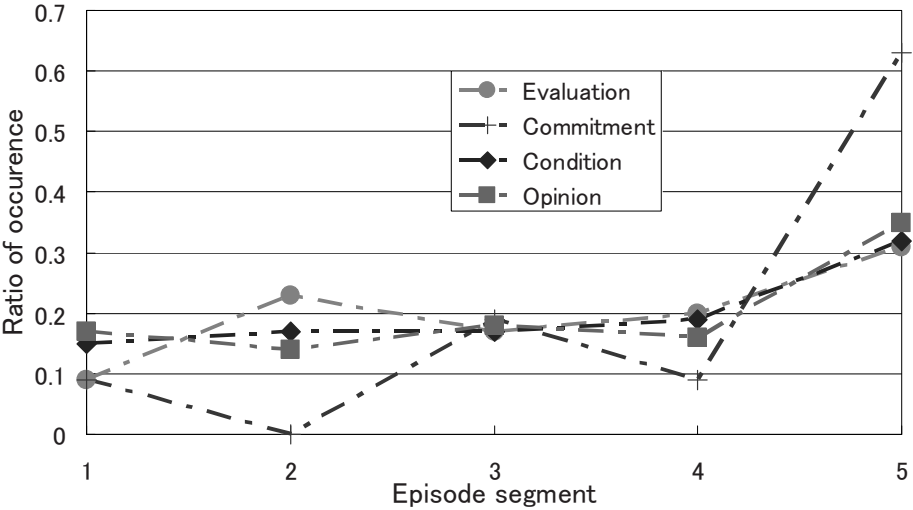


Figure 7. Tendency of SA tags: peak in the end.

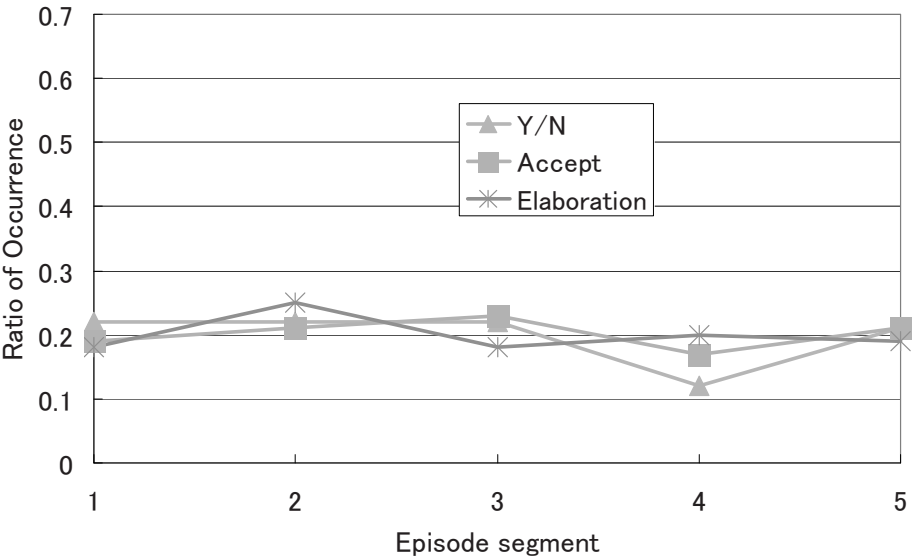


Figure 8. Tendency of SA tags: stable.

4.4 Preliminary Experiment to Estimate SA Tags via SVM

We also carried out a preliminary experiment to estimate the SA tags. The SA tagged corpus is being developed and the corpus may not be clean. However, we tried to construct an SA tagger via SVMs.

We can see SA tagging as a sequential labelling problem. We prepared 36 dialogues of the F2F corpus with SA tags, in which we used 34 dialogues as learning data and two dialogues as test data. We construct a classifier using only the labels of the General layer. The learning data and the test data include 16 and 13 labels of the *General* layer.

The features used to construct a classifier are as follows: the role of the speaker, length of the utterance (second), barge-in flag, last three morphemes of the utterance, etc. The feature vector for the label of utterance u_i is extracted from $u_{i-4}, u_{i-3}, \dots, u_i, u_{i+1}, u_{i+2}$. It should be noted that we tried to construct an off-line tagger to support the human annotation. When we use this tagger in a practical dialogue system to estimate the SA of user's utterance, the feature vector can be extracted from $u_{i-4}, u_{i-3}, \dots, u_i$. The kernel function of SVM is a 2nd-degree polynomial function. To achieve a multi-class classifier via SVM, we constructed the SVMs by the pairwise method. The accuracy of our first trial was 73.02%. We have to consider feature extraction to improve the accuracy. We will try to use the features of all sorts.

The SA tagged corpus should be brushed up, because the agreement ratio between human labellers as shown in Table 5 does not reach 90% for the general layer. In other words, the maximum accuracy is estimated at around 86%. From these numbers, the 73% accuracy of our first try seems very promising.

5. Semantic Content Tags

The semantic content tag set was designed to capture the content of an utterance. Some might consider semantic representations by HPSG (Pollard and Sag, 1994) or LFG (Dalrymple et al., 1994) for an utterance. Such frameworks require knowledge of grammar and experiences to describe the meaning of an utterance. In addition, the utterances in a dialogue are often fragmentary, which makes the description more difficult.

We focused on the dependency relations between two words to capture the semantic relations of the words. Annotating dependency relations is more intuitive and is easier than annotating the syntactic structure; moreover, a dependency parser is more robust for fragmentary expressions than syntax parsers.

We introduced semantic classes to represent the semantic content of an utterance. Semantic class labels are applied to each unit of the dependency structure of an utterance. The task that identifies the semantic classes is very similar to named entity recognition, because the classes of the named entities can be equated to the semantic classes that are used to express semantic content. However, both nouns and predicates are very important for capturing the semantic content of an utterance. For example, '10 a.m.' might denote the current time in the context of planning, or it might signify the opening time of a sightseeing spot. Thus, we represent the semantic content on the basis of the dependency

structure. Each element of a dependency structure is assigned a semantic category.

For example, the sentence “I would like to see Kinkakuji temple.” is annotated as shown in Figure 9. In this figure, the semantic content tag *(preference).action* indicates that the predicate portion expresses the speaker’s *preference* for the speaker’s action, while the semantic content tag *(preference).(spot).name* indicates the *name* of the *spot* as the object of the speaker’s *preference*.

Given sentence

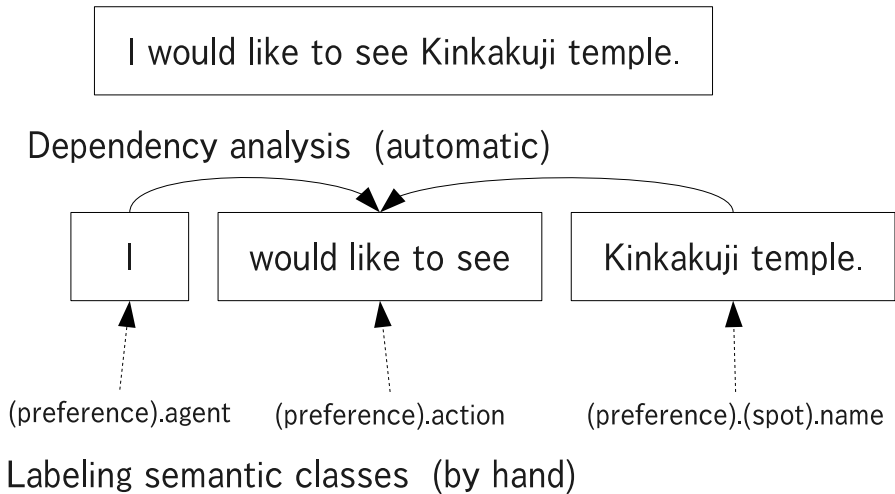


Figure 9. Example of an annotation with semantic content tags.

Although we do not define the semantic role (e.g. object (*Kinkakuji temple*) and subject (*I*)) of each argument item in this case, we can use conventional semantic role labelling techniques (Gildea and Jurafsky, 2002) to estimate them.

5.1 Tag Specifications

We defined the hierarchical semantic classes to annotate the semantic content tags. There are 33 labels (classes) at the top hierarchical level. The labels include **activity**, **event**, **meal**, **spot**, **transportation**, **cost**, **consulting**, and **location**, and are shown in Figure 10. There are two kinds of labels, nodes, and leaves. A node must have at least one child, a node, or a leaf. A leaf has no children. The number of types for nodes is 47, and the number of types for leaves is 47. The labels of the leaves are very similar to the labels for named entity recognition. For example, there are ‘year’, ‘date’, ‘time’, ‘organizer’,

‘name’, etc. in the labels of the leaves.

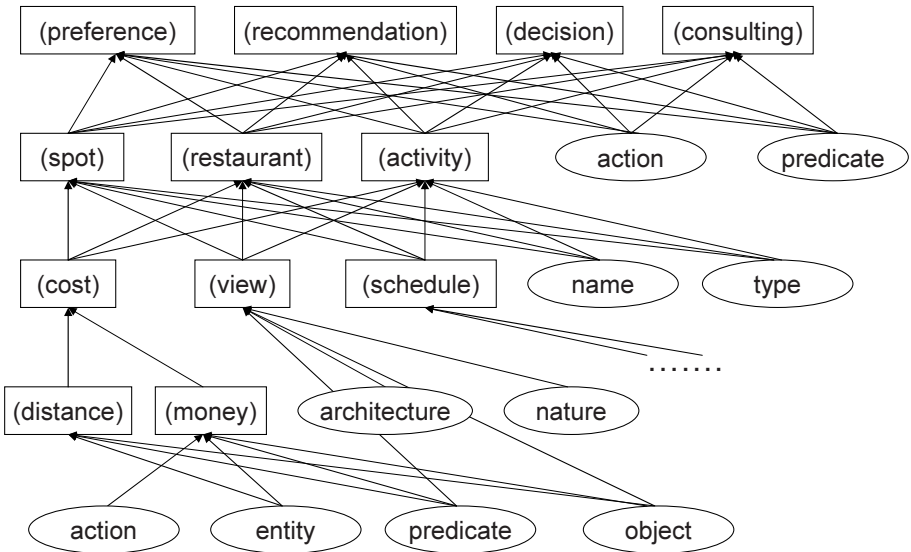


Figure 10. Part of the semantic category hierarchy.

One of the characteristics of the hierarchical structure of the semantic classes is that the lower level structures are shared by many upper nodes. Thus, the lower level structure can be used in any other domain or target task.

5.2 Annotation of Semantic Content Tags

The annotation of semantic content tags is performed in the following four steps. First, an utterance is analyzed by a morphological analyzer, ChaSen³. Second, the morphemes are chunked into the dependency unit (*bunsetsu*). Third, a dependency analysis is performed using a Japanese dependency parser, CaboCha⁴. Finally, we annotate the semantic content tags for each *bunsetsu* unit using our annotation tool. An example of an annotation is shown in [Table 6](#). Each row in the column ‘Transcript’ denotes the divided *bunsetsu* units.

The annotation tool interface is shown in [Figure 11](#). In the left side of this figure, the dialogue files and each utterance of the dialogue information are displayed. The dependency structure of an utterance is displayed in the upper part of the figure. The morphological analysis results and chunk information are displayed in the lower part of the figure.

Table 6. Example of semantic content tag annotation for the data shown in Table 2.

UID	Transcript	Semantic content tag
56	<i>Ato</i> (and)	null
	<i>Ohara ga</i> (Ohara)	(activity),location
	<i>dono heN ni</i> (whereabouts)	(activity),(demonst),interr
	<i>narimasuka</i> (be?)	(activity),predicate
57	<i>kono</i> (here)	(demonst),koso
	<i>heN desune</i> (around be)	(demonst),noun
58	<i>Ohara wa</i> (Ohara)	location
59	<i>Chotto</i> (a bit)	(trsp),(cost),(distance),adverb-phrase
	<i>hanaresugite masune</i>	(trsp),(cost),(distance),predicate
	(be too far)	
60	<i>A</i> (ah)	null
61	<i>kore demo</i> (it)	null
	<i>ichinichi dewa</i> (one day)	(activity),(planning),duration
	<i>doudeshou</i> (how about?)	(activity),(planning),(demonst),interr
62	<i>Soudesune</i> (let me see)	null
63	<i>Ichinichi</i> (one day)	(activity),(planning),(entity),day-window
	<i>areba</i> (if be)	(activity),(planning),predicate
	<i>jubuN</i> (enough)	(consulting),(activity),adverb-phrase
	<i>ikemasu</i> (can go)	(consulting),(activity),action
64	<i>Oharamo</i> (Ohara is)	(recommend),(activity),location
	<i>sugoku</i> (very)	(recommend),(activity),adverb-phrase
	<i>kireidesuyo</i> (beautiful)	(recommend),(activity),predicate
65	<i>Iidesune</i> (sounds nice)	(consulting),(activity),predicate

Moreover, there is another window that is used to select a semantic class for the annotation tool of the semantic content tag. This window is shown in Figure 12.

At present, the annotations of semantic content tags are being carried out for 40 dialogues. Approximately 26,800 paths, including paths that will not be used, exist if the layered structure is fully expanded. In the 40 dialogues, 1,980 tags (or paths) are used.

In addition, not only the annotation of semantic content tags but also the correction of the morphological analysis and dependency analysis results is being carried out. If we complete the annotation, we will also obtain the correctly tagged data of the Kyoto tour guide corpus. These corpora can be used to

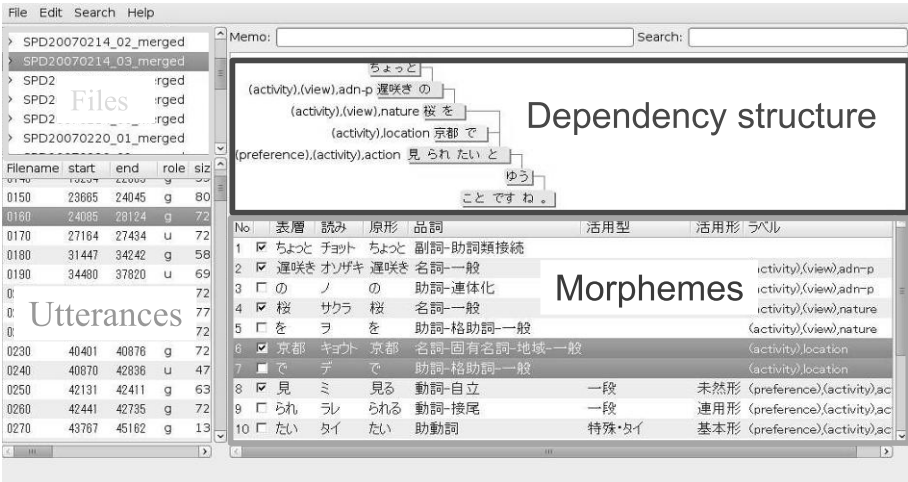


Figure 11. Annotation tool interface for annotating semantic content tags.

develop analyzers such as morphological analyzers and dependency analyzers via machine learning techniques or to adapt the analyzers for this domain.

6. Usage of the Kyoto Tour Guide Corpus

In this section, we discuss the usage of the Kyoto tour guide corpus. We can see that a dialogue system consists of a speech recognition module, a dialogue management module, a speech synthesis module, and a database for target domain. Recently, most of those modules have been based on statistical methods that require corpora. The relationship between a dialogue system and a dialogue corpus is shown in Figure 13.

6.1 Speech Recognition

We constructed the language model that is used in the speech recognition module of our dialogue system. To construct the language model, the morphological analysis results of the dialogue corpus were used. It is required that the domain specific n-gram entries are included in the language model to achieve high performance for speech recognition. Only maintaining the recognition dictionaries does not lead us to the satisfactory recognition results.

6.2 Dialogue Management

One of the most significant roles of the dialogue model in a spoken language dialogue system seems to appropriately represent a contextual interpretation of

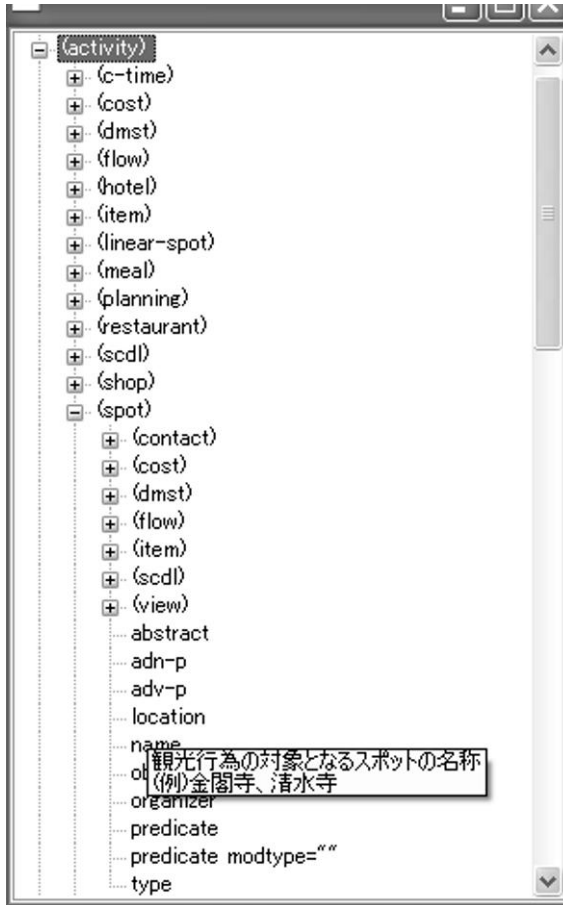


Figure 12. Window for semantic content tag hierarchy.

the user utterances. This allows the system to generate the most adequate system response without limiting the dialogue to a succession of questions and answers. This role should also enable the system to anticipate/predict, raise ambiguities, correct errors, explain system decisions, and trigger the corresponding actions throughout the dialogue to suitably manage other processing modules.

We are now adopting the corpus annotated with the DA tags to construct a dialogue system using WFSTs as dialogue management modules. To achieve dialogue management via WFSTs, we have to prepare not only the DA tags but also the tags for the system's action. As such, we are now preparing such action tags to construct a dialogue management module using WFSTs.

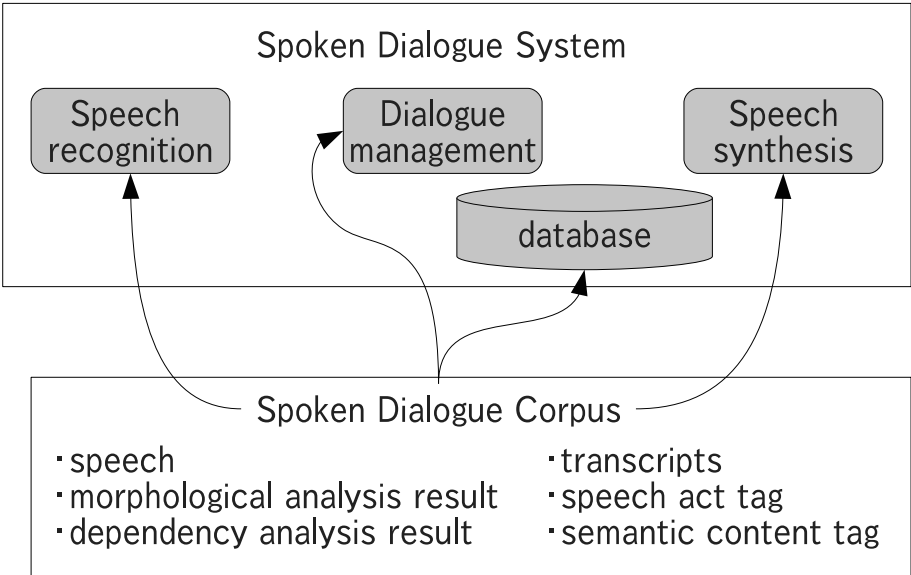


Figure 13. Relationship between a dialogue system and a dialogue corpus.

In addition, the corpus consists of real conversions between the guide and the travellers. Important and valuable information is buried in the corpus. If we apply data-mining techniques to the corpus, we will obtain much valuable information for travelling in Kyoto city and we can store this information in the database of the spoken dialogue system.

6.3 Speech Synthesis

Recent speech synthesis techniques such as concatenative synthesis or statistical parametric synthesis require large speech corpora. We can use conventional speech synthesis modules for a spoken dialogue system and the performance of the module as a text-to-speech module seems very high. However, we want to construct a more natural speech synthesis module that is suitable for a spoken dialogue system. Most of the conventional speech synthesis modules make only one speech from the text. In other words, it is hard to synthesize different speeches from the same text.

We have corpora with speech act tags, and we want to use this information to synthesize different speeches from the same text. In Japanese, “*hai* (yes)” is used in many ways, such as acknowledgment, back-channel, etc. We are now constructing speech synthesis modules using our dialogue corpus via two approaches. One is by constructing a speech synthesis system that directly

uses the recorded speech data of the guide. The other one is by constructing a speech synthesis system that uses a new speech corpus recorded with voice actors/actresses. For these recordings, we prepared the scripts from the transcripts of the corpus.

7. Conclusions

In this chapter, we have introduced our spoken dialogue corpus for developing consulting dialogue systems. We designed a DA annotation scheme that describes two aspects of a DA: SA and semantic content. The SA tag set was designed by extending the MRDA tag set. The design of the semantic content tag set is almost complete. If we complete the annotation, we will obtain the SA tags and the semantic content tags, as well as manual transcripts, morphological analysis results, dependency analysis results, and dialogue episodes. As a preliminary analysis, we have evaluated the SA tag set in terms of the agreement between labellers and investigated the patterns of tag occurrences. In addition, we tried to construct an SA tagger via SVMs as a first step to use the tagged corpus and the result was promising. We also mentioned the corpus usage in the development of our spoken dialogue system.

Next, we will investigate the features for SA tagging and semantic content tagging. We will construct a tagger for SA tags and semantic content tags using the annotated corpora and machine learning techniques. Our future work also includes the condensation or selection of DAs that directly affect the dialogue flow to construct a consulting dialogue system using the DA tags as an input.

Notes

1. <http://corpus.amiproject.org>
2. The perplexity was calculated by a 10-fold cross validation of the 30 dialogues.
3. <http://sourceforge.jp/projects/chasen-legacy/>
4. <http://chasen.org/~taku/software/cabocha/>

References

- Bangalore, S., Fabbrizio, G. D., and Stent, A. (2006). Learning the Structure of Task-Driven Human-Human Dialogs. In *Proceedings of COLING/ACL*, pages 201–208.
- Bouwman, G., Sturm, J., and Boves, L. (1999). Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, pages 493–496.
- Boye, J. (2007). Dialogue Management for Automatic Troubleshooting and Other Problem-Solving Applications. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 247–255.

- Bunt, H. (2000). Dialogue Pragmatics and Context Specification. In Bunt, H. and Black, W., editors, *Abduction, Belief and Context in Dialogue*, pages 81–150. John Benjamins.
- Dalrymple, M., Kaplan, R. M., III, J. T. M., and Zaenen, A., editors (1994). *Formal Issues in Lexical-Functional Grammar*. CSLI Publications.
- Ferguson, G. and Allen, J. F. (1998). TRIPS: An Intelligent Integrated Problem-Solving Assistant. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 567–573.
- Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.
- Hori, C., Ohtake, K., Misu, T., Kashioka, H., and Nakamura, S. (2008). Dialog Management using Weighted Finite-State Transducers. In *Proceedings of Interspeech*, pages 211–214.
- Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical report, University of Colorado at Boulder & SRI International.
- Kashioka, H. and Maruyama, T. (2004). Segmentation of Semantic Unit in Japanese Monologue. In *Proceedings of ICSLT-O-COCOSDA*.
- Lamel, L. F., Bennacef, S., Gauvain, J.-L., Dartigues, H., and Temem, J. N. (2002). User Evaluation of the MASK Kiosk. *Speech Communication*, 38(1):131–139.
- Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., Pianta, E., Cattoni, R., and Mana, N. (2002). Balancing Expressiveness and Simplicity in an Interlingua for Task Based Dialogue. In *Proceedings of ACL 2002 Workshop on Speech-to-speech Translation: Algorithms and Systems*.
- Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous Speech Corpus of Japanese. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2000)*, pages 947–952.
- Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.
- Rodriguez, K. J., Dipper, S., Götze, M., Poesio, M., Riccardi, G., Raymond, C., and Rabiega-Wisniewska, J. (2007). Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus. In *Proceedings of Linguistic Annotation Workshop*, pages 148–155.
- Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.
- Thomson, B., Schatzmann, J., and Young, S. (2008). Bayesian Update of Dialogue State for Robust Dialogue Systems. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Walker, M. A., Passonneau, R., and Boland, J. E. (2001). Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialogue Sys-

tems. In *Proceedings of the 39th Annual Meeting of the ACL*, pages 515–522.

Chapter 11

ON THE USE OF N-GRAM TRANSDUCERS FOR DIALOGUE ANNOTATION

Vicent Tamarit, Carlos-D. Martínez-Hinarejos, and José-Miguel Benedí
Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
Valencia, Spain

{vtamarit,cmartine,jbenedi}@dsic.upv.es

Abstract The implementation of dialogue systems is one of the most interesting applications of language technologies. Statistical models can be used in this implementation, allowing for a more flexible approach than when using rules defined by a human expert. However, statistical models require large amounts of dialogues annotated with dialogue-function labels (usually Dialogue Acts), and the annotation process is hard and time-consuming. Consequently, the use of other statistical models to obtain faster annotations is really interesting for the development of dialogue systems. In this work we compare two statistical models for dialogue annotation, a more classical Hidden Markov Model (HMM) based model and the new N-gram Transducers (NGT) model. This comparison is performed on two corpora of different nature, the well-known SwitchBoard corpus and the DIHANA corpus. The results show that the NGT model produces a much more accurate annotation than the HMM-based model (even 11% less error in the SwitchBoard corpus).

Keywords: Statistical models; Dialogue annotation.

1. Introduction

In the last few decades, the advances in speech technologies and natural language processing techniques have led to many natural-language-based solutions for several tasks. One of the most challenging examples of those solutions are dialogue systems.

A dialogue system is usually defined as a computer system that interacts with a human user to fulfil a task (Dybkjær and Minker, 2008). These systems

are of particular interest in many applications, like information systems that are accessed by telephone (Seneff and Polifroni, 2000; Aust et al., 1995) or assistant systems for people with special necessities (Wilks, 2006). Tasks such as ticket reservation or timetable consultation have usually been considered appropriate for these systems.

The speech act framework is the discourse theory in which many authors try to model the structure of dialogue. This theory (Austin, 1962) focuses on communicative acts performed through speech. Partially based on this theoretical approach, specific solutions have been proposed to model discourse in dialogue problems using a wide range of methods (dialogue grammars (McTear et al., 2000), information state (Bos et al., 2003), reinforcement learning (Williams and Young, 2007), etc.). Based on these solutions, a dialogue strategy can be defined. This dialogue strategy defines the way the system reacts to user inputs.

In any case, the strategies are based on the interpretation of the user input in terms of dialogue semantic units. These semantic units are usually coded in terms of Dialogue Acts (DA) (Bunt, 1994), which model the intention of the current user interaction along with its associated information. This concept can be extended to system responses. In an interaction, several sequences with a function from the viewpoint of the dialogue can be distinguished. These sequences are called segments (or utterances according to authors such as (Stolcke et al., 2000)), and each segment has associated only one DA label.

Data-based approaches to dialogue modelling such as (Stolcke et al., 2000) and (Young, 2000) have been developed in the last decade. These machine learning approaches rely on statistical models that can be automatically estimated from annotated data, which in this case, are dialogues from the task (knowledge domain). As a simplification, each situation in the dialogue can be associated with a specific label, and the models learn how to identify and react to the different situations by estimating the associations between the labels and the dialogue events (words, previous turns, etc.). Therefore, annotation schemes based on DA definitions must be defined to annotate the dialogues. Examples of DA annotation schemes developed in several projects are DAMSL (Core and Allen, 1997), *VerbMobil* (Alexandersson et al., 1998), DATE (Walker and Passonneau, 2001)), and DIHANA (Alcácer et al., 2005))

Consequently, the annotation of a dialogue corpus in terms of DA is an interesting problem for both the development of data-based dialogue systems and the study of discourse and dialogue structure. In the first case, the statistical models that implement the dialogue manager (Williams and Young, 2007; Meng et al., 2003; Stolcke et al., 2000) rely on annotated dialogues to estimate their parameters. This annotation process is developed by human experts and it is a hard and time-consuming task. The use of probabilistic models can provide a draft annotation of the corpus (Stolcke et al., 2000) that can make the manual annotation process faster.

Most of the previous works on the use of probabilistic models for DA annotation use segmented dialogue turns (Stolcke et al., 2000; Webb and Wilks, 2005; Rangarajan et al., 2007). However, this segmentation is not usually available in the initial transcription of a dialogue corpus. Other works propose a decoupled segmentation-annotation scheme (Ang et al., 2005). However, the ideal option is the use of models that can annotate unsegmented dialogue turns, giving the correct segments and labels. This option has been explored in a few previous works (Zimmermann et al., 2005; Martínez-Hinarejos et al., 2008), giving in any case (as could be expected) poorer results than when the segmentation is available.

The classical model for this task is based on Hidden Markov Models (HMM) (Stolcke et al., 2000). It offers good results when working with segmented dialogues but shows a dramatical decrease in performance when used in unsegmented dialogues to obtain their annotation and segmentation. An alternative model is the N-Gram Transducer (NGT) model, whose latest version was presented in (Martínez-Hinarejos et al., 2009). To improve the time and spatial complexity of the NGT model, some modifications can be performed on the search process, such as those presented in (Tamarit et al., 2009).

In this chapter we review the HMM-based model and the NGT model in depth, and we present exhaustive experiments to evaluate the performance (in both annotation accuracy and time) in two different corpora with very different features (SwitchBoard (Godfrey et al., 1992) and DIHANA (Benedí et al., 2006)). These results confirm the appropriateness of the NGT model for dialogue annotation.

This chapter is organised as follows: in Section 2 we present the baseline statistical model, which is based on HMM and N-grams, in Section 3 we define the NGT model as the new statistical model for dialogue annotation, in Section 4 we detail the corpora for the experiments, in Section 5 we describe the experiments and show their results, in Section 6 we draw conclusions and future lines of work.

2. The HMM-based Annotation Model

In this section we present the baseline statistical annotation model. This model is based on the combination of HMM and N-grams. Consequently, we will refer to it as the HMM-based model. This model is oriented to solve the optimisation problem of, given a word sequence $\mathcal{W}$ that represents a dialogue, obtaining the sequence of DA labels $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$.

We can express the complete sequences of words and DA in terms of the different turns in the dialogue: given a dialogue with T turns, we express its associated word sequence and DA sequence as $\mathcal{W} = W_1^T = W_1 W_2 \cdots W_T$

and $\mathcal{U} = U_1^T = U_1 U_2 \cdots U_T$, respectively. In this notation, W_t represents the sequence of words of turn t , and U_t the sequence of DA of turn t . Thus, we can express the optimisation problem as:

$$\hat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{U}|\mathcal{W}) = \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T|W_1^T). \quad (11.1)$$

From Equation 11.1, we can develop a model based on the application of the rule of Bayes on the formula. In this case, Equation 11.1 can be expressed as:

$$\begin{aligned} \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T|W_1^T) &= \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T) \Pr(W_1^T|U_1^T) = \\ &\underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^T \Pr(U_t|U_1^{t-1}) \Pr(W_t|W_1^{t-1}, U_1^T). \end{aligned} \quad (11.2)$$

A first reasonable assumption we can make is that DA sequences until turn t only affect to the first t turns in the dialogue. In addition, to simplify notation, we will use $W = W_t$ and $U = U_t$, and consequently we will express Equation 11.2 as:

$$\underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^T \Pr(U|U_1^{t-1}) \Pr(W|W_1^{t-1}, U_1^t). \quad (11.3)$$

Previous works such as that presented in (Stolcke et al., 2000) have proposed similar approaches to DA annotation. However, these previous approximations assume the availability of the segmentation of the turn to perform the DA assignment (Stolcke et al., 2000; Webb and Wilks, 2005), when segmentation of the turns is not usual in transcribed dialogues. In our case, we try to generalise the DA assignment problem in the case of unavailable segmentation.

We can develop the formulation to use the model in the unsegmented case. From this point, we introduce single words and DA in the formulation, instead of sequences of words and DA associated to a whole turn. Notice that in this reformulation w and u represent single words and DA, respectively, while W and U represent turns (sequences of words) and DA sequences. The problem is formulated as follows:

- the current word sequence $W = w_1^l = w_1 w_2 \cdots w_l$ is described in terms of a possible segmentation s_1^r as $W = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \cdots w_{s_{r-1}+1}^{s_r}$, where r is the number of segments and s_k is the index of the k -th segment (i.e., the k -th segment is the sequence of words $w_{s_{k-1}+1} \cdots w_{s_k}$), with $s_0 = 0$ and $s_r = l$;
- the DA sequence of turn t is expressed as $U = u_1^r = u_1 u_2 \cdots u_r$;

- the previous DA sequences are expressed as $U_1^{t-1} = U_1 U_2 \cdots U_{t-1}$.

Furthermore, since W_1^T is the sequence of given events, we can neglect the dependency between word sequences. Consequently, the terms in the product in Equation 11.3 are rewritten as:

$$\Pr(U|U_1^{t-1}) \Pr(W|W_1^{t-1}, U_1^t) \approx \sum_{r, s_1^r} \prod_{k=1}^r \Pr(u_k|u_1^{k-1}, U_1^{t-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}). \quad (11.4)$$

This model in Equation 11.4 can be simplified with some assumptions:

- the current DA depends only on the previous $n - 1$ DA:

$$\Pr(u_k|u_1^{k-1}, U_1^{t-1}) \approx \Pr(u_k|u_{k-n+1}^{k-1});$$

- the sequence of words of the current segment depends only on the current DA:

$$\Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \approx \Pr(w_{s_{k-1}+1}^{s_k}|u_k).$$

From this model we can formulate the search problem that looks for the segmentation and DA sequence with maximum probability. This search for the maximum probability is solved using the Viterbi process on the whole dialogue, which results in the following final model:

$$\hat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \max_{R, S} \prod_{t=1}^T \prod_{k=1}^r \Pr(u_k|u_{k-n+1}^{k-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_k). \quad (11.5)$$

In this formula, $R = \{r_1, r_2, \dots, r_T\}$ represents the set of number of segments for each turn and $S = \{s_1^{r_1}, s_1^{r_2}, \dots, s_1^{r_T}\}$ is the set of segmentations for each turn that maximise the product. Notice that to simplify notation, terms in the product are defined in terms of $r = r_t$ and $s_1^r = s_1^{r_t}$.

With respect to the estimation of the presented probability distributions, following the work of other authors (Stolcke et al., 2000; Young, 2000), the terms in Equation 11.5 are modelled as follows:

- $\Pr(u_k|u_{k-n+1}^{k-1})$ is usually represented by a statistical model of DA sequences (DA language model), generally an n -gram model;
- $\Pr(w_{s_{k-1}+1}^{s_k}|u_k)$ is usually modelled by a HMM.

This formulation searches, using a Viterbi process, for the DA sequence of a complete dialogue and gives as by-product a segmentation for each turn. In case there is an available segmentation, the maximisation step is overridden

and the values r and s_1^r are fixed to that provided by the segmentation. In any case, the influence of the models (specially the DA language model) can be tuned by using scaling factors, similar to the Grammar Scale Factor used in speech recognition.

The HMM-based model allows to keep the information on the dialogue history by means of the n -gram of DA, which is an important source of information for the annotation process (the previous DA determine in a great extent the current DA in a real dialogue interaction). The association between word sequences and DA is given by the HMM models, which take into account all the words of the segment; however, this can be a drawback due to data sparseness, especially for long segments. Another interesting feature on the HMM models is their topology. The conventional topology is the one-state-with-loop topology, since it allows the shortest possible segments (one word). However, this topology does not maintain information on the word order, which can be crucial to obtain an accurate segmentation.

3. The NGT Annotation Model

The alternative model is the NGT model, which directly estimates the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$ by means of an n -gram model which acts as a transducer. The definition of this model is based on a Stochastic Finite-State Transducer (SFST) inference technique known as GIATI¹. GIATI (Casacuberta et al., 2005) is a general technique to infer SFST whose first application was in Machine Translation. GIATI starts from a corpus of aligned pairs of input-output sequences. These alignments are used in a re-labelling process that produces a corpus of extended words as a result of a combination of the words of the input and output sentences. This corpus is used to infer a grammatical model (usually a smoothed n -gram). The inversion of the re-labelling process on the grammatical model results in the final SFST, although the use of smoothing techniques makes the conversion of the n -gram to an equivalent SFST difficult. The general GIATI process is presented in Figure 1.

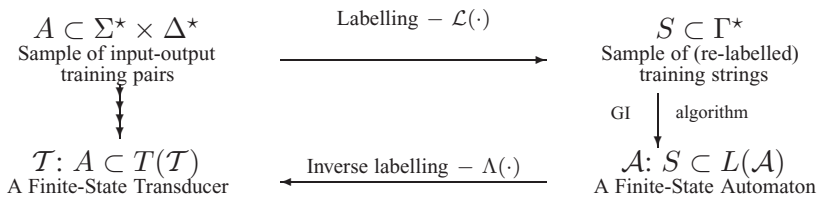


Figure 1. General scheme for the GIATI technique. Σ , Δ and Γ are the input, output, and extended set of symbols, respectively. A and S are the initial sets of aligned and re-labelled samples. $L(\mathcal{A})$ and $T(\mathcal{T})$ represent the languages derived from $\mathcal{A}$ and $\mathcal{T}$, respectively. The GI algorithm is usually the inference of a smoothed n -gram, and $\mathcal{A}$ is the automaton equivalent to the inferred n -gram. $\mathcal{L}$ and Λ are the labelling and inverse labelling functions.

In the case of dialogues, the input language is the sequence of words of the dialogue, the output language is the sequence of DA of the dialogue, and the alignment is between the last word of the segment and the corresponding DA. Thus, for each turn $w_1 w_2 \dots w_l$ and its associated DA sequence $u_1 u_2 \dots u_r$, the re-labelling step attaches the DA label to the last word of the segment using a metasympol ($@$), providing the extended word sequence $e_1 e_2 \dots e_l$, where:

- $e_i = w_i$ when w_i is not aligned to any DA;
- $e_i = w_i @ u_k$ when w_i is aligned to the DA u_k .

Figure 2 presents an example of alignment for a dialogue turn and the corresponding extended word sequence. After the re-labelling process, a grammatical model is inferred. The usual option is a smoothed n-gram.

Yes , uh , I don't work , though , but I used to work and , when	
↑ % ↑ sd	
I had two children .	
↑ sd	
Yes , uh ,@% I don't work , though ,@sd but I used to work and , when I had two children .@sd	

Figure 2. An alignment between a dialogue turn and its corresponding DA labels (from the SWBD-DAMSL scheme, %: uninterpretable, sd: statement-non-opinion), and the result of the re-labelling process, where @ is the attaching metasympol.

In the case of dialogues, the alignments between the words in the turn and the corresponding DA labels are monotonic (no cross-inverted alignments are possible). Consequently, no conversion to SFST is necessary to efficiently apply a search algorithm on the n-gram, since for each input word we can decide whether to emit or not a DA label without referring to posterior words. Therefore, this n-gram acts as a transducer and gives the name to the technique (NGT: N-Gram Transducers) (Martínez-Hinarejos et al., 2008).

The decoding in the NGT model is a Viterbi search which forms a search tree. The i -th level of the tree corresponds to the i -th input word in the sequence. Each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. For example, suppose that next input word w_i was aligned in the training corpora to outputs o_1 , o_2 and o_3 , apart from the empty output. In this case, all the nodes in level $i - 1$ will expand into four children nodes, each of which is associated to the corresponding output (w_i , $w_i @ o_1$, $w_i @ o_2$, $w_i @ o_3$). Therefore, the tree search is expanded in each node in several branches according to the number of outputs associated

to the word. Each new branch represents a possible output (DA), including the empty output (the word is not attached to any DA).

The probability of each branch is updated according to the corresponding n-gram probability using the following steps:

- we start from a parent node P (at level $i - 1$) associated to the sequence of extended words $e_1 e_2 \dots e_{i-1}$ and with an associated probability p_P ;
- if the new word to process w_i has associated in the training corpus o outputs and the empty output, we expand from P the children nodes $e_i^0 = w_i$, $e_i^1 = w_i @ u_1, \dots, e_i^o = w_i @ u_o$;
- the probability of the child node associated to the extended word e_i^j is computed as $p_P \cdot \Pr(e_i^j | e_{i-n+1} \dots e_{i-1})$, where $\Pr(e_i^j | e_{i-n+1} \dots e_{i-1})$ is given by the n-gram of the NGT model.

An example of NGT tree expansion is presented in [Figure 3](#). This expansion on complete dialogues produces a high temporal and spatial complexity, which is admissible in the off-line dialogue annotation framework. The search process can be applied to dialogues with unsegmented turns giving, as in the case of the HMM-based model, a segmentation as by-product. NGT can be applied on segmented turns by restricting the outputs to the end words of the segments.

The main drawback of this initial approach is its high locality: only the last $n - 1$ extended words are really taken into account to assign the DA labels. This makes the current DA independent from most of the previous DA in the dialogue history, and loses an important source of information. In the tree of [Figure 3](#) we can see that the last node of the best hypothesis (in boldface and marked by an arrow) calculates its probability based only on the two previous nodes values (“don’t work”), ignoring previous DA.

In (Martínez-Hinarejos et al., 2009), we proposed a modification of the basic search algorithm in which the probability of the different branches is not only computed from the n-gram transducer itself, but from an n-gram of DA as well (which acts as DA language model). Therefore, when expanding a branch of a word w_i with a DA u_j , the new probability is computed using both the n-gram transducer and the DA language model (a n-gram of degree m).

Consequently, the calculation of the probability for the child node associated to $e_i = w_i @ u_j$ is changed. In this case, it is given by $p_P \cdot \Pr(e_i | e_{i-n+1} \dots e_{i-1}) \cdot \Pr(u_j | u_{j-m+1} \dots u_{j-1})$. No change in the computation of the probability of the child node is produced when the output is empty (i.e., $e_i = w_i$).

The expansion process in this new version can be seen in the tree of [Figure 4](#). We can see that the probability of the last node of the best hypothesis is calculated using the values of the two previous nodes (“don’t work”) and the

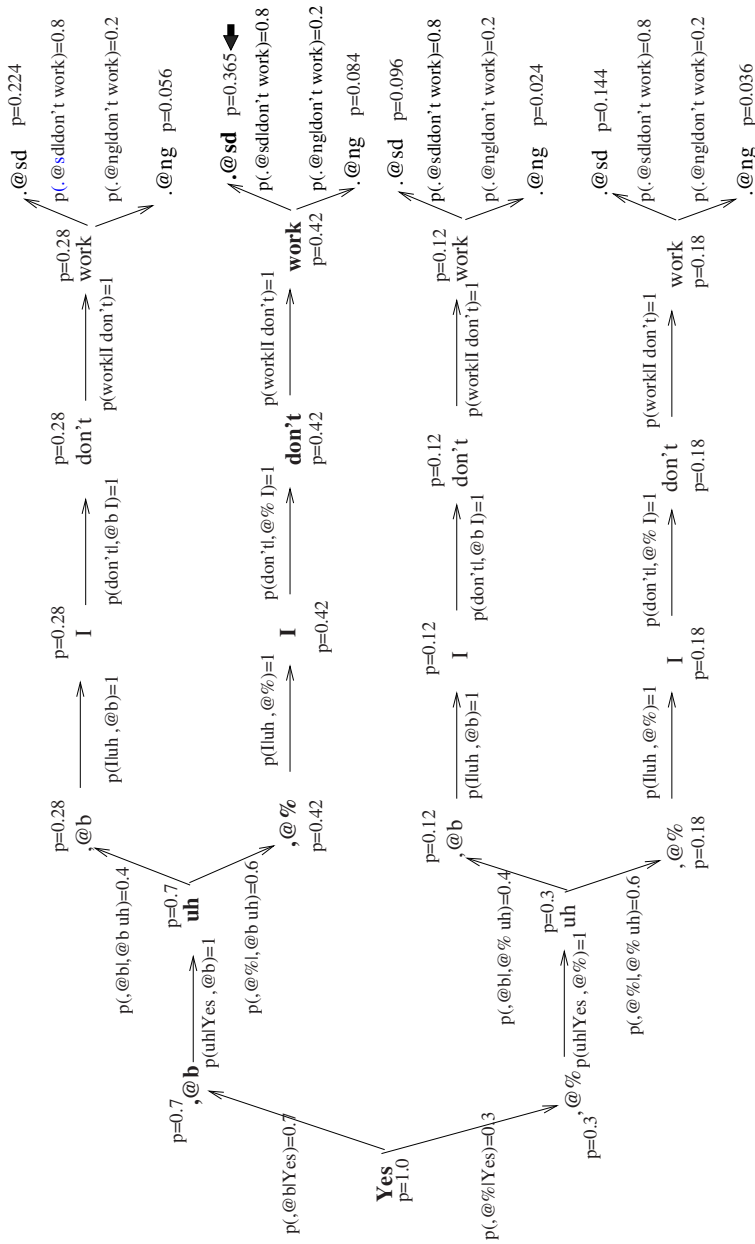


Figure 3. An example of the Viterbi tree search for the basic NGT model. In this case only the NGT probabilities are taken into account. Best hypothesis (in boldface and marked by the dark arrow) is (b % sd). In this example, the NGT model is a trigram.

two previous DA (“b %”). With this enhancement, the NGT model keeps information on the DA history and is competitive with respect to the HMM-based model, as the results in Section 5 will show.

In summary, the NGT model keeps information on the dialogue history (using the same n -gram of DA than the HMM-based model), but it only takes into account the last $n - 1$ words of a segment to perform the assignment of DA. Consequently, the number of events used to assign DA are lower than in the case of HMM, but this makes it more robust to data sparseness. With respect to the segmentation accuracy, the alignment-based nature of the NGT model makes it more suitable to obtain a more accurate segmentation of the unsegmented turns.

However, this model presents a high spatial and temporal complexity with respect to the HMM-based model: if the mean number of outputs for all words is k , in the i -th level of the exploration tree we will have k^i nodes on average. As the exploration tree must be maintained until the last word of the dialogue to retrieve the best path, this produces an exponential growth of the needed time and space, even when beam search is applied.

Although this exponential complexity is not especially important in the annotation process (since it is an off-line process without real-time requirements), a faster annotation process speeds up the construction of the global dialogue system. Thus, it is convenient to obtain variations of NGT which keep the annotation accuracy with a lower complexity. Following this idea, we proposed a modification of the search process in (Tamarit et al., 2009) which is based on only allowing the expansion of n branches in each node. These n branches correspond to those yielding the highest probabilities from the total number of expansion branches. The implementation consists of expanding all possible new branches but keeping only the n best in an auxiliary space and not linking them to the parent node. After all the expansion, only the branches in the auxiliary space (the n best branches derived from the parent node) are linked to the parent node. With this process we reduce the expansion on the i level of the tree to n^i , and when $n \ll k$, this provides an important reduction on space and time. The results presented in Section 5 will show that this reduction does not affect the quality of the results produced by the NGT model.

4. Corpora

In this section we present the features of the corpora we used to evaluate the performance of the presented statistical models. In our case, we used two corpora with very different features, in order to assess the robustness of the models against different conditions and to obtain more reliable conclusions. [Table 1](#) summarises the main features between the two corpora used in the experiments: the SwitchBoard corpus (Godfrey et al., 1992) and the DIHANA

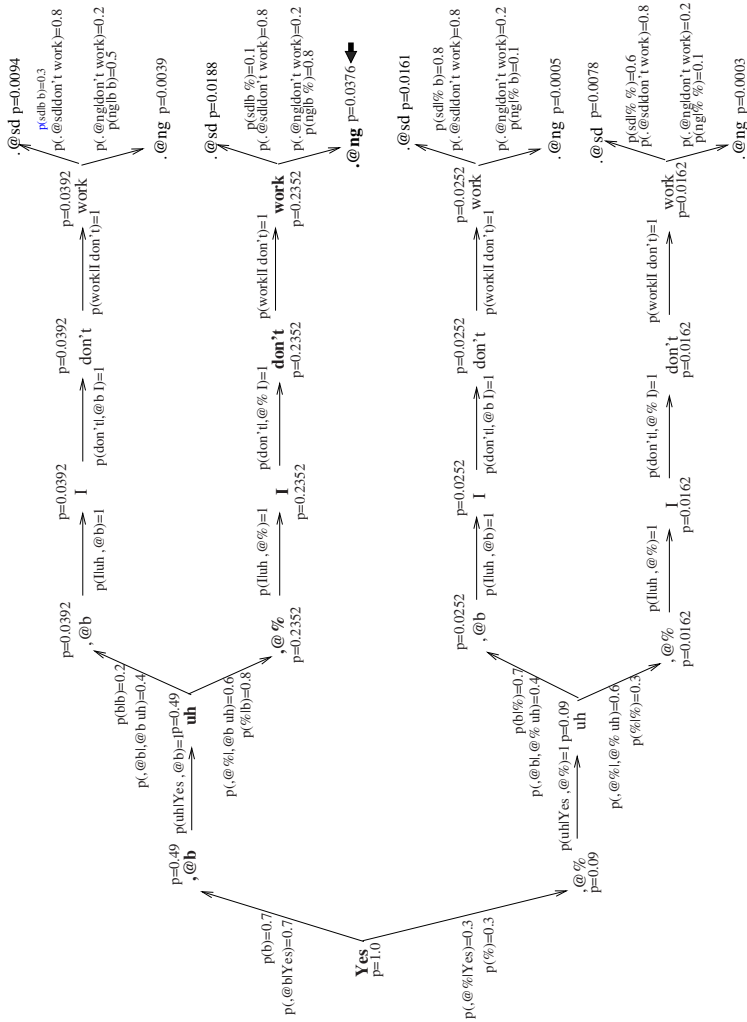


Figure 4. An example of the Viterbi tree search for the enhanced NGT model. In this case, both the NGT probabilities and the DA language model probabilities are used in the transitions where outputs are produced. In these nodes, the new probability is the result of multiplying the probabilities of the parent node, the NGT probability and the DA language model probability. Best hypothesis (in boldface and marked by the dark arrow) changes to (b % ng). In this example, both the NGT model and the DA language model are trigrams.

corpus (Benedí et al., 2006). The main difference between them is that DIHANA is a corpus with a clear semantic restriction, based on the definition of scenarios and objectives to accomplish, whereas SwitchBoard does not present a clear objective and the interactions present a broader semantic focus.

Table 1. Summary of the different features of the SwitchBoard and the DIHANA corpus. For DIHANA, U means user turns and S system turns.

Corpus	SwitchBoard	DIHANA
Language	English	Spanish
Nature	Human-human	Human-computer
Number of dialogues	1,155	900
Number of turns	115,000	6,280 U + 9,133 S
Annotation scheme	SWBD-DAMSL	IF-DIHANA (3 and 2 levels)
Number of DA labels	42	248 (3 levels) / 72 (2 levels)

4.1 SwitchBoard Corpus

The SwitchBoard corpus (Godfrey et al., 1992) is a well-known corpus of human-to-human telephone conversations in English. The conversations are about general topics, with no clear task to accomplish. This corpus recorded spontaneous speech, with frequent interruptions between the speakers, hesitations, non-linguistic sounds (laugh, cough) and background noises. The transcription of the corpus takes these phenomena into account, and it includes special notation for the overlaps and different noises produced in the recording.

The corpus consists of 1,155 conversations, with approximately 115,000 different turns with 16 words by turn on average. The amount of speech signal is about 95 hours. The vocabulary size is about 42,000 words. The dialogue annotation was performed using the SWBD-DAMSL scheme (Jurafsky et al., 1997), a simplified version of the standard DAMSL annotation set (Core and Allen, 1997). In the process, the dialogue turns were split into segments and each segment was annotated with one of the 42 different labels of the SWBD-DAMSL scheme. These labels represent several dialogue communicative categories such as statement, question, backchannels, etc., and the corresponding subcategories (e.g., statement opinion/non-opinion, yes-no/open question, etc.). An example of annotation is shown in [Figure 5](#). The manual labelling was performed by 8 different human labellers, with a Kappa value of 0.80 (Stolcke et al., 2000). After the annotation process there are 1.7 segments per turn on average.

To simplify the experimental framework, we preprocessed the SwitchBoard corpus to remove certain phenomena: interruptions and overlaps were erased

Speaker	Segment	Transcription
		Label
S1	S1-1	Yeah,
		aa
	S1-2	to get references and that,
		sd
	S1-3	so, but, uh,
		%
	S1-4	I don't feel comfortable about leaving my kids in a big day care centre, simply because there's so many kids and so many <sniffing> <throat_clearing>
		sd
S2	S2-1	I think she has problems with that, too.
		sd

Figure 5. An example of annotated turns in the SwitchBoard corpus. The meaning of the labels is statement-non-opinion (sd), uninterpretable (%) and agree/accept (aa).

(by joining the interrupted turns), all the words were transcribed to lowercase and punctuation marks were separated from the words. This preprocess is reasonable for the annotation of transcribed dialogues or for speech recognitions that can provide punctuation marks (e.g., based on prosodic features), but for most of the speech dialogues it should be changed, since punctuation marks are not usually part of speech recognisers outputs.

4.2 DIHANA Corpus

The DIHANA corpus (Benedí et al., 2006) is a set of 900 task-oriented human-computer dialogues in Spanish. The task is about railway information for timetables, fares and services for long-distance trains in the Spanish territory. The corpus was acquired from conversations with 225 voluntary speakers (153 male and 72 female), with small Spanish dialectal variants. The acquisition was performed using the Wizard of Oz technique (Fraser and Gilbert, 1991), and it only had semantic restrictions (the objective of the interaction was defined by mean of scenarios), but not lexical or syntactical restrictions.

The acquisition process resulted in 6,280 user turns and 9,133 system turns, with a vocabulary of approximately 900 words and a final amount of speech signal of about five and a half hours. On average, there are 15 words (9 for user turns and 20 for system turns) and 1.5 segments per turn (without variations for user and system). The dialogue annotation scheme was defined based on the Interchange Format (IF) (Fukada et al., 1998), which defines labels with three

different levels, called respectively speech act, concept and argument. The adaptation for the DIHANA corpus represents the general purpose of the segment (first level), as well as more precise semantic information that is specific to each task (second and third levels). The second level contains the repository of information implicit in the segment (i.e., the set of data used or modified according to the intention given by the first level). The third level represents the specific data present in the segment.

All the dialogues were manually transcribed. These transcriptions were used to annotate the corpus by means of a semiautomatic procedure (Alcácer et al., 2005). Next, all the dialogues were manually corrected by human experts using a very specific set of defined rules (Alcácer et al., 2005). The annotation of all the dialogues was consistently revised by a single expert. Figure 6 shows a sample of annotated dialogue (in English) from the DIHANA corpus.

The DA definition resulted in a set of 248 different 3-level labels (153 for user turns and 95 for system turns) (Alcácer et al., 2005). Due to the high specificity of the third level (which takes into account the specific data used or provided in the segment), an alternative labelling using only the first two levels is also considered in the experiments. In this 2-level case, there are 72 different labels (45 for user and 27 for system).

In summary, the DIHANA corpus can be viewed as a medium-sized, spontaneous-speech dialogue corpus involving a well-known task. Its features show DIHANA as one of the first and largest task-oriented corpora in Spanish. All the dialogues were transcribed and annotated. The DA annotation scheme follows many of the principles used in other projects with a three-level structure that covers the general intention as well as more specific details of the domain of the task.

To simplify the experimental framework, the DIHANA corpus was preprocessed to reduce its complexity. In this case, as in the case of the SwitchBoard corpus, all the words were transcribed to lowercase and punctuation marks were separated from the words. Additionally, a categorisation of sequences such as town names, dates, hours, etc. was performed, and the words were speaker-labelled (U for user and S for system).

5. Experimental Results

We propose a set of experiments to compare the performance of the two models introduced in Sections 2 and 3. The models were proved with the two corpora described in Section 4, and the experiments were made using a cross-validation approach². For both corpora, we present results for the annotation using the unsegmented version of the dialogues. In both cases the weight parameter of the HMM-based model (which scales the influence of the DA language model) was optimised for the whole cross-validation process. We only

Speaker	Segment	Transcription		
		Level 1	Level 2	Level 3
S	S1	Welcome to the railway information system. How may I help you?		
		Opening	Nil	Nil
U	U1	I want to know the departure times from Valencia		
		Question	Departure-hour	Origin
	U2	to Madrid		
		Question	Departure-hour	Destination
	U3	arriving on May the 15th of 2,004.		
		Question	Departure-hour	Day
S	S2	Do you want to leave on Saturday, May the 15th of 2,004?		
		Confirmation	Day	Day
U	U4	Yes.		
		Acceptance	Day	Nil
S	S3	Consulting times for trains from Valencia to Madrid on Saturday, May 15th of 2,004.		
		Confirmation	Departure-hour	Destination, Day, Origin
	S4	Wait a moment, please.		
		Waiting	Nil	Nil
	S5	There are several trains. The first one leaves at 7:45 and arrives at 11:14, and the last one leaves at 18:45 and arrives at 22:18.		
		Answer	Departure-hour	Arrival-hour, Departure-hour, Order-number, Number-trains
	S6	Do you need anything else?		
		Consult	Nil	Nil
U	U5	Yes, I want to know the fare for the train leaving at 7:45.		
		Question	Fare	Departure-hour
S	S7	That train in tourist class costs 35.50 euros.		
		Answer	Fare	Class, Fare
	S8	Do you need anything else?		
		Consult	Nil	Nil
U	U6	No, thank you.		
		Closing	Nil	Nil
S	S9	Thanks for using this service. Have a nice day.		
		Closing	Nil	Nil

Figure 6. An example of an annotated dialogue in English from the DIHANA corpus. *Nil* denotes the absence of information.

<i>Original annotated sentence</i>
Yeah , (aa) to get references and that , (sd) so , but , uh , (%) I don't feel comfortable about leaving my kids in a big day care centre , simply because there's so many kids and so many <sniffing> <throat_clearing> (sd)
<i>Derived sequence for SegDAER measure</i>
2-(aa) 8-(sd) 14-(%) 40-(sd)

Figure 7. An example of how to obtain the sequences used to calculate the SegDAER measure from an annotated turn of the SwitchBoard corpus.

show the results for the best weight parameter. The Viterbi search in the NGT model was a beam-search with a dynamic beam parameter.

In order to evaluate the models we compute the Segmentation DA Error Rate (SegDAER). The SegDAER is an average edit distance between sequences derived from the reference and the annotation result. In this case, the sequences are a combination of the DA label and its position (segmentation). Figure 7 shows how the sequences to be compared are obtained: each DA label is joined with the position of the last word of the corresponding segment, and these sequences of position-DA are compared using the edit distance. With SegDAER we can evaluate at the same time the quality of the annotation and the segmentation obtained. Although other evaluation metrics can be used (see (Ang et al., 2005)), we consider this metric a good choice to evaluate the quality of the techniques in the annotation and segmentation task. In all the experiments, confidence intervals of 90% were calculated using bootstrapping with 10,000 repetitions using the method described in (Bisani and Ney, 2004). We also measured the average seconds per dialogue required for each experiment. We launched the experiments in an Intel Xeon E5420 machine at 2.5GHz with 14GB of RAM.

In Table 2 we show the SegDAER for the experiments with the SwitchBoard and DIHANA corpora using the HMM-based model and the NGT model. The results for the HMM-based model are considered as baseline results. For the NGT model we include the results of the 3-gram and 4-gram, using a 3-gram and a 4-gram as DA language model, for different values of n -limited expansion (no limited, 3 and 5). In Table 3 we show the seconds per dialogue of each experiment. Additional results on only annotation errors (DAER measure, which only compares the DA sequences and not their positions) are presented in Table 4.

The results for the SwitchBoard corpus show that NGT produces a quite better annotation than the HMM-based model. In this case, the absolute SegDAER gets reduced in more than an 11% when using the NGT model (with respect to the SegDAER obtained with the HMM-based model, i.e., from 61.8% to 50.3%). The limited search in NGT produces a reduction in the annotation time that it affects in a low extent the accuracy of annotation. However, the

Table 2. Annotation and segmentation error (SegDAER) of the SwitchBoard and DIHANA corpora using the HMM-based and the NGT models. Confidence intervals are in all cases lower than ± 0.2 for the SwitchBoard corpus and lower than ± 0.6 for the DIHANA corpus. The table shows the results with different n-grams to estimate the DA language model. In boldface the best result for each corpora and model.

Model	DA N-gram	n-limit	SwitchBoard	SegDAER	
				DIHANA 2	DIHANA 3
HMM	3	-	61.8	29.1	34.4
	4	-	62.0	29.3	34.6
NGT-3g	3	No	51.0	9.1	18.7
		3	50.4	9.5	19.2
		5	50.3	9.1	18.8
		No	51.5	9.0	19.4
	4	3	50.5	9.4	19.8
		5	50.4	9.2	19.4
NGT-4g	3	No	53.6	8.3	17.9
		3	52.3	8.7	18.5
		5	52.3	8.4	18.1
		No	53.2	8.3	18.1
	4	3	52.4	8.8	18.7
		5	52.3	8.5	18.3

Table 3. Seconds per dialogue annotation of the SwitchBoard and DIHANA corpora using the HMM-based and the NGT models. The table shows the results with different n-grams to estimate the DA language model. In boldface the best result for each corpora and model.

Model	DA N-gram	N-best	Seconds/dialogue		
			SwitchBoard	DIHANA 2	DIHANA 3
HMM	3	-	3.1	1.1	20.0
	4	-	5.1	1.8	27.6
NGT-3g	3	No	94.0	23.3	26.8
		3	76.3	19.0	23.5
		5	82.0	20.2	24.4
		No	95.7	25.5	28.9
	4	3	76.3	21.7	26.8
		5	82.8	22.9	27.8
NGT-4g	3	No	114.5	26.5	30.4
		3	95.5	23.0	27.8
		5	103.8	24.3	28.7
		No	116.5	28.9	32.9
	4	3	95.8	25.8	30.9
		5	104.1	26.9	32.1

Table 4. Annotation error (DAER, comparing only DA labels of the sequences used to compute SegDAER) of the SwitchBoard and DIHANA corpora using the HMM-based and the NGT models. The table shows the results with different n-grams to estimate the DA language model. In boldface the best result for each corpora and model.

Model	DA N-gram	n-limit	SwitchBoard	DAER	
				DIHANA 2	DIHANA 3
HMM	3	-	55.5	8.0	15.5
	4	-	55.8	8.0	15.7
NGT-3g	3	No	47.3	8.6	18.2
		3	46.6	9.0	18.7
		5	46.5	8.7	18.3
	4	No	47.8	8.6	18.9
		3	46.7	8.9	19.3
		5	46.6	8.7	18.9
NGT-4g	3	No	49.3	7.9	17.3
		3	47.9	8.3	17.9
		5	47.9	8.0	17.5
	4	No	48.9	8.0	17.5
		3	47.9	8.3	18.0
		5	47.9	8.1	17.7

annotation time of NGT is really larger than that needed for the HMM-based model. This is caused by the natural exponential complexity with respect to the input length of the NGT technique, while the HMM-based model is linear with respect to this length. In any case, the convenience of using NGT in annotation is reinforced by the off-line nature of the corpus annotation task (i.e., there are no “real-time” requirements), although it would be desirable a reduction of the time to speed up even more the annotation process.

The results for the DIHANA corpus show that the NGT technique produces a dramatical increase (even higher than in SwitchBoard) on the annotation accuracy with respect to that obtained with the HMM-based model. In this case, the absolute SegDAER reduction is of a 21% in the 2-level version (from 29.1% to 8.3%) and of a 17% in the 3-level version (from 34.4% to 17.9%). Besides, as it happened in SwitchBoard, the limited search of NGT produces a moderate reduction of the annotation time, but it does not produce significant differences in annotation accuracy between experiments. The experiments also show that the HMM-based model obtains much better annotation times than NGT when using the 2-level version of DIHANA (but in a lower degree than in SwitchBoard). However, for the 3-level version of DIHANA, NGT obtains similar times to those of the HMM-based model. This is due to the dependency of time cost of the HMM-based model on the number of HMM models, whereas

in the NGT model complexity does not depend on the number of labels (it depends only on the length of the dialogue).

6. Conclusions and Future Work

In this chapter we compared the performance of a more classical model for dialogue annotation (the HMM-based model) against the new NGT model. The presented results show that the NGT model produces a high increment in the annotation accuracy with respect to that obtained by the HMM-based model. These results were common to two different corpora of very different nature: in the SwitchBoard corpus, an 11% of SegDAER reduction was achieved, while in the DIHANA corpus, a 21% and 17% of reduction was obtained for the 2-level and 3-level label versions, respectively. The main drawback of the NGT model is its high temporal complexity. Although the *n*-limited search in NGT produces time improvements, for long dialogues (such as SwitchBoard dialogues) and relatively small sets of DA labels (such as SwitchBoard or DIHANA 2-levels) the HMM-based model is one magnitude-order faster. This situation changes when the number of DA labels increases (as the time results for DIHANA 3-levels show), where the NGT model becomes as fast as the HMM-based model. However, since the nature of the dialogue annotation process does not require a fast response, the NGT model seems the most convenient technique for dialogue annotation.

The success of the NGT model in this task opens a broad variety of future applications. In any case, a first aim would be improving the time complexity of the NGT implementation used in these experiments, since a faster annotation is always more desirable for improving the human annotator performance. Some other improvements in the annotation application can be implemented to enhance the human annotator activity. One of them is the definition of confidence measures in the annotation, which can speed up the review of the draft annotation by the human expert. Another interesting point is the implementation of an annotation interface that uses NGT in an incremental fashion, i.e., a first set of dialogues is annotated from scratch and these dialogues are used to infer the models, which are used to automatically annotate a new set of dialogues that, after being corrected by the human expert, are added to the training set to infer new annotation models that will be used in the next set of dialogues. With respect to the selection of the partitions to be annotated, the use of active learning would be interesting in order to improve the process, because some subsets of dialogues could be more useful to infer the models than others.

Apart from the annotation of dialogues, one of the most interesting application of NGT is its use as language model in a speech recogniser. This is of particular interest in real dialogue systems, since the user input could be directly decoded into DA labels by using the NGT *N*-gram as language model.

In this case, since the input is an only turn for each decoding, the time complexity would be lower and real-time can be achieved. In any case, the use of n -limited search can aid the objective of achieving real-time decoding of the user input.

Acknowledgments

This research work was partially supported by the EC (FEDER) and the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme, the grant TIN2009-14633-C03-01 and the PROMETEO/2009/014 project. We also wish to thank the anonymous reviewers for their criticism and suggestions.

Notes

1. GIATI is the acronym for Grammatical Inference and Alignments for Transducer Inference.
2. The NGT software and the preprocessed SwitchBoard corpus and its cross-validation partitions are available at the web page <http://users.dsic.upv.es/cmartine/research/resources.html>.

References

- Alcácer, N., Benedí, J. M., Blat, F., Granell, R., Martínez, C. D., and Torres, F. (2005). Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In *Proceedings of International Conference on Speech and Computer (SPECOM)*, pages 583–586, Greece.
- Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., and Siegel, M. (1998). Dialogue Acts in VERBMOBIL-2 - Second Edition. Verbmobil Report 226, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany.
- Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1061–1064, Philadelphia.
- Aust, H., Oerder, M., Seide, F., and Steinbiss, V. (1995). The Philips Automatic Train Timetable Information System. *Speech Communication*, 17:249–263.
- Austin, J. (1962). *How to Do Things with Words*. Oxford University Press, London.
- Benedí, J. M., Lleida, E., Varona, A., Castro, M. J., Galiano, I., Justo, R., López, I., and Miguel, A. (2006). Design and Acquisition of a Telephone Spontaneous Speech Dialogue Corpus in Spanish: DIHANA. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639, Genova, Italy.
- Bisani, M. and Ney, H. (2004). Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation. In *Proceedings of International Conference*

- on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 409–412.
- Bos, J., Klein, E., Lemon, O., and Oka, T. (2003). DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124.
- Bunt, H. (1994). Context and Dialogue Control. *THINK Quarterly*, 3.
- Casacuberta, F., Vidal, E., and Picó, D. (2005). Inference of Finite-State Transducers from Regular Languages. *Pattern Recognition*, 38(9):1431–1443.
- Core, M. G. and Allen, J. F. (1997). Coding Dialogues with the DAMSL Annotation Scheme. In Traum, D., editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI.
- Dybkjær, L. and Minker, W., editors (2008). *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Springer, Dordrecht.
- Fraser, M. and Gilbert, G. (1991). Simulating Speech Systems. *Comp. Speech Lang.*, 5:81–99.
- Fukada, T., Koll, D., Waibel, A., and Tanigaki, K. (1998). Probabilistic Dialogue Act Extraction for Concept based Multilingual Translation Systems. In *Proceedings of International Conference on Spoken Language Processing*, volume 6, pages 2771–2774.
- Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'92)*, pages 517–520.
- Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual - draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science.
- Martínez-Hinarejos, C. D., Benedí, J. M., and Granell, R. (2008). Statistical Framework for a Spanish Spoken Dialogue Corpus. *Speech Communication*, 50:992–1008.
- Martínez-Hinarejos, C. D., Tamarit, V., and Benedí, J. M. (2009). Improving Unsegmented Dialogue Turns Annotation with N-gram Transducers. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, volume 1, pages 345–354, Hong Kong. City University of Hong Kong Press.
- McTear, M. F., Allen, S., Clatworthy, L., Ellison, N., Lavelle, C., and McCaffery, H. (2000). Integrating Flexibility into a Structured Dialogue Model: Some Design Considerations. In *Proceedings of International Conference on Speech and Language Processing*, pages 943–946.

- Meng, H. M., Wai, C., and Pieraccini, R. (2003). The Use of Belief Networks for Mixed-Initiative Dialog Modeling. *IEEE Transactions on Speech and Audio Processing*, 11(6):757–773.
- Rangarajan, V., Bangalore, S., and Narayanan, S. (2007). Exploiting Prosodic Features for Dialog Act Tagging in a Discriminative Modeling Framework. In *Proceedings of Interspeech*, Antwerp, Belgium.
- Seneff, S. and Polifroni, J. (2000). Dialogue Management in Mercury Flight Reservation System. In *Proceedings of ANLP-NAACL 2000, Satellite Workshop*, pages 1–6.
- Stolcke, A., Coccoaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., and Meteer, M. (2000). Dialogue Act Modelling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):1–34.
- Tamarit, V., Martínez-Hinarejos, C. D., and Benedí, J. M. (2009). Reduction of the Temporal Complexity of N-gram Transducers for Dialogue Annotation. In *Proceedings of the First International Workshop of Spoken Dialogue Systems Technology*, pages 302–305, Irsee (Germany). IWSDS.
- Walker, M. and Passonneau, R. (2001). DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. In *Proceedings of International Conference on Human Language Technology Research (HLT)*, pages 1–8, San Diego.
- Webb, N. and Wilks, Y. (2005). Error Analysis of Dialogue Act Classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD)*, pages 451–458.
- Wilks, Y. (2006). COMPANIONS: Intelligent, Persistent, Personalised Interfaces to the Internet. <http://www.companions-project.org>.
- Williams, J. D. and Young, S. (2007). Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Comput. Speech Lang.*, 21(2):393–422.
- Young, S. (2000). Probabilistic Methods in Spoken Dialogue Systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402.
- Zimmermann, M., Liu, Y., Shriberg, E., and Stolcke, A. (2005). A* based Joint Segmentation and Classification of Dialog Acts in Multi-Party Meetings. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 581–584.

Index

- acoustic and prosodic feature, 84, 102
- ambient intelligence, 135, 136, 143, 148, 150, 151, 157, 182
- anger recognition, 83–87, 89, 101–103
- Bayes risk, 29, 31, 37–42, 46, 48
- consulting dialogue, 231, 232, 235, 237, 238, 241, 244, 251, 252
- corpus annotation, 270
- DBN, 163, 165, 166, 169, 170, 172, 173, 175, 176, 178, 180, 181
- dialogue
 - act, 61, 63, 231, 255, 256
 - annotation, 255, 257, 262, 266, 267, 271–273
 - management, 29–31, 37–39, 42, 55, 56, 60, 68, 70, 77, 137, 140, 142–144, 146, 147, 149, 153, 154, 156, 164, 165, 248–250
 - modelling, 56, 140, 142, 155, 256
 - phenomena, 137, 140, 142, 144
 - system, 2, 24, 36, 53–56, 60–63, 65, 66, 68, 69, 72, 73, 76, 77, 221, 232, 237, 239, 245, 248, 249, 251, 252, 255, 256, 265, 273
- embedded system, 1, 7, 19, 24
- Embodied Conversational Agent, 151, 152, 154, 214, 215, 221, 226
- emotion
 - processing, 154
 - recognition, 84
- evaluation methodology, 54
- hedonic quality, 226, 227
- interactive presentation, 187, 189, 200, 203, 204
- IVR System, 84
- language model, 2, 36, 56, 57, 61, 68, 70–72, 74–77, 147, 149, 150, 156, 200, 204, 205, 244, 249, 259, 262, 264, 268, 270–273
- multilingual, 1, 3, 8, 9, 15, 19, 20, 24, 141
- MVC model, 189
- N-gram transducer, 255, 257, 261, 262
- non-native, 1, 4–8, 10, 11, 13, 14, 19, 21, 22, 24
- onomatopoeia, 108–110
- POMDP, 44, 60, 163–174, 176, 178–182, 232
- pragmatic quality, 224, 227
- projection, 5, 11, 15, 17–21, 24
- psychomime, 107–111, 115–120, 122–132
- rapid development, 187
- reward, 29, 31, 37–40, 43–48, 163, 165, 166, 168, 170, 171, 173, 176, 178–181
- selection and classification, 84
- self-organizing map (SOM), 107, 108, 111–119, 121, 122, 128, 131, 132
- semi-continuous, 2–5, 7, 12
- speech
 - corpus, 58, 204, 251
 - recognition, 1–4, 7, 20, 24, 36, 56, 68, 70, 71, 76, 77, 85, 132, 137, 141, 145–149, 200, 249, 259, 266
- spoken dialogue system, 1, 2, 14, 30, 38, 53–57, 66, 76, 108–110, 131, 132, 135–140, 142, 143, 145–147, 151, 152, 155, 156, 187–189, 197, 200, 231, 233, 239, 250, 251
- statistical
 - method, 137, 140, 142, 156, 249
 - model, 59, 255–257, 259, 265
- user
 - model, 61, 65, 149, 189, 201, 206–208
 - modelling, 54, 60, 156, 207, 208
 - simulation, 53, 55, 61, 62, 64, 65, 144
- value iteration, 163–166, 168, 170, 173, 176, 178, 182
- voice search, 135, 136, 143, 144, 147–150, 156, 157