Proceedings of the Fifth Annual

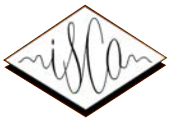
Young Researchers' Roundtable on Spoken Dialogue Systems'09

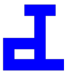
Queen Mary University of London, UK
September 13th–14th, 2009

http://www.yrrsds.org/

Sponsors

Endorsements





Organising Committee

David Díaz Pardo de Vera,
Signal Processing Applications Group (GAPS),
Politechnic University of Madrid, Spain

Milica Gašić,
Department of Engineering,
University of Cambridge, UK

François Mairesse,
Department of Engineering,
University of Cambridge, UK

Matthew Marge,
Language Technologies Institute,
Carnegie Mellon University, USA

Joana Paulo Pardal,
L2F INESC-ID and
IST, Technical University of Lisbon, Portugal

Ricardo Ribeiro,
L2F INESC-ID and
ISCTE-IUL, Lisbon, Portugal

Local Organisers

Arash Eshghi,
Interaction, Media and Communication Group,
Queen Mary University of London, UK

Christine Howes,
Interaction, Media and Communication Group,
Queen Mary University of London, UK

Gregory Mills,
Interaction, Media and Communication Group,
Queen Mary University of London, UK

Foreword

Welcome to the fifth annual Young Researchers Roundtable on Spoken Dialogue Systems, which builds on the good work of the previous workshops, held in Columbus (2008), Antwerp (2007), Pittsburgh (2006) and Lisbon (2005).

The Young Researchers Roundtable aims to promote networking and discussion in the rich and varied field of dialogue systems between students and young researchers in both academia and industry. It is our hope that our program will facilitate and encourage this, and that our time together will be worthwhile, productive, and fun!

We would like to thank our sponsors, Orange, Microsoft Research and AT&T Labs-Research, for giving us the opportunity to make the workshop fee as low as possible for our participants and for providing coffee breaks, lunch and dinner. Thanks also to ISCA, SIGDial, and IEEE SLTC for endorsing and promoting the event, and thus broadening our reach. Special thanks to Queen Mary University of London for enabling us to use the university's rooms. Further, we would like to particularly thank the members of our advisory committee for providing assistance and helping to promote the event. And last, but not least, we would like to thank all of you, this year's participants, for providing interesting and varied position papers and thought-provoking questions for discussion, and without whom this event would not be possible.

We hope you all enjoy the workshop, and have fun in London!

The organising committee of YRRSDS 2009

Contents

Program

Day before - September 12th, Saturday

19:30 YRRSDS get-together

First Day - September 13th, Sunday

08:30 Registration & Coffee

09:00 Reception session
(with photo and 30 sec introduction / participant)

09:30 Discussion session I
Topics: to be determined based on participants interests

11:30 Summaries & discussion

12:00 Industry/Company talks: sponsors presentation

- Philippe Bretier, *Orange*
- Tim Paek, *Microsoft Research*
- Sebastian Möller, *DE Telecom*

13:00 Lunch break

14:00 Discussion session II
Topics: to be determined based on participants interests

15:30 Summaries & discussion

16:00 Tea break

16:30 Industry & Academic Panel: session on career development

- Jason Williams, *AT&T*
- Michael McTear, *University of Ulster*

18:30 Dinner

Second Day - September 14th, Monday

09:00 Coffee

09:30 Demos and Posters

10:30 Frameworks and Grand Challenges for Dialogue System Evaluation

- Alan Black, *Carnegie Mellon University*
- Dan Bohus, *Microsoft Research*

11:00 Discussion Sessions on Evaluation

- Automative
- Human-Robot Dialogue
- Let's Go
- Multimedia
- Tutorial

12:00 Summaries & discussion with Panelists

12:30 Close up

13:00 Lunch break and farewell

Advisory Committee

Hua Ai,
University of Pittsburgh, USA

James Allen,
University of Rochester, USA

Alan Black,
Carnegie Mellon University, USA

Dan Bohus,
Microsoft Research, USA

Philippe Bretier,
Orange, France

Robert Dale,
Macquarie University, Australia

Maxine Eskenazi,
Carnegie Mellon University, USA

Sadaoki Furui,
Tokyo Institute of Technology, Japan

Luis Hernández Gómez,
Polytechnic University of Madrid, Spain

Carlos Gómez Gallo,
University of Rochester, USA

Kristiina Jokinen,
University of Helsinki, Finland

Nuno J. Mamede,
L^2F INESC-ID, Technical University of Lisbon, Portugal

David Martins de Matos,
L^2F INESC-ID, Technical University of Lisbon, Portugal

João Paulo Neto,
L^2F INESC-ID, Technical University of Lisbon, Voice Interaction, Portugal

Tim Paek,
Microsoft Research, USA

Antoine Raux,
Honda Research Institute, USA

Robert Ross,
University of Bremen, Germany

Alexander Rudnicky,
Carnegie Mellon University, USA

Mary Swift,
University of Rochester, USA

Isabel Trancoso,
L^2F INESC-ID, Technical University of Lisbon, Portugal

Tim Weale,
The Ohio State University, USA

Jason Williams,
AT&T Labs, USA

Sabrina Wilske,
Saarland University, Germany

Andi Winterboer,
University van Amsterdam, Netherlands

Craig Wooton,
University of Ulster, UK

Steve Young,
University of Cambridge, UK

Industry Speakers

Philippe Bretier

Affiliation: Orange

Presentation Abstract:
Recent research in spoken dialogue systems (SDS) has called for a synergistic convergence between research and industry. The talk presents one of the works Orange Labs is carrying under this motivation.

Manual dialogue design and adaptive dialogue management reflect more or less the respective paths that industry and research have followed when building their SDS. Dialogue evaluation is a common concern for both communities, but remains hard to put into operational perspectives. The talk shows how design tools, monitoring tools and dedicated reinforcement learning algorithms can be complementary and offer a new convergent experience for developers in designing and evaluating SDS. First, the integration of logs feedbacks into an industrial SDS design tool makes it a monitoring tool as well, by revealing the effective call flows and their associated Key Performance Indicators. Second, the integration of design alternatives into a single dialogue design makes the SDS developers capable to visually compare different design choices. Third, the integration of a reinforcement learning technique allows automatic optimisation of the SDS choices. Finally, the design/monitoring tool helps the SDS developers to understand the learnt behaviour. The SDS developers can then contrast the different indicators and control the further SDS choices by removing or adding dialogue alternatives.

Biographical Sketch:

Philippe Bretier heads the Natural Dialogue research group at Orange Labs, Lannion, France. He obtained Computer Science Engineering and Artificial Intelligence Master Degrees from the University of Rennes and received a PhD for his work on automatic reasoning on interaction theory for dialogue management from University of Paris Nord (1995). He joined France Telecom - Orange as a research engineer and developed the Artimis BDI-agent based dialogue system. Amongst work in research to industry transfers within the France Telecom - Orange operational and commercial divisions, he focused his research on industrial automata-based dialogue technology enhancement. His recent research interests include dialogue evaluation metrics and methodologies, dialogue design tools including usage analytics and feedbacks, machine learning, and modular architectures in dialogue systems.

Sebastian Möller

Affiliation: DE Telecom

Presentation Abstract:
Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin. Evaluating the quality and usability of spoken dialogue services is usually a cumbersome and expensive process. Whereas academic researchers are interested in quantifying the advances their research has made, industrial evaluations are commonly restricted to an absolute minimum, due to time and budget restrictions. This may however lead to low usability of the resulting commercial services. In my talk, I will address both academic and industrial approaches to evaluate and improve spoken dialogue services. I will highlight the current research topics of Deutsche Telekom Labs, a public-private partnership between Deutsche Telekom AG and TU Berlin, in the field of spoken dialogue systems.

Biographical Sketch:

Sebastian Möller was born in 1968 and studied electrical engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). From 1994 to 2005, he held the position of a scientific researcher at the Institute of Communication Acoustics (IKA), Ruhr-University Bochum, and worked on speech signal processing, speech technology, communication acoustics, as well as on speech communication quality aspects. 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.

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'Intélligence 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. In September 2008, we worked as a Visiting Fellow at MARCS Auditory Laboratories, University of Western Sydney (Australia) on the evaluation of avatars.

Sebastian Möller was awarded the GEERS prize in 1998 for his interdisciplinary work on the analysis of infant cries for early hearing-impairment detection, the ITG prize of the German Association for Electrical, Electronic & Information Technologies (VDE) in 2001, the Lothar-Cremer prize of the German Acoustical Association (DEGA) in 2003, and a Heisenberg fellowship of the German Research Foundation (DFG) in 2005. Since 1997, he has taken part in the standardisation activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals. He is currently acting as a Rapporteur for question Q.8/12.

Tim Paek

Affiliation: Microsoft Research

Presentation Abstract:

Current Directions in Dialogue Research at Microsoft. In this talk, I will provide a broad overview of current dialogue research at Microsoft as well as products and services offered by Microsoft that involve speech and dialogue technologies. I will also share my opinions about the opportunities and challenges of conducting dialogue research at an industry lab.

Biographical Sketch:

Dr. Tim Paek is a researcher in the Machine Learning and Applied Statistics group at Microsoft Research in Redmond, Washington. His primary research focus is on spoken dialogue systems, including both technologies that drive interaction (e.g., language modeling, dialogue management) as well as user interface design and experience. Since joining Microsoft Research in 2001, he has helped shipped several products, including Voice Command 1.6, Live Search Mobile, and the Windows Vista speech interface. He currently serves as the President of the Special Interest Group on Discourse and Dialogue (SIGDIAL) for Association for Computational Linguistics (ACL) and International Speech Communication Association (ISCA).

Invited Panelists

Jason Williams

Affiliation: AT&T

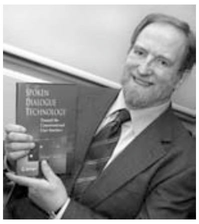
Biographical Sketch:

Jason Williams is Principal Member of Technical Staff at AT&T Labs Research, where he has been since 2006. He received a BSE in Electrical Engineering from Princeton University in 1998, and at Cambridge University he received an M Phil in Computer Speech and Language Processing in 1999 and a Ph D in Information Engineering in 2006. His main research interests are dialogue management, the design of spoken language systems, and planning under uncertainty. He is currently Editor-in-chief of the IEEE Speech and Language Processing Technical Committee's Newsletter. He is also on the Science Advisory Committee of SIGDIAL. Prior to entering research, he built commercial spoken dialogue systems for Tellme Networks (now Microsoft), and others. He also served as a consultant with McKinsey & Companys Business Technology Office.

Michael McTear

Affiliation: University of Ulster

Biographical Sketch:



Michael McTear is Professor of Knowledge Engineering at the University of Ulster. 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). 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).

Michael McTear has delivered keynote addresses at many conferences and workshops, including the EU funded DUMAS Workshop, Geneva, 2004, the SIGDial workshop, Lisbon, 2005, the Spanish Conference on Natural Language Processing (SEPLN), Granada, 2005, and has delivered invited tutorials at IEEE/ACL Conference on Spoken Language Technologies, Aruba, 2006, and ACL 2007, Prague. He was the recipient of the Distinguished Senior Research Fellowship of the University of Ulster in 2006. He is a member of the Scientific Advisory Group of SIGDial.

Frameworks and Grand Challenges for Dialogue System Evaluation: Speakers

Alan W. Black

Affiliation: Carnegie Mellon University

Biographical Sketch:

Alan W. Black is an Associate Professor in the Language Technologies Institute at Carnegie Mellon University. He previously worked in the Centre for Speech Technology Research at the University of Edinburgh, and before that at ATR in Japan. He is one of the principal authors of the free software Festival Speech Synthesis System, the FestVox voice building tools and CMU Flite, a small footprint speech synthesis engine. He received his PhD in Computational Linguistics from Edinburgh University in 1993, his MSc in Knowledge Based Systems also from Edinburgh in 1986, and a BSc (Hons) in Computer Science from Coventry University in 1984. Although much of his core research focuses on speech synthesis, he also works in real-time hands-free speech-to-speech translation systems (Croatian, Arabic and Thai), spoken dialogue systems, and rapid language adaptation for support of new languages. Alan W Black was an elected member of the IEEE Speech Technical Committee (2003-2007). He is currently on the board of ISCA and on the editorial board of Speech Communications. He was program chair of the ISCA Speech Synthesis Workshop 2004, and was general co-chair of Interspeech 2006 – ICSLP. In 2004, with Prof Keiichi Tokuda, he initiated the now annual Blizzard Challenge, the largest multi-site evaluation of corpus-based speech synthesis techniques.

Dan Bohus

Affiliation: Microsoft Research

Biographical Sketch:

Before joining Microsoft Research, Dan Bohus received his Ph.D. degree (2007) in Computer Science from Carnegie Mellon University. His dissertation work, supervised by Alex Rudnicky and Roni Rosenfeld was focused on developing techniques for increasing the robustness and reliability of spoken language interfaces. As part of this work, he developed a task-indepedent dialogue management framework called RavenClaw, that has been used to build and successfully deploy various dialogue systems spanning different domains and interaction types. Prior to CMU, Dan Bohus obtained a B.S. degree in Computer Science from Politehnica University of Timisoara, Romania.

List of Participants (ordered by last name)

Timo Baumann
University of Potsdam
Germany

Luciana Benotti
LORIA/INRIA
France

José Luis Blanco Murillo
Univ Politécnica de Madrid
Spain

Dan Bohus
Microsoft Research
USA

Okko Buss
Potsdam University
Germany

Dmitry Butenkov
Deutsche Telekom Labs
Germany

Caroline Clemens
Deutsche Telekom Labs
Germany

Klaus-Peter Engelbrecht
Deutsche Telekom Labs
Germany

Arash Eshghi
Queen Mary Univ of London
United Kingdom

Milica Gašić
University of Cambridge
United Kingdom

Florian Gödde
Berlin Institute of Technology
Germany

Christine Howes
Queen Mary Univ of London
United Kingdom

Srinivasan Janarthanam
University of Edinburgh
United Kingdom

Theodora Koulouri
Brunel University
United Kingdom

Catherine Lai
University of Pennsylvania
USA

Marianne Laurent
Orange Labs
Technopôle Brest-Iroise
France

Beatriz López Mencía
Univ Politécnica de Madrid
Spain

Syaheerah Lutfi
Univ Politécnica de Madrid
Spain

François Mairesse
University of Cambridge
United Kingdom

Toyomi Meguro
NTT Communication Science Labs, NTT Corporation
Japan

Teruhisa Misu
Nat Institute of Information & Communications Tech (NICT)
Japan

David Díaz Pardo de Vera,
Univ Politécnica de Madrid
Spain

Mara Reis
Federal Univ of Sta Catarina
Brazil

Nur-Hana Samsudin
University of Birmingham
United Kingdom

Niels Schütte
Dublin Institute of Technology
Ireland

Vasile Topac
Politehnica Univ of Timişoara
Romania

Ina Wechsung,
Deutsche Telekom Labs
Germany

Lluís Mas Manchón
Univ Autonoma de Barcelona
Spain

Gregory Mills
Queen Mary Univ of London
United Kingdom

Joana Paulo Pardal
L^2F INESC-ID
IST Tech Univ Lisbon
Portugal

Florian Pinault
LIA - University of Avignon
France

Sylvie Saget
LLI-IRISA
France

Alexander Schmitt
University of Ulm
Germany

Álvaro Sigüenza Izquierdo
Univ Politécnica de Madrid
Spain

Bogdan Vlasenko
Otto-von-Guericke University Magdeburg
Germany

Charlotte Wollermann
University of Bonn
Germany

Timo Baumann

University of Potsdam
Department for Linguistics
Karl-Liebknecht-Str. 24-25
D-14476 Golm

timo@ling.uni-potsdam.de
www.ling.uni-potsdam.de/~timo

1 Research Interests

My research is geared towards **interaction management** in spoken dialogue systems. Specifically, I am interested in the **timing** of dialogue and dialogue-related phenomena, and in taming dialogue systems to respond as quickly and in similar ways as humans. For a dialogue system to react to a user, while she is still speaking, it is necessary for the system to run **incrementally**, that is, to process the user's utterance while it is ongoing, and to come up with partial conclusions about what the user is saying, including how certain the system is about these conclusions or even predictions. I believe, that **prosody** analysis plays a vital role in improving incremental spoken dialogue systems' performance, as it conveys information that otherwise may only become clear later in the utterance.

1.1 Incremental Processing

In a modular system, an *incremental module* is one that generates (partial) output while input is still ongoing and makes these available to other modules. I have thoroughly investigated incremental ASR which outputs word hypotheses while recognition is still ongoing (Baumann et al., 2009a). We developed measures to be able to assess the incremental behaviour of ASR. These deal with how often hypotheses change (every change means that consuming modules have to re-process their input) and others describe timing properties of when words are first considered and first finally recognised by the ASR. I showed influences between optimising timing and change measures and derived a measure of certainty from the different timing measures. Having assessed incremental properties of ASR, I analysed methods to improve incremental behaviour by simple (and generic) filtering approaches (Baumann et al., 2009a).

Together with my colleagues, we applied the work on evaluation of incremental components to semantic interpretations (Atterer et al., 2009), the evaluation of incremental reference resolution (Schlangen et al., 2009), and to n-best processing in both ASR and semantic interpretation (Baumann et al., 2009b). As part of our venture into incremental analysis, we built a toolkit to process and visualise incremental data (Malsburg et al., 2009).

1.2 Prosody Analysis

Prosodic analysis of spoken input into a dialogue system is vital, if understanding and behaviour should depend on more than just words. We have investigated end-of-turn (EoT) prediction (Baumann, 2008) and end-of-utterance (EoU) detection (Atterer et al., 2008) using rather crude acoustic-prosodic features. While the need for EoT detection is obvious, EoU detection can be important for syntactic processing and semantic interpretation in more complex user turns.

Recently I have come back to the topic of prosody analysis, implementing a more phonologically sound incremental model of prosodic analysis (Baumann, 2009) that integrates with ASR output and generates word- and syllable-relative information which should be helpful in the NLU component of an SDS.

1.3 Incremental SDS

To show the end-to-end benefits of incrementality in spoken dialogue systems, it's best to show example systems. In (Baumann, 2008) I presented a system for turn-taking simulation, which stress-tests end-of-turn prediction. Two artificial dialogue participants converse with each other; the resulting turn-taking behaviour is similar to human-human dialogue.

Currently, I am working on a multi-modal command-and-control game, that exploits the timing advantages of incremental ASR with associated hypothesis certainties over regular incremental ASR (Baumann, 2009).

1.4 Future Work

Currently in a spoken dialogue system, good timing means "as quickly as possible". But with quicker reasoning and sufficiently accurate predictive processing coming up, good timing may become more specific. I would like to further analyse gaps and pauses in dialogue (at turn-boundaries and within turns) and find out whether onset-timings are related (and in which way) to rhythm patterns across speakers, so that these can be taken into account for system utterances.

2 Future of Spoken Dialogue Research

Currently deployed SDSs are mostly tailored towards information access and simple tasks. To some extent they can be seen rather as VUIs than as full dialogue.

I believe that deployed dialogue systems will appear as **conversational assistants** in hospitals and for elderly people, in tutoring (not only for foreign language learning, but in all areas), as characters in computer games and as more natural user interfaces for general-purpose personal digital assistants in smartphones.

Computer games make for an especially interesting sandbox for advanced SDSs, as domains are controlled and consequences of errors are small, while demand for "new stuff" and enthusiastic users are plentiful.

While human-like behaviour is not needed or could even distract in simple task-oriented systems, human-like, **intuitive behaviour** becomes more important for future applications, as they will be less recognised as tools but as real interlocutors. For better intuitivity, interaction behaviour (turn-taking, and -yielding, understanding and hinting below the content level) must be improved.

A further ingredient to better intuitivity is **adaptivity**. On the linguistic side, there is **entrainment** that should be followed (or deliberately broken) on all levels but cannot simply be ignored. On the non-linguistic side, the integration of different knowledge sources (the user's or her peer-group's calendar, previous e-mails, overheard conversations, . . .) will lead to better understanding.

In terms of SDS design, better modelling of control flow in the system, (no central control, but patterns for distributed control), and theories of what is possible under which architectural constraints, are needed.

3 Suggestions for Discussion

- *How to learn interaction behaviour from corpora?*
  Most often, a range of legal interaction decisions exist. Albeit, in a corpus of recorded dialogue, all but one decision have been blocked by the one that has been taken. How do we learn from such noisy data? How do we now, which decision (among an infinite set) is best or at least good enough?
- *Availability of data (and systems) for independent result verification!*
  Reproducibility of results and independent verification is standard in scientific work. Not so in computational linguistics where people are happy to develop different methods on different data and all claim to have found the one correct solution.
- *People are afraid of too-human machines.*
  When asked, people want to know whether they are interacting with machines. At the same time, they don't care that traffic lights have been computer-controlled for decades.

References

Michaela Atterer, Timo Baumann, and David Schlangen. 2008. Towards Incremental End-of-Utterance Detection in Dialogue Systems. In *Proceedings of COLING 2008*, Manchester, UK.

Michaela Atterer, Timo Baumann, and David Schlangen. 2009. No Sooner Said Than Done? Testing the Incrementality of Semantic Interpretations of Spontaneous Speech. In *Proceedings of Interspeech 2009*, Brighton, UK.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009a. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Timo Baumann, Okko Buß, Michaela Atterer, and David Schlangen. 2009b. Evaluating the Potential Utility of ASR N-Best Lists for Incremental Spoken Dialogue Systems. In *Proceedings of Interspeech 2009*, Brighton, UK.

Timo Baumann. 2008. Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking. In *Proceedings of the 13th ESSLLI Student Session*, Hamburg, Germany.

Timo Baumann. 2009. Integrating Prosodic Modelling with Incremental Speech Recognition. In *Proceedings of DiaHolmia (SemDial 2009)*, Stockholm, Sweden.

Titus von der Malsburg, Timo Baumann, and David Schlangen. 2009. TELIDA: A Package for Manipulation and Visualisation of Timed Linguistic Data. *Submitted.*

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. *Submitted.*

Biographical Sketch

Timo Baumann is a research assistant and PhD candidate at the University of Potsdam, Germany, working under the supervision of David Schlangen.

He previously studied computer science, phonetics and linguistics at Hamburg University and received his master's degree in 2007 for work on prosody analysis carried out at IBM Research Germany.

In his free time, Timo likes to go hiking or cycling and sings in a choir. Also, he enjoys to stay abroad and live on student grants (USA 1997, Switzerland 2003, Spain 2005). Continuing in this spirit, Timo is currently a guest researcher at KTH Stockholm's speech group, working with Jens Edlund on incremental prosody analysis.

Luciana Benotti

LORIA/INRIA Nancy Grand Est
615, rue du Jardin Botanique
54600 Villers les Nancy, France
luciana.benotti@loria.fr
www.loria.fr/~benottil/

1 Research Interests

More is involved in what one communicates than what one literally says; more is involved in what one means than the standard, conventional meaning of the words one uses. For instance, if I ask you to go to the cinema and you reply, ”I have a presentation tomorrow I'm not prepared for.” You have conveyed to me that you will not be coming to the cinema, although you haven't literally said so. You intend for me to figure out that by indicating a reason for not coming (the need to prepare your presentation) you intend to convey that you are not coming for that reason. The study of such **conversational implicatures**, and how they can be modelled computationally in a dialogue system is the subject of my research; which crucially involves the characterisation of the **context-representation** and **means-ends reasoning-tasks** involved. In other words, my research interests lie in the area of **Computational Pragmatics**.

1.1 Conversational implicatures as tacit acts

Modelling how listeners draw inferences from what they hear is a basic problem for theories of understanding natural language. An important part of the information conveyed is inferred in context, given the nature of conversation as a goal-oriented enterprise; as illustrated by the following classical example by Grice:

(1) A: I am out of petrol.
B: There is a garage around the corner.
$\rightsquigarrow$ B thinks that the garage is open and selling petrol.
(Grice, 1975, p.311)

B's answer **conversationally implicates** ($\rightsquigarrow$) information that is relevant to A. In Grice's terms, B made a **relevance implicature**: he would be flouting the conversational maxim of relevance unless he believes that it's possible that the garage is open and has petrol to sell.

I view such relevance implicatures as **tacit dialogue acts** that fill the gap between the context and the utterance (Benotti, 2008a). The tacit dialogue act is illustrated below in italics.

(2) A: I am out of petrol.
B: I know a garage that is open and selling petrol.
B: There is (such) a garage around the corner.

Human-human dialogue makes evident that such tacit dialogue acts are inferred and processed by the dialogue participants; as we argue in (Benotti, 2009a) presenting evidence from a dialogue corpus called SCARE (Stoia et al., 2008). A can well reject or ask clarifications on the tacit act content, as if the tacit act has been explicitly said. For instance, A can continue the exchange with follow-ups such as (a) or (b) below:

(3) a. A: No, I checked, it's closed.
b. A: Why do you think it's open at 4am?

I've integrated the inference and processing of (some kinds of) such conversational implicatures in two dialogue systems. I briefly describe each of them in the following section.

1.2 Conversational implicatures in dialogue systems

The first system to which I added the inference of tacit acts is called Frolog (Benotti, 2009b). Frolog implements a text-adventure game and the dialogue acts that it can handle are instructions that can be directly mapped into physical actions in the simulated game world. The inference of tacit acts requires **means-ends reasoning** on the possible dialogue acts and the representation of the **dialogue context**. The necessary reasoning tasks have been implemented integrating two planners into Frolog. The first planner that I used was Blackbox (Kautz and Selman, 1999) which is fast and **deterministic**; how a classical planner can be used to infer the conversational implicatures of an utterance is discussed in (Benotti, 2007). The other planner is called PKS (Petrick and Bacchus, 2004); PKS can reason over **non-deterministic acts**. For detailed discussion and examples including non-deterministic acts see (Benotti, 2008b).

Finally, I added support for some kinds of conversational implicatures to the ICT Virtual Human[1] dialogue manager (Traum et al., 2008). This dialogue manager allows virtual humans to participate in multi-party conversations in a variety of domains with dialogue acts at different levels. The dialogue manager is embedded within the Soar cognitive architecture (Laird et al., 1987), and decisions about interpreting and producing speech compete with other cognitive and physical operations. Soar

[1] http://ict.usc.edu/projects/virtual_humans

provides support for some kinds of means-ends reasoning useful for inferring tacit acts. However, the architecture of the dialogue manager is complex and the work presented in (Benotti and Traum, 2009) constitute just a first step that still needs to be properly evaluated.

2 Future of Spoken Dialogue Research

I think that a great challenge for this generation of young researchers is to start building dialogue systems that remember particular users and use that information to adapt in the following interactions with the same user (or group of users). This challenge could not only be useful application-wise but it could also have an impact in the theories of dialogue. I adhere to the approach that view conversational implicatures as imposing pressure on linguistic structures; pressure that many times leaves its imprint in the literal meaning of words. Studying semantic change in situated conversation, would make the concepts of semantic coordination and implicatures more concrete, and also clarify the relations between them.

This idea of learning from previous interactions is not new and it's the approach used by researchers such as those that apply reinforcement learning to learning dialogue policies. However, such approaches learn many different kinds of information at the same time including task dependent and task independent information. Moreover, since the search space they are dealing with grows with the amount of features that they include in each state, its information states are necessarily coarse-grained. What seems to be needed is to restrict the state space using some notion of coherence of sequences of acts in the state space. In other words, I believe that symbolic and statistical methods would need to be combined. If such a combination is possible, then systems that use expectation-based interpretation on a large scale seem not so far away. That is, systems that try to infer the next possible moves that the user can make and interpret the next utterance taking into account these possible moves. That we are approaching the point when we can begin to think about ways of combining such methods makes the next few decades in dialogue systems research seem very exciting indeed.

3 Suggestions for Discussion

Three possible topics for discussion:

- Challenges in combining reinforcement learning and symbolic methods for building adaptable dialogue systems.
- Expectation based interpretation and performance of NLU in dialogue systems.
- Using expectations for asking task-related questions to the user instead of signalling misunderstanding.

References

L. Benotti and D. Traum. 2009. A computational account of comparative implicatures for a spoken dialogue agent. In *Eighth Int. Conf. on Computational Semantics (IWCS-8)*.

L. Benotti. 2007. Incomplete knowledge and tacit action: Enlightened update in a dialogue game. In *The 2007 Workshop on the Semantics and Pragmatics of Dialogue (DECALOG)*.

L. Benotti. 2008a. Accommodation through tacit dialogue acts. In *Conf. on Semantics and Modelisation (JSM08)*.

L. Benotti. 2008b. Accommodation through tacit sensing. In *The 2008 Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL)*.

L. Benotti. 2009a. Clarification potential of instructions. In *The 10th Annual SIGDIAL Meeting on Discourse and Dialogue*, London, United Kingdom.

L. Benotti. 2009b. Frolog: An accommodating text-adventure game. In *The 12th Conf. of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

P. Grice. 1975. Logic and conversation. *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

H. Kautz and B. Selman. 1999. Unifying SAT-based and graph-based planning. In *Proc. of the IJCAI*, pages 318–325.

J. Laird, A. Newell, and P. Rosenbloom. 1987. SOAR: an architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64, September.

R. Petrick and F. Bacchus. 2004. Extending the knowledge-based approach to planning with incomplete information and sensing. In *Proc. of the Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 613–622.

L. Stoia, D. Shockley, D. Byron, and E. Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proc. of LREC*.

D. Traum, W. Swartout, J. Gratch, and S. Marsella. 2008. A virtual human dialogue model for non-team interaction. *Recent Trends in Discourse and Dialogue*. Springer.

Biographical Sketch

Luciana Benotti is currently a third year PhD Student at INRIA Nancy grand Est, in France. She is a member of the TALARIS Team[2] where she works under the supervision of Patrick Blackburn. She spent last summer at the Institute for Creative Technologies (ICT) of the University of Southern California (USC) working with David Traum. In 2006, she obtained an European Masters in Computational Logic (from the Free University of Bolzano in Italy and the Polytechnic University of Madrid in Spain). She also has a Masters in Computer Science from the National University of Comahue in Argentina (where she was born). She likes travelling, meeting people from different cultures and cooking. She enjoys hot weather, good food and life in the open air.

[2] http://talaris.loria.fr

José Luis Blanco Murillo

Universidad Politécnica de Madrid
ETSI de Telecomunicación
Avda. Complutense 30
28040 Madrid, Spain

jlblanco@gaps.ssr.upm.es

1 Research Interests

My main research interests lie in **spoken dialogue systems** (SDSs), incorporating **additional information from heterogonous sources**, focusing on the improvement in **representation** and **management** of this additional knowledge.

1.1 Past and Ongoing research

During the last decade spoken dialogue systems have had an enormous progression. From the first automatic speech recognition systems to the present conversational agents and interaction systems, our understanding of the nature of interaction has grown widely, allowing us to develop new standards and systems. Although there are still many questions to be answered, it is clear that the quality of our systems has improved noticeably, in terms of robustness and flexibility. However, new scenarios and applications have arisen for which there isn't a common concern about the way vocal interfaces have to be developed, neither the way information is to be handled.

At GAPS (the Signal Processing Applications Group at Universidad Politécnica de Madrid) we have been repeatedly leading with this problem during the last years we have overcome this same problem in various projects. As a matter of fact, each of those projects was meant to focus on a different aspect of the interaction. However, they all had something in common: vocal interaction was included to enhance human-machine interaction.

1.2 The in-vehicle interaction scenario

The main scenario we have been working out is the spoken interaction within the automobile. Our group has been collaborating in various publicly founded projects, both national and European, concerning this particular environment. Our main goal is, and has been, to analyse the opportunities that arise when we introduce vocal interfaces into the car, as well as the additional constraints this particular scenario imposes to the dialogue. The balance between those new possibilities and their limitations makes it an extremely open and challenging environment.

Several studies have been addressed related with this particular problem of the in-car interaction. However, there is a deep gap between academic improvements and the actual commercial applications. The most remarkable references we have had news of are those delivered from the European project TALK. In this consortium both industry and universities have considered several aspects about the in-vehicle interaction from a common perspective which is certainly the best way to bridge the gap (Becker et al., 2007). Their description of the problem human-machine interaction faces in the driving context, as well as their analysis of the alternatives the present state-of-the-art can provide, have taught us that it is worth to consider each situation specifically, rather than to ignore it's peculiarities. Since there isn't a common concern about this need, software developers are still reusing previous systems in quite different contexts from those for which they were designed without reworking those spoken dialogue systems as they require (Blanco et al., 2009).

Moreover, in this scenario most efforts have been directed towards the improvement of robustness in speech recognition systems, due to its particular noisy conditions. The adaptation of the available base technologies is certainly a key, highly correlated with users' perception of systems' performance, but there is much more to do. Adapting the interaction strategies and policies to the particular circumstances of the interaction is required if we want to achieve a successful improvement. Finally, the combination of both will help us to reduce the interference the spoken interaction might cause on the driving.

1.3 Reworking spoken dialogue systems

Reworking spoken dialogue systems has to be faced not just from this particular point for view of the in-vehicle scenario, but from a broader perspective.

Sources of information appear almost anywhere and are extremely heterogeneous: from social relations (i.e. contacts list or social agenda), to localisation, travelling speed, or users actual physical and emotional state. It seems reasonable to believe there might be some kind of correlation between interaction and its circumstances when thinking about those examples. For instance, if we know that a certain vocal interface is been accessed from a mobile phone, or from a land line, it might be useful to change the noise models in the speech recognition system. Or if the user is known to suffer from some kind of disease affecting the vocal cords, it might be useful to use other models or to ad-apt those being used. But also the dialogue might have to be adapted. If a user calls from a hands-free phone while driving, as his/her interaction abilities will certainly be diminished, shorter locutions from the speech synthesizer are recommended (Nishimoto et al., 2005).

There are just a few studies focusing on these particular aspects in a deep and sensible way. Actually the best alternative is to continue with the design of specific solutions, but obviously this isn't the best way to face the problem. A homogeneous framework for the available information will certainly help to adapt our systems in an efficient and reliable way.

2 Future of Spoken Dialogue Research

I believe we will soon have at our disposal efficient technologies for the representation of heterogeneous information, which will bring us the opportunity to model the variables affecting the curse of interaction. This will help us adapt our designs to make them much more flexible under adverse circumstances. Adapting to the particular circumstances of the interaction will certainly take longer. Though the present technologies are capable to do this after an extensive redefinition, new technologies and tools are to come which will help us design those new systems.

On the other hand, commercial solutions are improving their performance ratios and approaching the developments from the academic realm. Standardising those results and their description of the interaction beyond the present VoiceXML v.2 standard, as well as the representation of the context variables will bring the opportunity to finally bridge that gap. This will be take long, but will get our actual designs out of their knowledge bubble and into a see of shared knowledge.

The development of new methodologies to aim designers in their task will simplify the integration of systems at any level of the interaction. Open architectures such as the Olympus/Ravenclaw (Bohus et al., 2007) have provided an excellent precedent of this and shall be preserved. This same idea, but from a quite different perspective will bring to us new systems able to check the topics of the communication to decide how the interaction should be carried out. Those systems which multiplex various singular spoken dialogue systems designed for a certain task (Kerminen and Jokinen, 2003) will shed some light about the way in which systems should be integrated.

3 Suggestions for Discussion

My suggestions for discussion are the following:

- How could we combine the representation of the circumstances together with the interaction in a way that can be easily used to adapt our systems?
- How should we adapt interaction strategies to take advantage of this new knowledge?
- Standardisation of experiments for analysing correlation between dialogue progress and context.
- SDSs implemented as multiplex of singular conversational agents.

References

T. Becker, N. Blaylock, C. Gerstenberger, A. Korthauer, N. Perera, M. Pitz, P. Poller, J. Shehl, F. Steffens, R. Stegmann and J. Steigner. 2007. D5.3: In-Car Showcase Based on TALK Libraries. Talk and Look: Tools for Ambient Linguistic Knowledge. *IST-507802 Deliverable 5.3.*

J. L. Blanco, A. Sigüenza, D. Díaz, M. Sendra and L. A. Hernández 2009. Reworking spoken dialogue systems with context awareness and information prioritisation to reduce driver workload. In *Proceedings of the NAG-DAGA 2009*, International Conference in Acoustics.

T. Nishimoto, M. Shioya, J. Takahashi and H. Daigo. 2005. A study of dialogue management principles corresponding to the driver's workload. In *Biennial Workshop on Digital Signal Processing for In-Vehicle and mobile systems*, Sesimbra, Portugal.

D. Bohus, A. Raux, T. K. Harris, M. Eskenazi and A. I. Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Bridging the Gap: Academic and Industrial Research in Dialog Technology* workshop at HLT/NAACL 2007

A. Kerminen and K. Jokinen. 2003. Distributed Dialogue Management in a Blackboard Architecture. In *Proceedings of the EACL Workshop Dialogue Systems: interaction, adaptation and styles of management*, Budapest, Hungary. pp. 55–66.

Biographical Sketch

José Luis Blanco Murillo has a MEng in telecommunications engineering from Universidad Politécnica de Madrid. He is currently a PhD student and holds a research position in this same university, under the supervision of Dr. Luis Hernández.

Dan Bohus

Microsoft Research
One Microsoft Way
Redmond, WA, 98052
USA

dbohus@microsoft.com
research.microsoft.com/~dbohus

1 Research Interests

My research interests lie in the area of **situated interaction** and **open-world spoken dialogue systems**. The central question driving my long-term research agenda is: how can we develop systems that naturally embed interaction and computation deeply into the flow of everyday tasks, activities and collaborations? More specifically, some of the areas and problems that I am currently investigating are: **multimodal sensor fusion**, **conversational scene analysis**, **situated and open-world dialogue**, **turn-taking and engagement models**, **self-supervised (lifelong) learning and adaptation**.

1.1 Previous work

My dissertation work (Bohus, 2007) has focused on issues of **dialogue management** and **error handling** in spoken dialogue systems. As part of this work, I have developed RavenClaw (Bohus and Rudnicky, 2003; 2009), an open-source, reusable dialogue management framework, that has since been used to develop several spoken dialogue systems spanning different domains and interaction types. With respect to error handling, my dissertation addressed the following questions: (1) how can a system reliably detect potential errors? (Bohus and Rudnicky, 2005a; 2006; Bohus 2007), (2) what strategies can be used to recover from different types of errors? (Bohus and Rudnicky, 2005b), and (3) how should a system choose between multiple such strategies at runtime? (Bohus et al., 2006).

1.2 Current research

Most research to date on spoken dialogue systems has focused on the study and support of interactions between a single human and a computing system within a constrained, predefined communication context. Efforts in this space have led to the development and wide-scale successful deployment of telephony based, and more recently multimodal mobile applications. At the same time, numerous and important challenges in the realm of situated spoken language interaction remain to be addressed.

In particular, my current research is focused on the challenges of developing spoken language interfaces that operate continuously in open, dynamic, relatively unconstrained environments. The question that drives this long-term research agenda is: how can we develop systems that embed interaction and computation deeply into the natural flow of everyday tasks, activities and collaborations? Examples include interactive billboards in a mall, robots in a home environment, interactive home control systems, interactive systems providing assistance during procedural tasks, intelligent tutoring systems, etc.

Interaction in the open environments is characterised by two aspects that capture key departures from assumptions traditionally made in spoken dialogue systems (Bohus and Horvitz, 2009c). The first one is the *multiparty and dynamic* nature of the interaction, *i.e.*, the world typically contains not just one, but multiple agents that are relevant to the interactive system; agents may come and go, and their goals, needs, and plans change in time; relevant events might happen asynchronously with respect to the natural flow of dialogue. A second important aspect is the *physically situated* nature of these systems, *i.e.* the fact that the physical surroundings provide a continuously streaming, rich context relevant for organising and conducting the interactions. I believe these aspects of open-world interaction raise interesting research challenges and bring new dimensions to existing dialogue problems, such as engagement, turn-taking, language understanding, dialogue management and output planning.

As an example, consider the problem of establishing and maintaining engagement with an interactive system. In single-user, closed-world dialogue systems this problem often finds a relatively trivial solution. For instance, in telephony-based applications, engagement can be safely assumed once a call has been picked up by the system. Similarly, a push-to-talk button provides a clear engagement signal in multimodal mobile applications. While these solutions are appropriate, perhaps even natural in those contexts, they are insufficient for systems that must operate continuously in open, dynamic environments, where multiple participants may enter and leave conversations, and interact with the system and with each other. Such systems should ideally be able to fluidly engage, disengage, or re-engage with one or multiple participants, whether the participants are close by or at a distance, whether they have a standing plan to interact with a system, or opportunistically decide to do so, in-stream with other ongoing activities. To successfully manage engagement in this context, a system must continuously monitor its surroundings and fuse incoming sensory streams to infer engagement states, actions, and intentions of various participants, must be able to make engagement control decisions while obeying rules of etiquette and social interaction, and ultimately render these decisions in an appropriate set of behaviors.

This is just one example problem that lives at the lowest (Channel) level of grounding in communication. Similar new challenges arise however in turn-taking, spoken language and discourse understanding, interaction planning, etc. In (Bohus and Horvitz, 2009c) we have identified a number of such challenges and outlined a set of core competencies required for open-world spoken language interaction. In the Situated Interaction (2009) project at Microsoft Research, we have developed several real-world situated conversational agents, and are using them as an experimental platform for pursuing these challenges. Initial work in this space including a computational model for managing situated multiparty engagement (Bohus and Horvitz, 2009c) and an implicit learning approach for detecting engagement decisions (Bohus and Horvitz, 2009c) have been reported in SIGDial'2009. Current efforts are focused on tracking conversational dynamics in multiparty interactions and developing multi-participant turn-taking models.

2 Future of Spoken Dialogue Research

I believe that in the next 5–10 years more attention will shift towards issues in multi-modal, embodied and situated interactive systems, of the type described in the previous section. The challenges that lie ahead of us are exciting and many: situation awareness, multi-modal sensor fusion, scene analysis, behavior and intention recognition, situated dialogue management, situated grounding, engagement models, mixed-initiative and multi-participant interaction, life-long, open-domain and open-world learning and adaptation.

3 Suggestions for Discussion

- **dialogue systems and robots.** Discuss applicability and limitations of current dialogue technologies in the context of human-robot interaction (*e.g.,* multiple sensors, asynchronous events, multiple participants, etc.)

- **challenge problem (s) and evaluation.** Propose and discuss one or more challenge problem(s) for the field, and a corresponding evaluation process.

References

D. Bohus and E. Horvitz 2009c. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. In *Proc. of SIGDial'09*, London, UK.

D. Bohus and E. Horvitz. 2009b. Models for Multiparty Engagement in Open-World Dialog. In *Proc. Of SIGDial'09*, London, UK.

D. Bohus and E. Horvitz. 2009a. Open-World Dialog: Challenges, Directions and Prototype. In *Proc. of KRPD'09*, Pasadena, CA.

D. Bohus and A. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. In *Computer Speech & Language* 23, Issue 3, pp.332–361.

D. Bohus 2007. Error Awareness and Recovery in Conversational Spoken Language Interfaces. *Ph.D. Thesis, CS-07-124*, Carnegie Mellon University, Pittsburgh, PA.

D. Bohus, B. Langner, A. Raux, A. Black, M. Eskenazi and A. Rudnicky. 2006. Online Supervised Learning of Non-understanding Recovery Policies. In *Proc. of SLT'2006*, Palm Beach, Aruba.

D. Bohus and A. Rudnicky. 2006. A K-hypotheses + Other Belief Updating Model. In *Proc. of AAAI Workshop on Statistical Methods in Spoken Dialogue Systems.*

D. Bohus and A. Rudnicky. 2005b. Sorry I didn't Catch That: An Investigation of Non-understanding Errors and Recovery Strategies. In *Proc. of SIGDial'2005*, Lisbon, Portugal.

D. Bohus and A. Rudnicky. 2005a. A Principled Approach for Rejection Threshold Optimization in Spoken Dialog Systems. In *Proc. of Interspeech'2005*, Lisbon, Portugal.

D. Bohus and A. Rudnicky. 2003. RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. In *Proc. of Eurospeech'2003*, Geneva, Switzerland. `http://www.ravenclaw-olympus.org`

Situated Interaction Project web page, 2009 — `research.microsoft.com/~dbohus/research_situated_interaction.html`.

Biographical Sketch

Dan is currently a researcher in the Adaptive Systems and Interaction group at Microsoft Research. His current research agenda is focused on situated interactive systems and open-world dialogue. Prior to joining Microsoft, Dan obtained his Ph.D. degree from Carnegie Mellon University, where he investigated problems of dialogue management and error handling in task-oriented spoken dialogue systems.

Okko Buss

Potsdam University
Karl-Liebknecht Str. 24-25
14415 Potsdam
Germany

okko@ling.uni-potsdam.de
www.ling.uni-potsdam.de/~okko

1 Research Interests

My research is concerned with incremental human-machine dialogue processing. An incremental SDS can process and react to user input while the user is still speaking. In particular I'm focussed on modelling **interaction and dialogue management** from an incremental point of view. I hold these to be two separate tasks (or groups of tasks) that require individual lines of inquiry.

I'm further interested in what **types of control** a speech dialogue processing system must embody to exhibit human or human-like behavior. I believe this to be a prerequisite for accurate cognitive modelling of spoken language behavior.

Lastly I am interested in **standardisation** of speech dialogue processing.

1.1 Incremental Interaction and Dialogue Management

Our mission in the InPro project at the University of Potsdam is to build computational models of incremental dialogue processing. Thus far, a great deal of effort has gone into achieving incrementality at the levels of speech-recognition (ASR) (Baumann et al., 2009) and semantic interpretation/natural language understanding (NLU) (Atterer et al., 2009) and related issues, such as n-best lists (Baumann et al., 2009b). These modules can best be understood as point-of-contact technologies between the system and the user. Achieving incrementality at these levels is a prerequisite for exploring "later" issues such as interaction (IM) and dialogue management (DM).

IM and DM are often treated as part of an overall dialogue planning module. This is particularly true in commercial dialogues systems, such as telephony applications built on standards like VoiceXML.[1]

While it makes sense for standardisation to keep IM and DM closely linked, there are conceptual and computational benefits to separating them. While the former is conceptually concerned with low-level events, such as producing and playing back system prompts, aggregating user input and keeping timing measures, the latter is responsible for high-level tasks such as verifying user intentions against a dialogue strategy, issuing clarification requests and updating its own world knowledge. A two-tiered approach that separates IM and DM along these distinctions is described in (Lemon et al., 2003).

Making IM and DM incremental shows the need for this separation even more clearly. Incrementality places far more stringent requirements on interaction aspects of communication, especially where timing is concerned. While a non-incremental SDS may only need two timeout variables to handle turn-taking - for in-speech and after-speech states - an incremental system can exhibit any number of in-between states, each of which requires its own kind of timeout behavior. Moreover, these timeouts may change dynamically in-utterance. We describe this aspect in (Baumann et al., 2009a) for our research prototype.

Other aspects of IM and DM to investigate from an incremental perspective will include confidence scoring, output generation and alternative hypothesis processing.

1.2 Control in SDS

An incremental SDS by nature has a much more loose definition of control than a classical pipelined architecture where control is passed from one module to the next (e.g. the ASR listens, passes its result to the NLU and so forth.) Incrementality however requires that parallel actions and asynchronous events can occur at any given time, something that such an architecture prohibits. Yet completely removing any notion of control is computationally and cognitively unsatisfying. The question of how control can be embodied in a parallel, distributed architecture thus arises.

I believe that understanding control for cognitive modelling is crucial to understanding intelligent behavior, a view adopted with modification from (Sloman, 2003). Rather than focussing solely on behavior, a proper taxonomy of types of control may be necessary for implementing intelligent behavior in parallel, asynchronous agents,

[1] Here, IM and DM are reducible to XML tags for low-level interaction and ECMA script variables for high-level world-knowledge.

including some components in SDSs.

1.3 Standards for SDS

Having worked extensively with commercial voice applications, I'm keenly interested in extending the groundwork performed in standardising aspects of SDS with the incremental, parallel processing approach described above. Some such aspects include prompt typology, semantic representation, confidence scoring and n-best list processing.

2 Future of Spoken Dialogue Research

Until recently, most deployed SDSs were largely highly task-specific applications geared at automation of otherwise labor-intensive tasks performed by human operators. While these kinds of applications will remain prominent (if unpopular), **mobile computing** has opened opportunities to move speech and language technology to novel areas. Three things about mobile computing need to be explored for development of SDSs in mobile environments.

Mobile speech applications that exist today make use of a mixed bag of tools for language processing. Some rely on local others on distributed resources for language processing. A first line of inquiry is thus harnessing these resources for SDSs in a meaningful way.

Second is the ubiquity of internet access. This makes speech-enabled mobile applications available to interface with a host of data sources. Making use of these will certainly need to be explored and requires moving towards a data-centric view of dialogue.

A third feature is multi-modality. Mixing traditional interfaces with voice in an intuitive fashion is a goal far from being reached and poses interesting research questions from computational, a cognitive load, and integration point of views.

In addition to mobile computing, **paralinguistic input sources** will have an effect on SDS research. Language and communication consist of more than spoken strings. SDSs stand much to gain from research on affect recognition, prosodic cueing and gesture recognition.

3 Suggestions for Discussion

- *Learnability of behavior from data.*
  Statistical approaches in ASR and NLU are taken for granted. But what means are available to train an SDS's interaction and planning modules from data. To what extend can an SDS function without being informed by symbols and rules.

- *Cross-disciplinary evaluation metrics.*
  Evaluation of SDSs involves comparing simulated outcomes with hand-annotated ones. Usually this concerns ASR and NLU output only. Is it possible to evaluate performance of SDSs in terms of metrics from other scenarios, such as psycholinguistic ones (e.g. eye-tracking for reference resolution)? How would an SDS fare compared to a human agent faced with an identical task.

- *Standardisation.*
  How relevant are standards to SDS research? Who uses them and to what effect?

References

Michaela Atterer, Timo Baumann, and David Schlangen. 2009. No sooner Said Than Done? Testing the Incrementality of Semantic Interpretations of Spontaneous Speech. To appear *Proceedings of Interspeech 2009*, Brighton, UK.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Timo Baumann, Michaela Atterer, and Okko Buss. 2009a. Incremental ASR, NLU and Dialogue Management in the Potsdam INPRO P2 System. Invited talk at the *Workshop on Incrementality in Verbal Interaction*, Bielefeld, Germany.

Timo Baumann, Okko Buss, Michaela Atterer, and David Schlangen. 2009b. Evaluating the Potential Utility of ASR N-Best Lists for Incremental Spoken Dialogue Systems. In *Proceedings of Interspeech 2009*, Brighton, UK.

Oliver Lemon, Lawrence Cavedon, and Barbara Kelly. 2003. Managing Dialogue Interaction: A Multi-Layered Approach. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.

Aaron Sloman. 1996. Beyond Turing Equivalence. In *Machines and Thought: The Legacy of Alan Turing*, 1:179–219.

Biographical Sketch

Okko Buss is a research assistant and PhD candidate at the University of Potsdam, Germany, working under the supervision of David Schlangen in the InPro project. He holds a B.A. in Linguistics from McGill University, Montreal, Canada, as well as an M.A. in Cognitive Science from UNSW, Sydney Australia. In the years prior to joining the team in Potsdam he held various positions at speech and language technology companies, such as Speechworks, Scansoft and Mundwerk building voice applications.

He also enjoys skiing, hiking and cooking.

Dmitry Butenkov

Deutsche Telekom Laboratories
Quality & Usability Lab
Berlin Institute of Technology
Ernst-Reuter-Platz 7
10587 Berlin Germany

dmitry.butenkov@telekom.de

1 Research Interests

My current research interests include statistical user simulation and modelling, decision making under uncertainty, uncertainty models and fuzzy applications in spoken dialogue systems. I share the research view-point of Berkley Initiative in Soft Computing. Besides that, I consider the Robotics and Machine Learning fields as tightly related to user simulation and I will try to make use of the results from both. I mostly work on an abstract, user intentions level of Human-Computer Interaction, trying to understand and formalise how the real users think and behave.

1.1 Past Research

I started my research career in early 2002 from the field of intelligent overtaking and navigation systems (Butenkov and Finaev 2003). The project addressed the optimal models of traffic streams in local city area.

Slightly after that, I worked in cooperation with Laboratory of Mathematical Problems in AI (based on TSURE) toward modern fuzzy methods and approaches to various natural problems formalisation and interpretation (Butenkov et al. 2004, 2005, 2008). Most of topics addressed image and scene analysis and decision making problems.

1.2 Current Research

The focus of my current research is data-driven user simulation for spoken dialogue systems evaluation. The plenty of attempts were made in this direction so far (Schatzmann et al. 2005; Pietquin, 2008), but re-search results indicate that considerable research work is still necessary in order to develop a flexible user simulation which would be valid for a range of systems and users (Chung 2004). However, evaluation is vital for systems which should be usable for real users. These facts explain growing interest in automatic, model-driven spoken dialogue system evaluation approaches in recent years (Möller et al., 2006; Ai and Weng, 2008; Eskenazi et al., 2008; Jung et al., 2009).

The specific attention is put to the usability assessment in such an evaluation. My current work is done within the frame of the SpeechEval project (Butenkov and Möller, 2009). This is a joint project between T-Labs[1] and the German Research Center for Artificial Intelligence (DFKI)[2], sponsored by the European Found for Regional Development[3].

Working on intentions level, we develop an inter-action paradigm that should allow us to simulate user behavior more naturally than classic approaches. The data-driven scope of the project is based on a large Voice Awards corpus. This data was collected during several years of open contest with real, deployed commercial German spoken dialogue systems (Steimel et. al, 2007).

The entire solution is based on an Ontology-based Dialogue Platform by SemVox GmbH[4] (Pfalzgraf et al., 2008). The Nuance Recognizer 9.0 by Nuance Communications[5] is used as high-quality ASR back-end.

1.3 Future Research

I plan to explore and extend the possibilities of modern user simulation for spoken dialogue systems. During the long time its role was underestimated. Thus, today it is mostly used as an assessment tool for spoken dialogue management development.

However, I would like to demonstrate the significant usefulness of user simulation as an evaluation tool and address the fundamental generalisation problem, i.e. the transferability of a simulation from one system (version) to another, and from one group of users to another.

2 Future of Spoken Dialogue Research

In my opinion, the information access and retrieval via spoken language will play a more and more important role in the near future. Databases become aggressively huge causing the rapid grow of the semantic web search field. I assume it is very attractive to design speech applications for this. Therefore, we can expect to have public

[1] www.t-labs.tu-berlin.de
[2] www.dfki.de
[3] europa.eu/scadplus/leg/de/lvb/l60015.htm
[4] www.semvox.de
[5] www.nuance.com

spoken voice search services based on e.g. Google Search or MSN engines in next 5-10 years

3 Suggestions for Discussion

Data-driven user simulation
Can we reliably test and evaluate modern commercial spoken dialogue systems with the help of a user simulation? Does the generalisation problem evaluating multiple spoken dialogue systems still hold (even within same domain)?

Complex natural problems formalisation
Is the reduction of communication and discourse phenomena to simple slot-filling an oversimplification (in terms of modern commercial SDS)? How to efficiently extract and formalise real user motivations?

The trends in the field from business point of view
Are state-of-the-art commercial spoken dialogue systems, in principle, much more complex then 10 years ago? Do the semantic technologies help to significantly improve those systems? What other technologies are useful or have potential impact?

References

H. Ai and F. Weng. 2008. User Simulation as Testing for Spoken Dialog Systems. In *Proceedings of the 9th SIGDial Workshop on Discourse and Dialogue*, Columbus, Ohio, USA.

D. Butenkov, S. Möller. 2009. Towards a Flexible User Simulation for Evaluating Spoken Dialogue Systems. In *Proceedings of INTERACT 2009*, Uppsala, Sweden.

D. Butenkov. 2008. Fuzzy geometric models of multidimensional data in intelligent data processing systems. In *Proceedings of 2nd National conference "Fuzzy Systems & Soft Computing"*, Uljanovsk, Russian Federation

D. Butenkov and V. Finaev. 2003. The development of intelligent systems of overtaking and traffic streams modeling. In *Proceedings of winter scientific session "MEPHI-2003"*, Moscow, Russian Federation

S. Butenkov, V. Krivsha, and D. Butenkov. 2005. The measures for multilevel information granulated problems formalization in the systems computing with words. In *Proceedings of International conference "Intellectual Information Analysis – 2005"*, Kiev, Ukraine

S. Butenkov, V. Krivsha, D. Butenkov, and A. Kurbesov. 2004. The adaptation of fuzzy relations for case-based reasoning in medical diagnostics and problem solving. In *Proceedings of 9th International Conference AIS-2004*, Moscow, Russian Federation

G. Chung. 2004. Developing a Flexible Spoken Dialog System Using Simulation. In *Proceedings of ACL-04*, Barcelona, Spain

M. Eskenazi, A. Black, A. Raux, and B. Langner. 2008. Let's Go Lab: a platform for evaluation of spoken dialog systems with real world users. In *Proceedings of InterSpeech'2008*.

S. Jung, C. Lee, K. Kim, M. Jeong, and G. Lee. 2009. Data-driven user simulation for automated evaluation of spoken dialog systems. In *Computer Speech & Language* 23.

S. Möller, R. Englert, K.-P. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, and N. Reithinger. 2006. MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations. In *Proceedings of InterSpeech'2006*, Pittsburgh, USA

A. Pfalzgraf, N. Pfleger, J. Schehl, and J. Steigner. 2008. ODP: ontology-based dialogue platform. Technical report, SemVox GmbH

O. Pietquin. 2008. User Simulation/User Modeling: State of the Art and Open Questions. CLASSiC Project Consortium Meeting, Issy Les Moulineaux.

J. Schatzmann, K. Georgila, and S. Young. 2005. Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. In *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal

B. Steimel, A. Jameson, O. Jacobs, and S. Pulke. 2007. Voice Awards 2007: Die Besten deutschsprachigen Sprachapplikationen. Project Deliverables

Biographical Sketch

Dmitry Butenkov currently works toward a PhD at Berlin Institute of Technology under supervision of Sebastian Möller. He joined T-Labs in October 2008 holding BSc in Computer Engineering and a Diploma in Computer Science from Taganrog State University of Radioengineering (now Taganrog Institute of Technology, Southern Federal University), Russian Federation. He was born in 1984 in Taganrog, USSR.

Caroline Clemens

Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
Germany

caroline.clemens@telekom.de
www.qu.tu-berlin.de/menue/team/forscher/caroline_clemens/
www.laboratories.telekom.com

1 Research Interests

My focus of work is the **user interface design** of spoken dialogue systems. One of my main research fields in the last years has been **automatic user classification**. I develop voice applications that support the process of **call handling in call centers**. As a **speaker** and **voice coach** I am working on the production of speech recordings.

1.1 User Classification

The aim of my dissertation is to show, how information about the users and their behavior can be gathered automatically. The results can be used for automatic user classification. As fields of benefit, this enables adaptivity of dialogue systems, gives clues for the designers, and helps marketing to know the users. The results support a prospective and therefore more efficient dialogue development process in early phases of dialogue development.

I focused on log files as a source of information. Log files are created during running dialogues and contain detailed entries about events of the dialogue with precise timestamps. I collected log files of test person calls together with data from questionnaires and interviews. The combined analysis including demographic data, personality profiles, and features of interaction behavior offer a deeper understanding and interpretation of the resulted correlations.

I worked on my PhD thesis "User classification for automatic speech dialogue systems" (see Clemens and Hempel, 2008) in the *DFG* (*Deutsche Forschungsgemeinschaft*) research training group *Prometei*[1] (Prospektive Gestaltung von Mensch-Technik-Interaktion) at the *Center of Human-Machine-Systems*[2] at the Technische Universität Berlin.

1.2 Call Center Supporting Applications

In several projects at *T-Labs*[3] (*Deutsche Telekom Laboratories*) we develop applications that support the process of call handling in call centers (for example Clemens et al., 2009). Speech based classification is already used for routing purposes in automatic voice portals. We develop concepts for the optimisation of costumer concern handling. I am working on design guidelines for a consistent voice user interface design of such applications.

1.3 Questionnaire for Technical Affinity

With colleagues of the *Center of Human-Machine-Systems* we set up a workgroup to develop a questionnaire to obtain the user's attitude towards and contact with menu based electronical devices in every day live. The questionnaire is ready and will be published soon (Karrer et al., 2009). It can be used for free for research and product development and evaluation.

2 Future of Spoken Dialogue Research

I think that the design of spoken dialogue systems has to be more flexible and adaptive in the future due to the diversity of the users and system contexts. There will be:

- heterogeneous user groups,
- less stand-alone applications and more integrations in complex systems and processes,
- combinations with others modalities and cross-device interaction.

3 Suggestions for Discussion

- Adaptivity of spoken dialogue systems: To which kind of information can a system adapt? Which kind of information retrieval will be used in the future? How can the system adapt in style, speed, content etc.?
- Combination of spoken dialogue systems with other modalities like audiovisual interaction: Which chances and which limitations do we see?
- Intuitivity of spoken dialogue systems: Is talking to a machine intuitive in general? Which aspects are important to increase intuitive use?

[1] www.zmms.tu-berlin.de/prometei
[2] www.zmms.tu-berlin.de
[3] www.laboratories.telekom.com

References

C. Clemens and T. Hempel. 2008. Automatic User Classification for Speech Dialog Systems. In *Usability of Speech Dialog Systems – Listening to the target audience*. Berlin, Springer.

C. Clemens, S. Feldes, K. Schuhmacher and J. Stegmann 2009. Automatic Topic Detection of Recorded Voice Messages. In *Proc. Interspeech 2009*.

C. Hipp et al. 2008. *Qualitätsleitfaden. Kochbuch für gute Sprachapplikationen*. Initiative VOICE BUSINESS, Fraunhofer-institut für Arbeitswirt-schaft und Organisation (IAO), Stuttgart. Qualitäts-leitfaden für Sprachanwendungen.

K. Karrer, C. Glaser, C. Clemens and C. Bruder. 2009. *Technikaffinität erfassen – der Fragebogen TA-EG*.

Biographical Sketch

Caroline Clemens studied Science of Communication and Linguistics focussing on phonetics and speech. As a spoken dialogue designer she gathered experience in dialogue design and dialogue production at *Mundwerk AG*[4] where she worked on the whole process of system development from requirement analysis, architecture design, dialogue flow, storyboard, grammar writing, to testing und evaluation. She worked at *Siemens AG, User Interface Design*[5], where she also was a PhD student with a doctoral thesis about spoken dialogue systems that was supported by the research training group *Prometei* at Technische Universität Berlin. For several years she was member of the scientific board of the *Centre of Human-Machine Systems*, State Berlin. She participated in the development of a handbook for speech applications (Hipp et al. 2008) Currently she works as a research scientist at *Deutsche Telekom Laboratories*, *Quality and Usability Lab*[6] at Technische Universität Berlin (*Institut für Softwaretechnik und Informatik*). For the second time she is in the program committee of the *Berliner Werkstatt Mensch-Maschine-Systeme*[7] and also organising a workshop[8] on *Exploring Design Criteria for Intuitive Use* at the *M&C*[9] conference.

[4] The Mundwerk AG developed and ran speech recognition based telephone applications.

[5] Siemens AG, Corporate Technology, User Interface Design.

[6] `www.qu.tu-berlin.de/menue/home/parameter/en/`

[7] `www.tu-berlin.de/zentrum\_mensch-maschine-systeme/menue/veranstaltungen/berliner\_werkstaetten\_mms/8\_berliner\_werkstatt\_mms`

[8] `www.iuui.de/workshops/mc2009/Home.html`

[9] `www2.hu-berlin.de/mc2009/`

Klaus-Peter Engelbrecht

Quality and Usability Lab
Deutsche Telekom Laboratories
Berlin University of Technology
Ernst-Reuter-Platz 7
Berlin, Germany

klaus-peter.engelbrecht@telekom.de

1 Research Interests

My research interests lie in **measurement of the usability and quality** of spoken dialogue systems (SDSs). In this, I mainly focus on the **prediction of subjective judgments**.

1.1 Context of my Work

I am working at Deutsche Telekom Laboratories in the field of automated usability evaluation methods. We aim at modelling human machine interactions with the help of user simulation, focusing on spoken dialogue systems and multimodal systems. I participated in the MeMo project, in which a workbench for early usability testing was developed. In order to simulate the interaction, the device under test is modeled as a state chart, representing the system behavior, and annotations of features of the dialogue elements, e.g. prompts or widgets. The user model searches a path through the state-chart, based on an agenda, and rules reflecting usability guidelines may cause deviations from that path. Interactions are on the concept level, and errors can be generated by randomly deleting, substituting or inserting concepts with a defined likelihood.

In another project currently running at T-labs, a user simulation interacting with real SDSs over the telephone is developed. Accordingly, the interaction is fully verbally, i.e. spoken prompts and utterances.

Apart from quantitative and qualitative parameters such as duration of the interaction, number of turns to accomplish the goal or task success, we would like to obtain an estimation of subjective measures like user satisfaction or cognitive demand during the interaction.

1.2 Previous work

I started my work on the research area with my Magister thesis, in which I developed a classification scheme for user behavior in interactions with spoken dialogue systems (Engelbrecht, 2006). This served as a basis or the behavior that should be modeled in the MeMo workbench (or generally in a user simulator for SDS evaluation).

My work in the MeMo project laid mainly in the evaluation of the simulation approach. I modeled an existing system with the workbench, and compared the simulations to real user data gathered with this system. In addition, I compared two real corpora according to the same criteria. Criteria were interaction parameters, user ratings (predicted with PARADISE (Walker et al., 1997) in case of the simulation), and coverage of utterances in the corpora. In addition, I analysed in depth which of the behavior classified in my Magister thesis needs to be added to the simulation.

1.3 Prediction of user judgments

User judgments provide a valid and easy-to-compare means of determining the overall quality of a system. Therefore, as our simulations focus on the evaluation of systems, a prediction model for user judgments can be very helpful.

There is one previous approach to such predictions, which is the PARADISE framework by Walker et al. (1997). The basic assumptions are that judgments depend on task success and dialogue costs, and that task success and dialogue costs can be measured instrumentally in the form of interaction parameters. Linear Regression is used to derive the prediction function. The accuracy of the predictions, however, is relatively low. Models can explain around 50% of the variance in the judgments. An extensive evaluation of such models showed that prediction results can hardly be improved, even if more interaction parameters are considered (Möller et al., 2008) or the classifier is changed.

I have analysed the value of PARADISE-style models for the task desired in our user simulations, i.e. providing an overall measure of the system's quality from the perspective of the user. I found that the predictions result in relatively accurate mean values for system configurations, despite the low accuracy in predicting the ratings for individual dialogues. My assumption is that users differ in their judgment behavior, depending on their current mood, attitude towards SDSs and many other factors which are hard to measure or even to identify. As these differences are random, they are equalised when average judgments are considered (Engelbrecht and Möller, 2007).

In a follow-up study, I analysed an experiment where 14 tasks were judged by each user (Engelbrecht et al., 2008). Correlations between interaction parameters and judgments could be calculated. The finding was that the parameters correlated with the judgment were similar for all users. However, the strength of the correlations differs considerably among the users, namely depending on the users' age, affinity towards technology, and their cognitive abilities. However, even for the groups which could be predicted best, the accuracy of predictions was relatively low.

Therefore, a new approach was created to more accurately model the interrelations between events in the dialogue, and to take into account the inter-individual differences in judging things. In this approach, a HMM is used to model the evolution of the judgment over time. In the model, the judgment depends on the current combination of dialogue events, as well as the judgment of the dialogue so far (i.e. the previous judgment). To attain data for the model, I conducted a WOZ experiment in which I asked the users for a quality rating after each dialogue turn. The model predicts the probability for each judgment ("bad".."excellent") at each turn, reflecting the differences between the users.

In the future, I plan to test the model on data from other databases and systems. Also, I will investigate the possibilities of training the HMM on dialogues for which only the final judgment is available. Finally, I plan to integrate the model into the user simulator and compare estimations of quality from the simulation to measurements in an experiment with real users.

2 Future of Spoken Dialogue Research

From my perspective, commercially applied SDSs have made major progress in the past two years. With a well designed system, speech recognition errors are seldom, and efficient and reliable recovery/prevention strategies are at hand. However, for many tasks the systems are applied for, the internet is far more efficient.

I believe there are two possible lines research should follow. Firstly, improved user simulation might enable extensive testing, and thus allow to create systems with reliable, but more complex (and efficient) dialogue strategies. These might include adoption to the users, e.g. in terms of default tunings or expectation of typical behavior patterns or tasks.

On the other hand, mobile internet, especially on small devices, might be more attractive in many cases. However, Speech technology might add comfort to such interfaces and services.

3 Suggestions for Discussion

- Perspectives for SDSs. Which applications will overcome mobile internet?
- System performance, user and expert judgments. Which of them determines system quality?
- Modelling of SDS users. How accurate models are needed?

References

K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabdar and S. Möller. Modelling User Satisfaction with Hidden Markov Models. *Submitted to SIGDial 2009.*

K.-P. Engelbrecht, S. Möller, R. Schleicher and I. Wechsung. 2008. Analysis of PARADISE Models for Individual Users of a Spoken Dialog System. In *Proc. of ESSV'2008.*

K.-P. Engelbrecht and S. Möller. 2007. Pragmatic Usage of Linear Regression Models for the Prediction of User Judgments. In *Proc. of SIGDial'07.*

K.-P Engelbrecht. 2006. Fehlerklassifikation und Benutzbarkeits-Vorhersage für Sprachdialogsysteme auf der Basis von mentalen Modellen. *Magister thesis.* Deutsche Telekom Labs, TU Berlin.

S. Möller, K.-P. Engelbrecht and R. Schleicher. 2008. Predicting the Quality and Usability of Spoken Dialogue Services. In *Speech Communication 50*, pp. 730-744.

M. Walker, D. Litman, C. Kamm and A. Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proc. of ACL/EACL*, Madrid, pp. 271–280.

Biographical Sketch

Klaus-Peter Engelbrecht is working as Wissenschaftlicher Mitarbeiter at the Quality and Usability Lab at Deutsche Telekom Laboratories, TU-Berlin. He studied Com-munication Research and Musicology and received his Magister degree in 2006 from the TU-Berlin. At T-Labs, he is working towards his PhD thesis in the domain of automated usability evaluation for spoken dialogue systems. His extracurricular interests include playing in rock bands and sports.

Arash Eshghi

Queen Mary University of London
Interaction, Media and Communication Group
Mile End Road
London, E1 4NS

arash@dcs.qmul.ac.uk

1 Research Interests

My main research interests are in the area of **human-human interaction**. I am interested in how the conversational **context** evolves and **coordination of agents** is achieved in **multi-party dialogues** (henceforth multilogue). My research has so far been experimental, but I'm also interested in how the phenomena observed via experimentation may be captured within a formal contextual framework.

2 Past, Current and Future Work

Models of dialogue, both psychological (e.g. (Clark and Schaefer, 1989)) and computational (e.g. Ginzburg's KoS (Ginzburg 2009)), have been developed primarily to account for how context is built up through direct interaction between *pairs* of participants. Multilogue, however, is conditioned by an overall lack balance in the participants' levels of direct interaction or involvement with one another. It is not only possible but likely that a participant will be left out temporarily or during a whole conversation/topic.

This immediately raises the question of what, if anything, is the difference between those participants who are in direct interaction for a stretch of talk and those who provide little or no feedback during it. Are the latter at a disadvantage in terms of the levels of grounding and semantic coordination that they reach as regards the material they provided little or no feed back on? Studies have shown that they could be (e.g. (Schober and Clark, 1989; Healey and Mills, 2006)). But would these results persist even if this lack of involvement was only 'temporary'? If not, then what are the differences between a participant who is only momentarily inactive in order to avoid interruptions or overlapping talk, with one who plays a completely passive role? More broadly, how does shared context evolve and when and for whom is it shared in multi-party dialogue? These are questions which my research has attempted to address.

2.1 Collective states of understanding

A corpus analysis of the different surface forms of context-dependency (elliptical expressions) shows that there are dialogues in which a third silent participant, despite the lack of overt feedback, reaches the same level of grounding or mutual understanding as the dyad actively engaged in talk. These dialogues lead to collective contexts that are not reducible to the component dyadic interactions that gave rise to them (Eshghi and Healey, 2007).

2.2 Grounding by proxy

It is then proposed that such collective contexts are the effect, at least, of those dialogues in which subgroups of the participants are organised into *parties* (Schegloff, 1995), such that there are fewer parties than there are individuals. The claim is developed that grounding operates in such dialogues, *in the first instance*, between parties rather than individuals, such that the party acts as a unified aggregate to carry the conversation forward. That is, the party is responsible as a whole, for grounding and thus satisfying the constraints imposed by the current context of the conversation. In this manner, one member can *stand proxy* for others in doing this, such that it doesn't matter which member is doing the talking as the rest effect the same contextual increments as that individual. This provides a pragmatic parallel to Schegloff's (1995) proposal that turn-taking operates in the first instance between parties. An experimental test of this idea is carried out which provides causal evidence for the practical reality of parties, in the form of 'extra' *intra*-party interactions, which, all things being equal, are caused by the failure of one of the party members to ground.

2.3 Emergence of unevenly distributed shared contexts

The claim is then developed and tested experimentally that in contrast to the dialogues leading to the evenly shared irreducible collective contexts just mentioned, multi-party dialogues can also under some circumstances, give rise to multiple shared contexts that are *unevenly* distributed across the (ratified) participants (Eshghi and Healey, 2009). This is manifested in systematically divergent interpretations of one and the same utterance by different participants owing to the distinct contexts against which the utterance is assessed, a phenomenon we dub 'Pragmatic Pluralism' in dialogue. This occurs we argue, when a participant falls outside the interacting parties for a stretch of talk, and as a result is more weakly attached to the local context than the members of the interacting parties.

Arguably, the results summarised above present significant challenges for how formal models of dialogue context - underpinning dialogue systems - characterise the contextual increments arising from each individual contribution. Within Ginzburg's (1996, 2009) question based approach to modelling contextual change, these results indicate the possibility of a systematic divergence between the participants' Dialogue Game Boards (DGB). More specifically, they point to the need for the utterance by utterance DGB Update Rules (specifically FACTS-incrementation and QUD-downdate rules) to be relativised to *participant role*, understood here, NOT in terms of Goffman's participation framework, but in terms of party membership.

3 Future of Spoken Dialogue Research

Spoken Dialogue Systems are becoming increasingly commonplace, and yet, they seem to be hampered greatly by the fact that the semantic ontologies underpinning them are often, if not always, ad-hoc (domain-specific) and static. That is, the theoretical models that they're based on, fail to account for the ways in which shared semantic ontologies or conventions emerge from the interaction itself (the view of language as an adaptive system) via local coordination mechanisms such as repair, clarification and feedback. This requires a formal concept of context that is sufficiently rich to encode these essentially emergent and dynamic conventions and their transitions at a local level. A more generic dialogue system would certainly need to address these problems. In the case of multi-party dialogues, the situation is even more complex, as different sub-groups of the participants may reach divergent levels of semantic coordination (Healey and Mills, 2006) contingent upon their levels of direct interaction. Moreover, sub-groups may effect divergent commitments (contextual updates) as a result of one and the same utterance (Eshghi and Healey, 2009) depending on their prior knowledge of and publicly presumed stance towards the topic at hand. This latter problem is particularly significant in the development and implementation of automatic meeting assistants or summarisers.

4 Suggestions for Discussion

- How adaptive should a dialogue system be? How can we make dialogue systems more generic? How can the semantic coordination of agents be formalised?
- Apart from the current difficulties in speech recognition, what are the most pressing problems in the development of automatic meeting summarisers/assistants? How can they be addressed given the extreme complexity and contextual variability inherent in multi-party dialogues?
- How can different forms of non-verbal behaviour be recognised, and in turn how do they help the system in choosing the right course of action, which otherwise could involve needless clarification sub-dialogues?

Références

H. H. Clark, and E. A. Schaefer. 1989. Contributing to discourse. In *Cognitive Science* 13, pp. 259-294.

A. Eshghi, and P. G. T. Healey. 2007. Collective states of understanding. In B. H. Keizer S. and T. Paek (Eds.), In *Procs. of the 8th SIGDial workshop on discourse and dialogue*, pp. 2–9. Association for Computational Linguistics.

A. Eshghi, and P. G. T. Healey. 2009. What is conversation? distinguishing dialogue contexts. In *Proceedings of the annual cognitive science conference*.

J. Ginzburg. 1996. Dynamics and the semantics of dialogue. In J. Seligman (Ed.), *Language, Logic and Computation*.

J. Ginzburg. 2009. The interactive stance. *CSLI: Center For The Study of Language and Information*.

P. G. T. Healey, and G. J. Mills. 2006. Participation, precedence and co-ordination in dialogue. In *Proceedings of the annual cognitive science conference*.

E. A. Schegloff. 1995. Parties talking together: two ways in which numbers are significant for talk-in-interaction. In P. Ten Have and G. Psathas (Eds.), *Situated order: Studies in social organization and embodied activities*, pp.31–42. University Press of America.

M. F. Schober, and H. H. Clark. 1989. Understanding by addressees and overhearers. In *Cognitive Psychology* 21, pp 211–232.

Biographical Sketch

Arash Eshghi has just defended his PhD thesis in the Interaction, Media and Communication Group at Queen Mary University of London. His first degree was in Computer Science, and he has a Masters in Artificial Intelligence. Given the highly experimental nature of his thesis, his interests are, at least in principle, cross-disciplinary. He greatly enjoys philosophical debates and loves skiing and camping.

Milica Gašić

Department of Engineering
University of Cambridge
Trumpington street
Cambridge

mg436@cam.ac.uk
mi.eng.cam.ac.uk/~mg436/

1 Research Interests

My research interest lies in the area of **statistical** approaches to **Dialogue Management**. I am primarily focused on **Partially Observable Markov Decision Process framework to Spoken Dialogue Systems**. This framework has the potential to enable learning from data, learning from interaction with both simulated and real users and, also, to be robust to recognition errors.

1.1 Modelling Dialogue as a Partially Observable Markov Decision Process

One of the main problems that real world applications of Spoken Dialogue Systems are encountering is the difficulty to deal with the errors from the speech recogniser in a noisy environment. As almost any real application is not noise free, this is a large obstacle preventing dialogue systems from wide use.

It has been suggested that the Partially Observable Markov Decision Process (POMDP) can provide a principled mathematical framework to deal with the uncertainty that originates from errors in the speech recognition by retaining and updating the distribution over all states in each turn (Young, 2002; Williams and Young, 2007). The idea is that the dialogue state is hidden and therefore the policy should not be based on a single state but on the distribution over all possible states.

The main weakness of POMDP approach is its computational intractability for even very simple domains. However, it has been suggested that factoring the states into summary states can enable using POMDPs for policy optimisation (Williams and Young, 2007). It has also been shown that the policy learning can be performed online in interaction with a simulated user.

1.2 Current Work

My work is focused on the Hidden Information State system (HIS) (Young et al., 2009). What makes HIS different to other POMDP-based dialogue managers is its representation of user goal in the dialogue state. It retains the full representation of user goal though partitions of the user goal space. The partitions are created based on the information that the user has provided. After each dialogue turn the partitions and their probabilities are updated. The partitions have the ability to represent any user goal based on the domain ontology. As the noise in the input increases the number of partitions increases exponentially. Therefore, a pruning technique has been devised that can deal effectively remove low probability partitions and still retain the ones that are represent plausible user goal.

In order to reduce the dimensionality, the state space (the master space) is mapped into a much smaller summary space. Reinforcement learning is performed on the probability distribution over the summary space – the belief space. Since the belief space is continuous, it is discretised into grid points. Then, Monte Carlo Control algorithm (Sutton and Barto, 1998) is performed on these points. Finally, the outcome of the policy is mapped back into the master space.

The results of my current work suggest that the training in noise leads to better performance at higher semantic error rates then in training in noise free environment. Also, the ability of a dialogue manager to make use of N-best inputs from a recogniser improves in the performance at higher error rates (Gašić et al., 2008).

In addition to that, the performance can be improved by ordering the actions in an N-best list using the Q-values from the Monte Carlo Control algorithm. The idea is if the top action is not mappable to master space, to use the next-best one rather than the default action.

1.3 Future Work

What I encountered during the training of the dialogue manager with a simulated user is training a very small number of grid points can reach the same performance as using a larger number of grid points for learning. This does not only suggest that there the current choice of summary space can be enriched in order to be more informative, but that the Q-value function can be more effectively estimated in various parts of the summary space. Therefore, rather then computing the exact Q-value on grid point and generalising over a certain part of space as is done in grid based Monte Carlo Control algorithm, it would be better to have a functional estimate of Q-value function over summary space with a measure of certainty.

This can then be used for the exploration/exploitation trade-off during learning and also as a basis for policy adaptation with real users.

2 Future of Spoken Dialogue Research

In the next decade, I expect noise robustness issues that prevent dialogue systems from a wider application to be resolved. Also, it is very important to define the framework which is domain independent and easily transferable to other domains. POMDPs seem to have the potential for this, but additional research is needed in order to further investigate the possibilities to obtain desirable results with a computationally tractable approximations. It would be desirable that the process of building a Statistical Spoken Dialogue System in the next decade becomes fully automatic. This would mean that the simulated user can be trainable from real data, that the dialogue manager can learn from the interaction with the simulated user and finally that is also able to learn with real users. Therefore the future of the Dialogue Manager is highly dependent on the simulated user.

3 Suggestions for Discussion

- Machine Learning Techniques in Dialogue Modelling
- Data acquisition and reliability
- Generalisation across different domains

References

R. S. Sutton and A. G. Barto 1998. Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA.

S. J. Young. 2002. Talking to machines (statistically speaking). In *Proc. of Int. Conf. Spoken Language Processing*, Denver, Colorado.

J. D. Williams and S. J. Young. 2007. Partially observable Markov decision processes for spoken dialog systems. In *Computer Speech and Language* 21 (2), pp. 393–422.

M. Gašić, S. Keizer, B. Thomson, F. Mairesse, J. Schatzmann, K. Yu, and S. Young. 2008. Training and evaluation of the HIS-POMDP dialogue system in noise. In *Proc. 9th SIGDial*, Columbus, OH.

S. Young, M. Gašić, S. Keizer, B. Thomson, F. Mairesse, J. Schatzmann, and K. Yu. 2009. The Hidden Information Statemodel: A practical framework for POMDP-based spoken dialogue management. In *Computer Speech and Language*. doi:10.1016/j.csl.2009.04.001.

Biographical Sketch

After graduating in Computer Science and Mathematics from the University of Belgrade in 2006, Milica Gašić did MPhil in Computer Speech, Text and Internet Technology at the University of Cambridge. She completed the course in 2007 with the thesis Limited Domain Synthesis of Expressive Speech. In October 2007 she became a PhD candidate in Dialogue Modelling under the supervision of Prof Steve Young. She is working in the Dialogue Systems Group: http://mi.eng.cam.ac.uk/research/dialogue/.

Florian Gödde

Berlin Institute of Technology
Quality & Usability Lab
Ernst-Reuter-Platz 7
10587 Berlin - Germany

florian.goedde@tu-berlin.de

1 Research Interests

Building user-friendly spoken dialogue systems (SDS) is not trivial, as one can see that many deployed systems miss a clear, understandable dialogue structure, informative help prompts that focus on the necessary information and individual error recovery strategies for different errors. Since user tests within the development cycle of SDSs are very expensive and time consuming, it is of interest to do usability testing in the early stages of development in an automated way, to provide guidance for developers throughout the whole process and not only at a very late stage where some flaws may not be fixed that easily. I am working on semi-automatic evaluation methods for spoken dialogue systems that might provide a fast and cheap way to test early prototypes of SDSs against design and usability flaws. Furthermore, I am interested in SDSs for different user groups. One of the most promising applications for SDSs is the home-care sector, where easy and simple access to information technology is mandatory to help older users to interact with computer systems. Since speech is something that most of the potential users do not have to learn (in contrast to keyboard and mouse usage), SDS are a promising alternative. But frustration with systems that do not work as expected by the users might be very frustrating, so a clear understanding of the needs and abilities of older users is mandatory.

1.1 Past Research

To predict user satisfaction with SDSs it is necessary to understand how users perceive these systems. Different users have different needs regarding their interaction with computers: For instance, novice users should be provided with extensive help during a dialogue, whereas expert users should be able to perform their task as fast as possible. In the frame of my diploma thesis I analysed the interactions of two user groups using the smart-home system INSPIRE, namely younger and older users (Gödde et al., 2008). Older users are of particular interest for the industry, since SDSs might help older users to successfully integrate modern technology into their daily life, which can help them to stay independent as long as possible in their own homes. As an outcome of this work I showed, that the behavior and acceptance of SDSs does not mainly depend on the age of the users, but on a more general affinity towards technology, and that older users tend to speak in a more natural way to this system, comparable to human-human communication. Furthermore, older users benefit more from extensive help prompts, since it supports fast alignment to the system vocabulary (Wolters et al., 2009).

1.2 Current and Future Research

Currently I am working towards my PhD in a project named “SpeechEval”. We are building a framework for semi-automatic quality testing of spoken dialogue systems, which uses user simulation and usability prediction based on data-driven machine learning algorithms. The idea is to build a realistic user simulation that learns interaction behavior out of a dialogue corpus containing a variety of dialogues with commercial systems, and to generalise from that corpus to be able to interact with new systems. I am developing a method that uses dialogue corpora together with user ratings to automatically assign real user ratings to the simulated dialogues. Furthermore, I will try to measure the quality of the simulated dialogues independently of user characteristics. That means, the quality rating of the system is valid for all simulated user groups. As an outcome of the project, we build a workbench that utilises user simulation to interact with the SDS under test, by learning the dialogue structure in the first dialogues through trial and error. The user simulation recognises keywords in help prompts and uses ontologies to build utterances that the system should understand. After some dialogues, the user simulation gets more confident in the interaction, and the generated dialogue logs can then be used to measure the quality, usability and user satisfaction of the system. The developer of the SDS can then use this report to improve his/her system, and test it again, and so on. A user test with real users might then be only necessary at the end of the development cycle to get real ratings from the target group.

2 Future of Spoken Dialogue Research

For the next years, not only improving techniques for all parts of a SDS will be of interest, but also research on overall acceptance of these systems. Since speech will be more common, for instance as input modality in smart homes, it is necessary to understand how users perceive these systems, how users want to interact and what makes a dialogue system acceptable and comfortable to use.

We have to find out

- what is important for the users, “natural” –human-human like- dialogues or fast and error-free communication
- which factors are really influencing the satisfaction with SDSs. This question is not yet answered, since in every model of user satisfaction prediction there is a large unexplained variance
- how can SDSs be seamlessly integrated into modern technology, for instance in the home care sector

3 Suggestions for Discussion

- What do people really want and what not? How to increase the acceptance of SDSs by looking more closely on the users needs.
- How can real user satisfaction be measured? Do field tests provide more realistic judgments then laboratory experiments?
- Human-Human communication is very flexible, dialogues concerning the same topic will never be actually identical, in contrast to human-machine dialogues. Is it feasible for spoken dialogue research to work on dialogue systems with flexible prompts to bring this variety into HMI?

References

F. Gödde, S. Möller, K.-P. Engelbrecht, C. Kühnel, R. Schleicher, A. Naumann, and M. Wolters. 2008. Study of a Speech-based Smart Home System with Older Users. In *Proc. Int. Workshop on Intelligent User Interfaces for Ambient Assisted Living* (IUI4AAL 2008), Canary Islands, Jan. 2008, Frauenhofer IRB Verlag, Stuttgart, pp. 17–22.

M. Wolters, K.-P. Engelbrecht, F. Gödde, S. Möller, A. Naumann, and R. Schleicher. 2009. Making it Easier for Older People to Talk to Smart Homes: The Effect of Early Help Prompts. Accepted for *Universal Access in the Information Society*.

Biographical Sketch

Florian Gödde graduated in Computer Science in 2007 with his diploma thesis “Evaluation of a Smart-Home System for Elderly Users”. After graduation, he worked at the Carmeq GmbH on spoken dialogue systems for Volkswagen. Since September 2008 he is working towards his PhD at the Berlin Institute of Technology on usability of spoken dialogue systems under the supervision of Prof. Sebastian Möller.

Christine Howes

Queen Mary University of London
Interaction, Media and Communication Group
Mile End Road
London, E1 4NS

chrizba@dcs.qmul.ac.uk
www.dcs.qmul.ac.uk/~chrizba

1 Research Interests

My main research interests are in the area of **human-human interaction**. Specifically, I am concerned with how **coordination** of agents is achieved via the building up of semantic knowledge and/or situation models using strictly **incremental** mechanisms of syntax (using the formalism of **Dynamic Syntax**). I am currently investigating these questions via the study of **split utterances**.

1.1 Background and related work

The pioneering work of Clark (1996) initiated a broadly Gricean program for dialogue modelling, in which coordination in dialogue is said to be achieved by establishing recognition of speaker-intentions relative to what each person takes to be their mutually held beliefs (*common ground*). However, computational models in this vein have largely been developed without explicit high-order meta-representations of others' beliefs or intentions and it is arguable that the Gricean assumptions underpinning communication should be re-considered. In Kempson et al. (2009), the groundwork is laid for an interactive model of communication using Dynamic Syntax (DS: Cann et al., 2005), which examines its application to the tightly interactive dialogue phenomena that arise in cases of utterances split between speakers. In this model, each interlocutor interprets the signals they receive and plans the signals they send, egocentrically, without explicit representation of the other party's beliefs or intentions. Nevertheless, the effect of coordinated communication is achieved by relying on ongoing feedback and the goal directed action-based architecture of the grammar. The claim is that communication involves taking risks: in all cases where an interlocutor's system fails to fully determine choices to be made (either in parsing or production), the eventual choice may happen to be right, and might or might not get acknowledgement; it may be wrong and (possibly) get corrected; or, in recognition of the nondeterminism, they may set out a sub-routine of clarification, before proceeding.

1.2 Split Utterances

Split utterances (*SUs*) – single utterances split between two or more dialogue turns/speakers – have been claimed to occur regularly in dialogue. They are of interest to dialogue theorists as a clear sign of how turns cohere at all levels – syntactic, semantic and pragmatic. They also indicate the radical context-dependency of conversational contributions. Turns can be highly elliptical and nevertheless not disrupt the flow of the dialogue. SUs are the most dramatic illustration of this: contributions spread across turns/speakers rely crucially on the dynamics of the unfolding context, linguistic and extra-linguistic, in order to guarantee successful processing and production.

Utterances that are split across speakers also present a canonical example of participant coordination in dialogue. The ability of one participant to continue another interlocutor's utterance coherently, both at the syntactic and the semantic level, suggests that both speaker and hearer are highly coordinated in terms of processing and production. The initial speaker must be able to switch to the role of hearer, processing and integrating the continuation of their utterance, whereas the initial hearer must be closely monitoring the grammar and content of what they are being offered so that they can take over and continue in a way that respects the constraints set up by the first part of the utterance.

Currently, I am investigating the impact that SUs have on ongoing dialogues using the DiET chat tool (Healey et al., 20003). The experiment reported in Howes et al. (2009) tests the effects of artificially introduced SUs on groups of people engaged in a task-oriented dialogue. Results show reliable effects on response time and on the number of edits in formulating subsequent turns. In particular, if the second part of an utterance is 'misattributed' people take longer to respond, and responses to utterances that appear to be split across speakers involve fewer deletes. This provides evidence that: a) speaker switches affect processing where they interfere with expectations about who will speak next and b) the pragmatic effect of a split is to suggest to other participants the formation of a coalition or 'party' (Schegloff, 1995; Lerner, 1991).

Complimentarily, a preliminary corpus study on a subsection of the British National Corpus (Purver et al. 2009) indicates that splits can occur anywhere in a string. This is consistent with models that advocate highly coordinated resources between interlocutors and, moreover, the need for highly incremental means of processing (Purver et al., 2006; Skantze and Schlangen, 2009).

From a computational modelling point of view, the results of the corpus study indicate that continuations usually start in an incomplete way, which means that a dialogue system has a chance of spotting them from surface characteristics of the input. However, this is hampered by the fact that the split can occur within any type of syntactic constituent, hence no reliable grammatical features can be employed securely. On the other hand, 8% of all utterances do not end in a complete way, and only 63% of these ever get continued. Contrarily, the vast majority of continuations continue an already complete antecedent and long distances between antecedent and continuation are possible. In this respect, locating the antecedent is not a straightforward task for automated systems, especially again as this can be any type of constituent, and many SUs are split over more than two contributions.

2 Future of Spoken Dialogue Research

I see Spoken Dialogue Systems proliferating in the future, and feel that if they are to be accepted by the wider community then we need to get a handle on how human-human communication progresses so effortlessly, and which of these techniques can be effectively utilised in dialogue systems. This should include (but not be restricted to) a more detailed understanding of context (and how dialogue contexts can vary), feedback (in the form of backchannels etc - are there specific points when these are somehow ‘more’ appropriate), prosody, and the need for parsing and processing (and possibly the grammar itself) to be incremental and predictive. A good way to get a handle on some of these questions is through the development of spoken dialogue systems in severely restricted domains.

3 Suggestions for Discussion

- How general (and/or human-like) should a dialogue system be.

- How can we model a system that is flexible enough to process all the fragmentary, multiply ambiguous data that humans produce on a regular basis, and, if necessary, seek clarification?

- What feedback should a system give the user, and when is it appropriate to do so? And can this be determined by e.g. syntactic constraints, or intonational contours, or (potentially partial) semantic interpretations?

References

R. Cann, R. Kempson, and L. Marten. 2005. The Dynamics of Language. Elsevier, Oxford.

A. Gargett, E. Gregoromichelaki, C. Howes, and Y. Sato. 2008. Dialogue-grammar correspondence in dynamic syntax. In *Proceedings of the 12th SEMDIAL (LONDIAL).*

E. Gregoromichelaki, Y. Sato, R. Kempson, A. Gargett, and C. Howes. 2009. Dialogue modelling and the remit of core grammar. In *Proceedings of IWCS.*

P. Healey, M. Purver, J. King, J. Ginzburg, and G. Mills. 2003. Experimenting with clarication in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society.*

C. Howes, P. G. Healey, and G. Mills. 2009. A: An experimental investigation into ... B: ... split utterances. In *Proceedings of SIGDial’2009.*

R. Kempson, E. Gregoromichelaki, M. Purver, G. J. Mills, A. Gargett, and C. Howes. 2009. How mechanistic can accounts of interaction be? In *Proceedings of the 13th SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue (DiaHolmia).*

G. Lerner. 1991. On the syntax of sentences-in-progress. Language in Society, pp.441-458.

M. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. In *Behavioral and Brain Sciences* 27, pp. 169-226.

M. Poesio and H. Rieser. to appear. Completions, coordination, and alignment in dialogue. Ms.

M. Purver, C. Howes, P. G. Healey, and E. Gregoromichelaki. 2009. Split utterances in dialogue: a corpus study. In *Proceedings of SIGDial’2009.*

G. Skantze and D. Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009).*

Biographical Sketch

Christine Howes is a 2nd year Ph.D student in the Interaction, Media and Communication Group at Queen Mary University of London. Her first degree was in Artificial Intelligence and Psychology, and she has a Masters in Computational Linguistics and Formal Grammar. She thus likes to think of herself as inter-disciplinary.

Outside of research, she is a keen follower of the mighty Brentford F.C., and also loves the escapism of the cinema.

Srinivasan C Janarthanam

School of Informatics
University of Edinburgh
Edinburgh EH89AB
Scotland, UK

s.janarthanam@ed.ac.uk
srinivasancj.googlepages.com

1 Research Interests

Spoken dialogue systems must ideally adapt to the users based on their domain expertise. Although such adaptation can happen at different levels, I am interested in how referring expressions for domain objects can be appropriately chosen in a user-specific manner. In a technical support dialogue task, expressions can be chosen between technical terms and descriptions based on the user's experience in the domain, i.e. descriptive expressions for beginners and technical expressions for experts. Similarly, in a city navigation task, the system could choose to use proper names for locals while foreign tourists get more descriptive expressions. Natural language generation policies to choose the right set of referring expressions to use in a dialogue can be learned or hand-coded easily for different users separately. However, such policies cannot handle a user whose expertise is unknown. Ideally, the system needs a single policy that can adapt to all different kinds of users during the course of the dialogue based on the evidence presented by the users. Reinforcement learning methods can be used to learn adaptive referring expression generation policies (Janarthanam and Lemon, 2008, 2009a). Using a user simulation that is sensitive to the choice of referring expressions that the system is using, it is possible to learn a referring expression generation policy. The policy learned will be adaptive if the simulation produces users with different levels of expertise in different cycles. For this purpose, we propose a two-tiered simulation (Janarthanam and Lemon 2009c). We use a wizard-of-Oz environment to collect dialogue corpora to model the user's behavior in the simulation. (Janarthanam and Lemon, 2009b).

In addition to the above, I am also generally interested in other areas of dialogue systems research like tutorial dialogue systems, learning dialogue management and NLG policies, incremental processing in NLU and DM, etc.

2 Future of Spoken Dialogue Research

Dialogue systems community is slowly moving from information seeking tasks, like town-info, flights, etc to troubleshooting and technical support tasks. This will open up more research questions in terms of task management moves.

In terms dialogue processing, incremental processing (Skantze and Schlangen, 2009) is the direction to take. It avoids unnecessary delays and makes the conversation more natural. However, incremental processing must be implemented at all steps like parsing, dialogue management and generation.

More features like back-channeling, turn-taking, time management strategies must be explored in dialogue management. The use of multi-dimensional dialogue acts (Bunt and Girard, 2005) must be explored to make the dialogue systems more natural.

Dialogue systems have to be adaptive to the user to make the conversation less frustrating. Adaptation can happen over different dimensions like user's expertise in the domain, their age group, etc.

3 Suggestions for Discussion

Following are the topics that I think that needs to be discussed in the workshop:

- Adaptive systems: What are the features to look for to adapt in a user?
- Multi-functional dialogue acts: What dimensions must be considered in dialogue management?
- Enriched speech recognition: Can speech recognisers tell more than just the sequence of words and the confidence scores? Can prosodic features be detected and used to make more natural dialogue management decisions?
- Corpus collection strategies: How to collect large dialogue corpora for different needs quickly and inexpensively?

References

Harry Bunt and Yann Girard. 2005. Designing an open,multidimensional dialogue act taxonomy. In Gardent, C., and Gaiffe, B. (eds). *Proc. 9th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 37–44.

Skantze, G. and Schlangen, D. 2009. Incremental dialogue processing in a micro-domain. In *Proc. EACL 2009*, Athens, Greece.

Srinivasan Janarthanam and Oliver Lemon. 2008. User Simulation for knowledge-alignment and on-line adaptation in Troubleshooting Systems. In *Proc. SEMDial 2008*, London, UK.

Srinivasan Janarthanam and Oliver Lemon. 2009a. Learning Lexical Alignment Policies for generating Referring Expressions for Spoken Dialogue Systems. In *Proc. ENLG 2009*, Athens, Greece.

Srinivasan Janarthanam and Oliver Lemon. 2009b. A Wizard-of-Oz Environment to study Referring Expression Generation in a Situated Spoken Dialogue Task. In *Proc. ENLG 2009*, Athens, Greece.

Srinivasan Janarthanam and Oliver Lemon. 2009c. Learning Adaptive Referring Expression Generation policies for Spoken Dialogue Systems. In *Proc. SEMDial 2009*, Stockholm, Sweden.

Biographical Sketch

Srinivasan C Janarthanam is a third year Ph.D student at the University of Edinburgh. He is a UKIERI (2007-10) scholar and is funded by the British Council. Previously, he worked as a research associate in Amrita University, India and as an applications developer in iNautix Technologies, India. He has an M.Sc in Intelligent Systems from the University of Sussex, UK and a B.E in Computer Science and Engineering from Bharathiyar University, India. Currently, his work relates to Natural Language Generation in Dialogue Systems. Previously, he has worked on Dialogue Management, Tamil Named Entity Recognition, English-Tamil Machine Translation, Tamil Dependency Parsing and Morphological Analysis.

Denise Cristina Kluge

Federal University of Paraná-Litoral
Applied Linguistics
Matinhos, PR, Brazil

klugedenise@hotmail.com

1 Research Interests

My research interests lies in practical uses of spoken dialogue systems, towards studies that investigate **foreign language speech production and perception**, **perceptual training**, **audiovisual speech perception** and **speech intelligibility.**

1.1 Previous work

As the interrelationship between perception and production has been discussed in the second/foreign (L2) language phonetics and phonology literature and some studies have shown that perception plays a very important role in the production of second language sounds (Flege, 1995; Kuhl and Iverson, 1995; Best and Tyler, 2007), I have investigated the relationship between perception and production of Brazilian learners of English as a foreign language.

The study aimed at investigating the perception and production of the English nasals /m/ and /n/ in syllable-final position by Brazilian learners of English (Kluge et al., 2007) as those coda nasals have different patterns of phonetic realisations across languages, whereas they are distinctively pronounced in English, in Brazilian Portuguese they are not fully realised.

It might be expected that for accurate production, the learner would need accurate perception, which was the case of the study conducted, considering the two perception tests: an identification and a discrimination test. The results indicated that there was some relationship between the identification/ discrimination of the target coda nasals and their accurate production. However, the tendency of this study was for production to be more accurate than perception.

1.2 Current and future work

Recent studies have included visual cues as a variable to investigate the perception of L2 contrasts (Hardison, 1999; Öhrström and Traunmüller, 2004; Hazan et al., 2006) and they have shown that L2 listeners seemed to benefit from an Audio/Video presentation in the identification of visually distinctive L2 contrasts (Hazan et al., 2006). Bearing this in mind, my colleagues and I are currently working with the use of visual cues in the perception of nonnative contrasts.

We have recently investigated whether Brazilian learners of English would benefit from AV presentation when they had to identify a visually distinctive L2 contrast such as the labial/alveolar nasal consonants contrast in word-final position in English (Kluge et al., 2007). Brazilian learners of English have difficulties with English nasals in syllable-final position due to phonological differences between the two languages. In this study, the English monosyllabic words with either /m/ or /n/ in word-final position were presented in three different conditions (*Audio only*, *Audio/Video* and *Video Only*). Results not only showed that the Audio/Video condition seemed to favor the accurate identification of both word-final nasal consonants when compared to the *Audio only* condition, but also showed a slight tendency for the *Audio only* condition to disfavor the accurate identification of both bilabial and alveolar nasal consonants compared to the *Audio/Video* condition.

Taking into consideration the findings that L2 speakers may benefit from audiovisual input when they perceive visually distinctive L2 contrasts and the fact that a more accurate perception may lead to a more accurate production, our current work aims at investigating whether audiovisual perceptual training may be able to alter production. In order to do so, we are investigating whether Brazilian learners of English may be able to improve their perception and production of English word-final nasal consonants /m/ and /n/ by means of audiovisual perceptual training with this visually distinctive contrast. We are also aiming at investigating whether audiovisual perceptual training would alter perception and production of Brazilian learners of English with contrast such as the voiceless TH, S and F in word-final position which are not too visually distinctive contrasts as compared to the English nasal consonants in word-final position.

As regards the application and impact of the results of this study on research on spoken language systems, if the findings show that L2 speakers are able to better understand L2 speech in a more multimodal approach I would like to discuss whether this is applicable to any dialogue system.

2 Future of Spoken Dialogue Research

Taking into consideration the increasing use of speech technology and the increasing number of studies in the area of L2 speech perception and production, I would like to think that further research could somehow discuss the findings of language studies in order deepen the discussion of human-machine interaction.

As for the improvement of human-machine interaction, I believe that machines should be more able to offer an output that is more easily recognised copping with differences between a native speaker and a nonnative speaker of a given language. As regards dialogue for language learning, further studies could investigate how teachers and students, for instance, could benefit and make use of dialogue systems according to their learning needs.

3 Suggestions for Discussion

- Is the multimodal approach applicable and suitable to any dialogue system?
- If a multimodal approach is applicable, how could dialogue systems benefit from that?

References

C.T. Best and M.D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, Ocke-Schwen and Murray J. Munro (Eds.), *Second language speech learning: The role of language experience in speech perception and production*, pp. 13–34. Amsterdam: John Benjamins.

J.E. Flege. 1995. Second Speech Learning: Theory, Findings, and Problems. In W. Shange (Ed.), *Speech Perception and Linguistic Experience – Issues Cross-Language Research*, pp. 233–277. Timonium: York Press.

D. Hardison. 1999. Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. In *Language Learning* 49, pp. 213–283.

V. Hazan, A. Sennema, A. Faulkner, M. Ortega-Llebariad, M. Iba, and H. Chung. 2006. The use of visual cues in the perception of non-native consonant contrasts. In *Journal Acoustical Society of America*, 119 (3), pp. 1740–1751.

D. C. Kluge, A. S. Rauber, M. S. Reis, and R. A. H. Bion 2007. The relationship between perception and production of English nasal codas by Brazilian learners of English. In *Procs. of Interspeech 2007*, pp. 2297–2300.

D. C. Kluge, M. S. Reis, D. Nobre-Oliveira, and M. Bettoni-Techio. 2007. The use of visual cues in the perception of English syllable-final nasals by Brazilian EFL learners. In *New Sounds 2007: Procs. of the 5th International Symposium on the Acquisition of Second Language Speech*, pp. 274–281.

P. K. Kuhl and P. Iverson. 1995. Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, pp.121–154. CityTimonium, StateMD: York Press.

Biographical Sketch

Denise Cristina Kluge is a Professor of Applied Linguistics at the Federal University of Paraná, Brazil particularly interested in English Phonetics and Phonology learning. She received her PhD and her MA in the same field in 2009 and 2004, respectively. Her BA is in Portuguese and English Teaching in 2000. In 2007 she completed part of her PhD studies at University of Amsterdam under the supervision of Prof. Paul Boersma.

Theodora Koulouri

Department of Information Systems and
Computing, Brunel University
Uxbridge, Middlesex
UB8 3PH
UK

theodora.koulouri@brunel.ac.uk

1 Research Interests

My research interests lie in the area of **natural language interfaces** in **Human-Robot Interaction**. In particular, my research efforts are currently focusing on enriching the dialogue manager of a mobile personal robot with capabilities of **natural (mis)communication management**. My thesis follows an empirical approach and is motivated by **collaborative models** of human communication, also viewing Human-Robot Interaction (HRI) as a primarily bilateral process. It explores the linguistic resources employed by **users** in dialogue-based navigation of a robot. On the other hand, building upon human dialogue strategies, my aim is to identify the **feedback and repair mechanisms** that a **robot** should possess within the domain of goal-oriented HRI. The platform and testbed of my research is a speech-enabled robot that is capable of executing and learning navigation tasks by means of unconstrained natural language (Lauria et al., 2001).

1.1 Past and Current Work

We conducted a series of Wizard of Oz simulations in order to obtain information on the range of utterances that users would produce when interacting with the robot as well as specify task and system requirements (Koulouri and Lauria, 2009b). As mentioned above, I am looking at how a robot should initiate repairs and provide feedback, and therefore, the wizards were also subjects of the study. The study also explored how visual shared space and monitoring shaped the interaction patterns of both participants. Thus, the experimental design involved two conditions in which the user could or could not monitor the robot. The study yielded a corpus of route instructions and also defined a closed set of robot error handling strategies (Koulouri and Lauria, 2009a). Moreover, differences in coordination patterns and linguistic choices were observed depending on condition. In brief, when the robot was executing without the direct supervision of the user, the responsibility for establishing task status and understanding shifted towards the robot that provided rich descriptions and feedback. Yet, when users monitored the robot, verbal feedback by the wizards was often omitted. Similarly, when visual information was available, users' spatial descriptions tended to be less detailed and precise. The results indicate that the demands for spatial reasoning and inferential capacities could be higher for collocated, supervised robots. On the other hand, (semi-) autonomous execution requires more sophisticated interactive abilities.

1.2 Future Work

As suggested in the previous section, providing effective and timely feedback is crucial for task-oriented interactions, and especially in the dynamic setting of HRI, in which the user's instructions can be incomplete or outdated. The amount and placement of feedback should be decided upon several knowledge sources combined in a single criterion that is adaptive both within and between interactions. These sources could be the history of the on-going dialogue (e.g., how many times so far the robot and user have initiated repair), model of the environment (e.g., is the robot at home, outdoors or at a crowded workplace) and the task (e.g., is the route well-known, what are the consequences of errors). My future work will deal with the implementation of such functionality.

Moreover, according to a study on spatial descriptions by Mills and Healey (2006), clarification requests have a direct effect on the processes of coordination. They observed that as the dialogue progresses, the interlocutors converge in the use of more complex and efficient spatial descriptions. However, after clarification subdialogues, the instructor shifts to more conservative descriptions. These insights coming from human communication present interesting questions and rich opportunities for investigation in HRI. The next step in my research is to examine whether and how users revised and adapted their strategies, within the course of the dialogue, in response to particular robot utterances and in the presence of miscommunication. Finally, it would be interesting to determine whether certain user responses (as primed by previous repair initiations by the robot) are more efficient in terms of recovery from error.

2 Future of Spoken Dialogue Research

Spoken interfaces have been successfully embedded in numerous practical systems, which are mostly telephone-based, information-seeking applications. Recent progress in robotic technologies brings closer the vision of ubiquitous and commercial use of robots. Natural language apart from being the most intuitive and expressive means of communication is also a powerful tool for instruction, most appropriate for robots that essentially have to "learn on-the-job". Therefore, I believe that interaction with embodied agents, able to assist and collaborate with people in their everyday activities, is what the near future demands and holds for the dialogue systems community.

However, dialogue-based HRI, in which robots and users coordinate verbal and physical actions sharing time and space, poses several hard but exciting challenges. Unlike most dialogue systems, the deployment environment of a robot cannot be modelled a priori. Moreover, robots are typically built on agent-based architectures (that is, the systems consist of several components managing the dialogue and the robot actions and possibly sensor devices). Situated dialogue entails instantaneous synchronisation and updating of these components to include a continuous flow of information. Further, physical co-presence enhances the user's perception of common ground, increasing the use of spatiotemporal reference and, thus, the occurrence of mismatches. In particular, spatial language can be underspecified and arbitrary requiring the application of multiple layers of discourse and situational context. Thus, in navigation tasks, the robot needs to have understanding of language, spatial actions and relations as well as perception of the world. Finally, misunderstandings in HRI might have safety implications. In conclusion, the scope, occurrence and costs of miscommunication increase, impelling researchers to integrate miscommunication management into the design of spoken interfaces for robots.

Such dialogue models should certainly be computationally practical but should also be based on models of human communication as well as empirical studies in the domain of HRI. Endowing artificial agents with "human-inspired" strategies holds the promise of natural and robust management of miscommunication. Moreover, in HRI, like human collaborative behaviour and communication, omniscience is not required, but coordination of information and knowledge states. Thus, I believe that redefining HRI as collaboration and embracing errors as inherent in communication could enable us to transfer persistent problems within speech recognition technology and HRI (arising from uncontrolled environments, unknown tasks, system synchronisation, interpretation of raw sensor data etc.) into the dialogue domain and deal with them as triggers for further interaction in the form of verbal instruction, feedback, clarification and grounding. Therefore, research efforts should focus on developing dialogue models that equip embodied agents with such interactive capabilities.

3 Suggestions for Discussion

- Evaluation metrics for assistive robots with speech and multimodal interfaces.
- To what extent dialogue models and miscommunication frameworks from computer-based dialogue systems and human interaction can be applied to embodied agents.
- What additional challenges does long-term interaction entail for the dialogue manager of personal/service robots. How can the dialogue manager accommodate emergent language and social relationships.

References

Gregory J. Mills and Patrick G. T. Healey. 2006. Clarifying Spatial Descriptions: Local and Global Effects on Semantic Co-ordination. In *Procs. of the 10th Workshop on the Semantics and Pragmatics of Dialogue*.

Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and Ewan Klein. 2001. Training Personal Robots Using Natural Language Instruction. In *IEEE Intelligent Systems*, pp 38–45.

Theodora Koulouri and Stasha Lauria. 2009a. Exploring Miscommunication and Collaborative Behaviour in Human-Robot Interaction. In *Procs. of SIGDial'09*.

Theodora Koulouri and Stasha Lauria. 2009b. A Corpus-based Analysis of Route Instructions in Human-Robot Interaction. In *Procs. of TAROS'09*.

Biographical Sketch

Theodora Koulouri is a PhD student at Brunel University under the supervision of Stanislao Lauria. For the requirements of her MSc degree in Speech Sciences (UCL), she implemented a system for far-field recognition of spontaneous speech. She also holds a BA in Linguistics from the University of Athens. Among other happy distractions from research and teaching, she enjoys doing sports and drawing (mostly robots and llamas).

Christine Kühnel

Technische Universität Berlin
Deutsche Telekom Laboratories
Quality and Usability
Berlin, Germany

christine.kuehnel@telekom.de

1 Research Interests

My main research interest is on **multimodal human-machine interaction** in the context of a smart-home dialogue system. My dissertation work focuses on three major topics: (1) I am currently developing a multimodal smart-home dialogue system, based on the EU-funded IST-project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces; IST 2001-32746). The system extends home appliances with a speech interface. During the past months and the next year both, new output and input components are added with the aim to develop a multimodal smart-home system for research purposes.

(2) The **evaluation** of the system and its components is conducted either with standardised or – where necessary – new methods. The first part of it, the evaluation of system output will be briefly described in the next section.

(3) Based on earlier work on the quality of spoken dialogue systems (Möller, 2005) Sebastian Möller and colleagues proposed a taxonomy of performance and quality aspects, including as well influencing factors of the system, the user, and the context of use (Möller et al., 2009a). Part of my dissertation work is to further analyse the interrelations between the different aspects as given in the taxonomy.

1.1 Previous Work

The last year I focused on the output side of human-machine interaction. Four experiments on the quality of talking heads as output components to the smart-home system have been conducted by me and my colleague Benjamin Weiss. The results of a passive listening-only experiment (Kühnel et al., 2008), a web-based experiment (Weiss et al., 2009a; Weiss et al., 2009c), an interactive experiment with a simulated smart-home system (set in a laboratory) (Weiss et al., 2009b) and an interactive experiment with a fully functional smart-home system (set in a living room) (Kühnel et al., 2009) were analysed and compared to address the following main questions:

- Concerning the quality of talking heads, which aspects of talking heads contribute to the perceived quality and how?
- Concerning evaluation of talking heads, is a web-based experiment as valid as a laboratory-based experiment and how does interaction influence the quality ratings?
- Concerning the smart-home domain, is a talking head a suitable user interface and how does it influence system quality?

1.2 Current and future work

The next step is to focus on the input side of the INSPIRE system. The original system offers speech input only, depending on automatic speech recognition. Given the low recognition rates achieved, especially under the conditions found in the home (multiple speakers, background noise, distant speech) analysing different interfaces seems worthwhile. Furthermore, the majority of potential users is still not used to have a conversation with a machine. On the other hand, speech input offers advantages – as for example in hands-busy situations – that can not easily be provided by other means. Combining multiple input modalities might allow to compensate for the weakness of one modality by the strength of another and vice versa. It is often argued that multimodal interaction is intuitive because it resembles natural interaction. Apart from a talking head that can show facial expression – used for example for back channeling during a conversation – this would imply to render the system sensible to information conveyed via facial expressions and gestures. This seems a promising approach in dialogue-heavy applications and a lot of work is being done in the area of conversational agents. But in the smart-home domain this would on the one hand require omnipresent cameras while on the other hand quite a few interactions are task-driven and might not depend on a sophisticated dialogue were natural interaction is especially helpful.

But there are other means of interaction being or becoming more prevalent in recent years: GUI-based interaction and mobile interaction with a mobile phone as user interface. With the recent smart phones a broad range of user interfaces are feasible. With accelerometers and touch screens, 2D and 3D gesture recognition can be used for user input while the screen offers another mode for system output.

While writing this I am working on coupling an iPhone with the INSPIRE system both as a sole interface as well as combined with the speech input and output modality already existent. A major part of this work is the development of a fusion module on a semantic level. As the system is mainly developed for experiments a wizard interface to bypass parts of the system (ie. the ASR, the gesture recognition or the fusion) is developed at the same time.

The final system will be used for experiments with the aim to find relations between the quality of system modules and the overall system quality, as well as correlations between user ratings accumulated with questionnaires and system performance. Based on the taxonomy introduced above quality aspects important for multimodal systems will be identified and characterised and their interrelations defined.

2 Future of Spoken Dialogue Research

With ever evolving new user interfaces, dialogue systems relying purely on speech will most likely be confined to niche applications were its advantages outweigh the usability constraints of dissatisfying recognition results (compared for example to GUI-based interaction). But the need for advanced dialogue systems will increase as more and more of our every day world is equipped with machines replacing for example the woman behind the counter or combining different household equipment to complex robots. My grandfather is not longer able to by a ticket for a train ride and even people of younger age might be unable to cope with new technology. At the same time these technologies might offer a new freedom to disabled people granted an intuitive interface is provided. This does not necessarily require human-like interaction in the way that human interact amongst themselves. As long as the user knows a machine is involved (s)he will act accordingly. But dialogue systems that offer a choice of input and output modalities to cater for different user needs, different tasks and different situations could be the answer to many problems. This, of course requires a thorough understanding of the requirements and well established methods and metrics for evaluating systems and system components.

3 Suggestions for Discussion

- Does an interaction need to be human-like to be 'natural' and intuitive?

- How to achieve realistic setups for evaluation studies and how to motivate participant to give 'real world' ratings.

- Privacy issues: should the dialogue always be commenced by the user? Is it desirable to have an ever alert system that is always eavesdropping on you?

References

Christine Kühnel and B. Weiss and I. Wechsung and S. Möller and S. Fagel. 2008. *Evaluating Talking Heads for Smart Home Systems*. 10th Int. Conf. on Multimodal Interfaces (ICMI 2008), GR-Chania, 20-22 Oct. 2008, 81-85.

Christine Kühnel and B. Weiss and S. Möller. 2009. *Talking Heads for Interacting with Spoken Dialog Smart-Home Systems* Interspeech, Brighton, GB (accepted).

Sebastian Möller. 2005. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, US-New York NY.

Sebastian Möller and K.-P. Engelbrecht and C. Kühnel and I. Wechsung and B. Weiss. 2009. *Evaluation of Multimodal Interfaces for Ambient Intelligence*. accepted for: Human-Centric Interfaces for Ambient Intelligence (H. Aghajan, R. López-Cózar Delgado and J. C. Augusto, eds.), Elsevier.

Benjamin Weiss and C. Kühnel and I. Wechsung and S. Möller and S. Fagel. 2009. *Comparison of Different Talking Heads in Non-Interactive Settings* 13th Int. Conf. on Human-Computer Interaction (HCI International 2009), 19-24 July, US-San Diego CA.

Benjamin Weiss and C. Kühnel and I. Wechsung and S. Möller and S. Fagel. 2009. *Quality of Talking Heads in Different Interaction and Media Contexts*. Special Issue of Speech Communication (accepted).

Benjamin Weiss and C. Kühnel and I. Wechsung and S. Möller and S. Fagel. 2009. *Web-based evaluation of talking heads: How valid is it?* 9th Int. Conf. on Intelligent Virtual Agents, 14-16 September, Amsterdam, Netherlands (accepted).

Biographical Sketch

I am working as a research assistant at the Quality and Usability Lab of Deutsche Telekom Laboratories, TU-Berlin. I studied Electrical Engineering and Business Administration and received my diploma degree in 2007 from the Christian-Albrechts University of Kiel. I am working towards my PhD thesis in the domain of evaluation of multimodal systems. My extracurricular interests include tango argentino, violin playing, sports and outdoor activities like cycling, hiking and canooing.

Catherine Lai

Department of Linguistics
University of Pennsylvania
Philadelphia PA 19104, USA

laic@ling.upenn.edu
www.ling.upenn.edu/~laic

1 Research Interests

My research interests lie in modelling the contribution of **prosody** to meaning in dialogue. In particular, I am interested in how prosody can be integrated into **formal models of dialogue**. This involves studying how prosodic features fit in with theoretical concepts like an **intonational lexicon**, how well these sorts of theoretical models fare on speech data from real dialogue, and how measurable phonetic data can best be analysed and represented to address these questions. I am also interested in how **machine learning** and **speech recognition** tools can be used to investigate these problems.

In general, these studies are about understanding how prosody can be used determine different types of **speech acts** and to structure dialogue. I also hope they will shed light on some theoretical questions about questions themselves: what makes an utterance a question, what makes an answer, and when can an utterance with question characteristics safely be ignored.

1.1 Previous Work

The overarching goal of my research project is to provide a robust and testable framework of prosody in dialogue. To develop robust models, we need to develop methods of analysis on data sets which may not be annotated for prosody or discourse metadata. However, before making hypotheses about the full range of dialogue moves, we need to consider smaller sets of data which are tractable in terms of investigating fine phonetic detail, semantic/pragmatic contributions, and their interaction. More importantly we want to study data that are consistently and frequently used to shape the structure of the dialogue.

As such, my previous work has focused on studying the semantics, pragmatics and prosody on **cue words**. These **discourse markers** are important for maintaining dialogue coherence and have a wide range of uses. For example, the set of English cue words includes backchannels like *uh-huh* and *okay*, agreements like *right*, and questioning particles like *really*. Understanding these sorts of markers is important because they indicate both when things are going well and when things are going badly.

Individual cue words can express more than one interpretation and prosody contributes to this interpretation (Gravano, 2009). However, to understand this contribution we need appropriate dimensions of meaning with which to view the data. In Lai (2008), I investigated *really* which, as a one word turn, is classified as both a backchannel and question in a metadata annotation of the Switchboard (MDE 2003, LDC2004T12). That study was an attempt to find out if prosodic features (e.g. pitch contour type, intensity, duration) could distinguish these two categories of *really*s. Statistical analysis and machine learning attempts indicated that this was not the case. In fact, *really* appears to span a spectrum of meanings/uses ranging from acknowledgement to questioning to exclamative.

So, to model how these turns are interpreted we need to know not only what dimensions of meaning prosody *can* work on, but also how this combines with the syntax/semantics of the turn. *Really* and *right* were used to probe this interaction in Lai (2009). *Really* acts as a check on the common ground and seems to be interrogative as a one word turn. On the other hand, *right* is an agreement word which can appear as questioning in very specific contexts: as a tag question with rising intonation. A perception experiment of *really* and *right* found that ratings of surprise were positively correlated with ratings of how questioning the cue word sounded. However, the ratings for *really* did not translate to ratings for *right* with similar prosodic values. So, the semantics/pragmatics of cue words suggest different thresholds for surprisingness when they are heard in *isolation*. The next step is to look at how these perceptions are integrated in the dialogue.

1.2 Current and Future Work

My previous studies were small steps in understanding the prosody of dialogue alignment. What we want from this project is a way to detect when there is agreement in dialogue and when there is a need for repair. My current work is focuses on cue words in formal models of dialogue (e.g. Fernandez (2006), Schlangen (2004)). I am currently performing a corpus study looking at the rhetorical relations that structure the discourse around *really* and *right*, how this structure differs with varying prosody, and how this directs the Question Under Discussion or *discourse topic*. This feeds into further perception experiments of cue words in context. The goal of this is to understand how the range of speaker affect and attitude available with words like *really* translates into the categorisations and relations presented in semantic models.

A further goal of my project is to see whether the experimental and corpus results can be used to guide larger scale corpus analysis for better automatic classification of speech acts. However, I hope the theoretical background together with the corpus coverage will also help us choose between models of how we represent time varying prosodic features for dialogue like pitch contours (Prom-on et al. (2006), Kochanski and Shih (2003), *interalia*).

2 Future of Spoken Dialogue Research

For spoken dialogue systems to be usable they must be able to detect when things are going wrong and provide strategies for correction. This means detecting speech production errors as well as pragmatic misalignment. Clearly, given my line of research, I think prosody is key for handling this. However, it is still unclear what (abstract) representation is appropriate for prosodic features. Moreover, current research is significantly slowed by the fact that creating manually annotated corpora is an extremely painstaking process. So, the development of automated methods for extracting prosodic information is extremely important. In particular, this means dimensionality reduction for features like F0 contours. Development of such techniques seems viable in the next 5-10 year, especially as statistical/speech technology methods start to creep into theoretical linguistic research on discourse and phonology.

3 Suggestions for Discussion

- How can we model the distribution of the load between prosody and non-speech features in dialogue? When is prosody crucial for understanding dialogue and when can it be ignored?

- How can test theories of prosodic meaning and dialogue structure using real conversational corpus data? How far can we get with unannotated speech corpora?

- How can we model/detect speaker affect and attitude? How do we know when a dialogue is going well? How does this affect turn taking and the choice of conversational moves?

References

R. Fernandez. 2006. Non-Sentential Utterances in Dialogue: Classification, Resolution and Use. *PhD thesis*, Department of Computer Science, Kings College London, University of London.

A. Gravano. 2009. Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue. *PhD thesis*, Columbia University.

G. Kochanski and C. Shih. 2003. Prosody modeling with soft templates. *Speech Communication*, 39 (3-4), pp.311–352.

C. Lai. 2008. Prosodic Cues for Backchannels and Short Questions: *Really?* In *Proceedings of Speech Prosody*, Campinas, Brazils.

C. Lai. 2009. Perceiving Surprise on Cue Words: Prosody and Semantics Interact on *Right* and *Really*. In *Proceedings of Interspeech*, Brighton, UK.

S. Prom-on, Y. Xu, and B. Thipakorn. 2006. Quantitative Target Approximation Model: Simulating Underlying Mechanisms of Tones and Intonations. In *Procs. of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 1.

D. Schlangen. 2004. A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue. *PhD thesis*, School of Informatics, University of Edinburgh.

Biographical Sketch

Catherine is pursing a PhD in linguistics at the University of Pennsylvania. Her main interest is the interface between semantics, pragmatics and prosody, but she has also worked on the phonetics of speech disfluencies. Her other research hobby is formally modelling language change. Before this, she received a BSc (Hons) in Mathematics and Computer Science and an MSc in Computer Science at the University of Melbourne, Australia, where she worked on querying linguistic trees, amongst other things.

Marianne Laurent

Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France

marianne.laurent@orange-ftgroup.com
perso.telecom-bretagne.eu/
mariannelaurent/

1 Research Interests

The deployment of efficient Spoken Dialogue Systems (SDS) requires the use of a **dialogue design methodology**. Actual methods are not guided enough by an anticipation of a design adapted to the **user's practices**. Consequently they require numerous iteration cycles based on real interactions data to detect and correct design issues. As a result, these ad hoc processes mainly lead to long and costly development cycles.

In this framework, the goal of my thesis is to work on the definition of dialogue design methodologies able to put together some knowledge on the technical components, on the typology of services, and on the user's practices. My work relies on an **explicit formalisation of service evaluation**, with a view to identify problems more easily, to compare iterative versions of a service, and, if possible, validate best practices applicable to the design of future services.

My work should include three stages. A first stage dedicated to gather an understanding of several aspects of the dialogue between two people. A second one focused on the design of a SDS's logic of dialogue and its enhancement supported by evaluation tools. And the third one dedicated to the validation of best practices to feed the design process for future services.

1.1 A data driven approach

My study has started with the **collection of a corpus** of telephone communications held by a school secretary. The transcription reveals very rich interactions, addressing many kinds of subjects, with both known and unknown interlocutors. Our objective is to analyse these dialogues according to several point-of-views (theme, strategy, rhythm, etc.) to identify a few trends to be taken into account in the design of SDS dialogues. In this purpose, a deep reflexion has been con-ducted on the **transcription formalism** so that a statistical study of the data is made possible.

These findings will then be used to feed the design of a service related to some of the observed secretary task.

1.2 Evaluation for enhancement

The second stage, led by a **learning-by-doing approach**, consists in the design of a design, and its enhancement. This will be the occasion to both integrate the knowledge gathered thanks to the analysis of the human-human interactions, and think about the evaluation tools helping the iteration cycles.

Keeping the data driven approach, the sets of indicators will be deduced from success or failure situations identified in corpora. This opens two opposite perspectives for the evaluation process.

On the one hand, I will work on the identification of **red-light indicators**. Their goal will be to draw attention to the presence of problematical situations experienced in dialogues. They should be mainly used to spot issues to be fixed within the design of service which leads to iteration in conception to propose an alternative design.

On the other hand, work will be pursued on **performance indicators** to be used for either a/b testing of alternatives or automatic reinforcement learning integrated in the SDS. In (Laurent et al., 2009) we suggested a change in evaluation paradigm that consists in measuring the user's conformance to the system in the **system referential** instead of evaluating the system as regard to the users referential. This change suggests the selection of the indicators according to the service specifications, and not to a presupposed user satisfaction. Performance indicators will be obtained from the targeted functionalities of the service.

This work on evaluation will face a lot of issues. Among others, each intuitively selected metric will have to be challenged carefully as regard the performance objectives and user practice. We can also think about the problem of dispassion. Even if we deal with quantitative metrics, to what extend the designed evaluation paradigms can be objective? Last, we might have to handle the problem of commensurability and portability of the evaluation paradigms, question remarkably raised by (Paek, 2007).

1.3 Evaluation for learning

Beyond the use of evaluation for the comparison of alternatives, an interesting stake would be to broaden its role to the identification of clues for the design of services. I anticipate two perspectives to address this idea. The first one could be to **validate intuited best practices** on a certain amount of distinct contexts to consider them on trust for future developments. The second one would be a procedure that, on the basis of evaluation output, spots the element(s) to be modified within the design of service. Can evaluation help to recognise the causes of failure or success within a dialogue?

As a result, if we come to some results with these perspectives, and add them up, we could come to valuable tools to **industrialise the dialogue design process**.

2 Future of Spoken Dialogue Research

- A great challenge to cope with is undoubtedly in the vocal recognition performance. This might come from the new STT technologies, such as the one used by Spinvox, who still has to adapt to the real time needs of the SDS services. Anyway, we can anticipate a little revolution in the way of apprehending dialogue when recognition possibilities will skyrocket!

- Above the recognition of words pronounced by the user, we can hope for the future ability of system in the recognition of nonverbal elements within the user's utterances. Prosody or rhythm could strongly help the service in both understanding the intentions, and synchronising with the user's tense.

- Last I believe that another change to pay attention to is the evolution in the way the mainstream perceives SDS solutions. Will they always be compared to the preexisting human operator's services, or will they on day be seen as another human-machine interface, just as the web is. This could change the definition given to naturality in interaction, shifting from a search for humanlike behaviour to an attention given to interface usability.

3 Suggestions for Discussion

- Data-driven analysis: To what extend human-human dialogues corpora can be helpful for the study of human-machine dialogue?

- Evaluation: What are the issues triggered by the absence of dispassion in evaluating SDS? Is dispassion of the evaluation a goal in itself?

- Limitations of Spoken Dialogue Systems: To what extent a SDS performance could be enhanced by taking into account additional aspects of the users' utterances that are ignored today (prosody, paralanguage, etc.)?

References

Marianne Laurent, Ghislain Putois, Philippe Bretier, Thierry Moudenc. 2009. Nouveau paradigme d'évaluation des systèmes de dialogue humain-machine. In *TALN*.

Sebastian Möller. 2005. Quality of Telephone-based Spoken Dialogue Systems. Springer.

Tim Paek. 2007. Toward evaluation that leads to best practices: Reconciling dialogue evaluation in research and industry. In *Proc. of the NAACL-HLT Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*.

Biographical Sketch

Marianne joined the Orange Labs in Lannion, France, as a Ph.D. student in 2008, her thesis being supervised by Philippe Bretier (Orange Labs) and Ioannis Kanellos (Telecom Bretagne). After a Master in Management from Grenoble Ecole de Management, she graduated in 2007 with a Master of Engineering from Telecom Bretagne. Her studies, both technical and managerial, mainly focused on the Management of Enterprise Information Systems. She started her career in IS project management within Thales in the UK before she joined the dialogue team of Orange.

Beatriz López Mencía

Universidad Politécnica de Madrid
Ciudad Universitaria s/n
28040 Madrid
Spain

beatriz@gaps.ssr.upm.es

1 Research Interests

I am pursuing a research career in the field of Human-Computer Interaction, and my special interest is in the design and **usability evaluation of multimodal interfaces and embodied conversational agents**. My thesis topic explores the use of embodied conversational agents (ECAs) and their visual communicative ability to **improve typical interaction problems and the users' overall experience**.

1.1 Past and ongoing research

The main line of research I have been following at Grupo de Procesado de Señales(GAPS), Universidad Politécnica de Madrid(UPM), is the comparative study of a variety of contexts of multimodal human-machine interaction, with a special focus on embodied conversational agents (ECAs). For this purpose we have built several platforms to test interaction through various modalities in the context of use imposed by different applications, including biometric identification and remote access to home devices. My primary goal is to improve the robustness of human-computer communication and to foster a better user experience, and to explore how ECAs might be useful in these respects.

Firstly we focused on evaluating multimodal biometric systems, specifically, identity verification system using voice recognition. We sought for effects on system usability efficiency and user acceptance related to the presence of an embodied conversational agent (ECA). We designed a spoken dialogue interface which incorporates an animated figure making gestures tuned to specific verification dialogue stages. The results obtained suggest that, while interaction with the animated agent was more efficient and enjoyable, users had stronger privacy concerns (Hernández et al., 2007).

A following study focussed on robustness problems with spoken language dialogue systems. Adding a visual channel with an animated character that personifies the artificial system could make it possible to convey communication cues through body and facial gestures which might help the user follow the interaction more easily (e.g., by displaying metacognitive gestures suggestive of the mood and perhaps also of the content of what the ECA intends to communicate, reinforcing the information conveyed verbally). For this purpose we identified typical interaction problems with SLDSs and associated with each of them a particular ECA gesture or behaviour. The goal was to keep users in a positive frame of mind when recognition errors occurred. We also tested a proxemic 'code' to mark dialogue turns in order to make the interaction more fluent and the users more confident (López Mencia et al., 2008).

Our evaluation work was focused on special features that should be taken into account with a face-to-face communication system. This time we tested an application to remotely control home devices (i.e., a domotic system). Our evaluation scheme was chiefly based on ITU standards for the evaluation of (strictly) spoken dialogue systems (ITU, 2005). Our test results revealed that ECA users scored better on certain important objective performance parameters. We emphasised the need for in-depth factor analyses of the users' subjective experience with face-to-face dialogue systems. The results suggested that rejection is a dimension in its own right that may strongly influence subjective evaluation parameters closely related to user acceptance. Some of the extracted 'subjective' dimensions show that our ECA seems to help users to better understand the flow of the dialogue and reduce confusion.

My present research is focussed on annotation procedures, data processing and statistical analysis of data (user opinion and system performance). Another ongoing research line is a joint work between UPM and several schools of disabled children in Madrid. This work is based on the collaborative design, development and evaluation of applications that might reinforce learning in children with several disabilities (autism, cerebral palsy...). We hope that these children may benefit from learning improvements in specific areas and increase their learning motivation by using new advanced technologies (such as ECAs). At present we are working on the evaluation of a specific application using ECA technology and webcams in order to boost the learning of emotions in physically and cognitive disabled children with severe speech and motor limitations.

2 Future of Spoken Dialogue Research

A new generation of multimodal interactive systems is in the making that promises to boost user context awareness and multidevice interaction. I believe, according to this line, that we should take advantage of new communications channels such us the visual one. The integration of a visual channel to spoken dialogue systems could **enrich the interaction**. An interesting possibility is the incorporation of Embodied Conversational Agents to Spoken Dialogue Systems using the visual channel. These ECAs may contribute to boosting the social abilities of the system and also humanising the system in specific applications (learning with children, applications for old people...) in which these aspects are very important. A visual communication channel could also be exploited to carry supralinguistic information regarding **affective interaction components** (expectations, intentions, mental processes and emotions). The goals of interaction could then become less task- and domain-specific and more **open and social in nature**, perhaps even allowing meaningful long-term relationships to develop between the user and an animated agent that shows some sort of 'personality' (López Mencia et al., 2008). Another research challenge is to design an interaction adapted to users' needs, and offering accessibility for all users. Minimising interaction errors is also a key aspect of user experience that should be dealt with. A visual channel would introduce new problems, however, such as the possibility of misinterpreting the other party's gestures. In any case, it seems worthwhile to explore the possibilities that interaction so enriched may offer. In order to tackle these challenges, standardisation work is very important. Defining standard evaluation parameters would allow comparing systems in terms of their usability and user-centred quality parameters. Also, research in standards and the implementation of nonverbal information (e.g. generation of nonverbal output information [EML, 2008]) would be desirable for the future of spoken dialogue research.

3 Suggestions for Discussion

Some suggestions for discussion are the following:

- How can we identify interaction errors in a spoken dialogue system enriched with a visual communication channel?

- Definition of metrics for evaluating spoken dialogue systems with a face-to-face communication channel.

- How does the way nonverbal behaviour is implemented influence the interaction? Is there a compromise that has to be reached between attaining naturalness in nonverbal behaviour and obtaining good values in dialogue metrics for performance?

- Evaluation: How to define standard procedures to collect corpora? What methodologies are most appropriate for analysing objective and subjective data (e.g., questionnaire responses)?

References

A. Hernández, B. López, D. Díaz, R. Fernández, L. Hernández, and J. Caminero. 2007. A person in the interface: effects on user perceptions of multibiometrics. *Workshop on Embodied Language Processing, in the 45th Annual Meeting of the Association for Computational Linguistics*: 33–40, ACL, Prague.

B. López Mencia, D. Díaz de Vera, A. Hernández Trapote, M. C. Rodriguez Gancedo, J. Relaño, and L. Hernandez Gomez. 2008. Evaluation of ECA Gesture Strategies for Robust Human-Computer Interaction. *International Journal of Semantic Computing*, Vol. 2, No. 1: 1–24.

I.-T. S. t. P.-S. Rec. 2005. *ITU-T Suppl. 24 to P-Series Rec. Parameters Describing the Interaction with Spoken Dialogue Systems. International Telecommunication Union, Geneva (2005).*

EML. 2008. *EML: Elements of an EmotionML 1.0. W3C Incubator Group Report.* November. `http://www.w3.org/2005/Incubator/emotion/XGR-emotionml-20081120/`

Biographical Sketch

Beatriz López Mencía has an MEng in telecommunications engineering from Universidad Politécnica de Madrid. She is currently a 3rd year PhD student and holds a research position in the same university, under the supervision of Prof. Luis Hernández.

Syaheerah L. Lutfi

Speech Technology Group (GTH)
Department of Electronic Engineering
Universidad Politécnica de Madrid
Spain

heerah@die.upm.es
www-gth.die.upm.es

1 Research Interests

Human-Machine Interaction in general and specifically **Affective Computing** is my field of interest. Currently I am involved in a project for **modelling affect in a robotic agent** for domestic purposes. In a smaller scope, I am interested in researching and developing **culture-sensitive emotive interfaces**. This is motivated by the fact that perception and expression of emotions vary among people from different cultures, and therefore designs of emotive computer interfaces should be tailored to conform to the different needs according to cultural background.

1.1 Previous Experiences

Emotional prosody modelling for a Malay TTS

Emotional prosody modelling for a Malay TTS I became involved with speech studies when I worked on my Masters dissertation. I worked on a project in collaboration with the R&D Company, MIMOS Malaysia to incorporate emotions into a Malay TTS. It involved the incorporation of an affective component to the Malays TTS system, in order to produce a system that is more expressive in nature. I introduced a new template-driven method for generating expressive speech by embedding an 'emotion layer' called **eX**pressive **T**ext **R**eader **A**utomation, abbreviated as eXTRA. The module is an independent component that can serve as an extension to any Malay TTS system that uses Multiband Resynthesis Overlap Add (MBROLA) engine for diphone concatenation. Details can be found in (Lutfi et al., 2006).

Culture-sensitive TTS

It is vital to ensure that intelligent interfaces are also equipped to meet the challenge of cultural diversity. Studies show that the expression and perception of emotions may vary from one culture to another (Silzer, 2001) and that the localised synthetic speech of software agents, for example, from the same ethnic background as the interactor are perceived to be more socially attractive and trustworthy than those from different backgrounds (Nass and Lee, 2002). Based on these studies and personal experiences, we realised that it is crucial to infuse a more familiarised set of emotions to a TTS system whereby the users are natives. We further worked on establishing a localised TTS by concentrating on the culture-specific manner of speaking and choices of words when in a certain emotional state. Evaluations from a localised TTS that we established show that the risk of evoking confusions or negative emotions such as annoyance, offense or aggravation from the user is minimised (Syaheerah et al., 2008).

1.2 Recent work

Emotion Identifications in Speech

In my first year of PhD I worked on identifying emotions from speech for a couple of class projects. One of them was concerned with obtaining the best parametric model for emotion recognition, based on a Hidden Markov Models (HMMs) classifier. The optimised parameters for the emotion identification task were determined empirically, across two representations of observations, the mel-scale cepstral co-efficient (MFCC) and also Perceptual Linear Prediction (PLP), and were improved using well-known normalisation techniques. A more important finding showed that certain representations were better or more precise at identifying a certain type of emotion over the other. This and other findings from the experiments are discussed in (Syaheerah et al., 2009). In the article, it is also proposed that the findings could be applied to speech-based affect identification systems as the next generation biometric identification systems that are aimed at determining a person's 'state of mind', or psycho-physiological state.

Emotional TTS

In GTH we also work on Emotional Spanish TTS and participated in a number of Expressive TTS competitions such the Spanish Albazyn competition and INTERSPEECH Emotional Challenge 2009 (Barra-Chicote et al., 2009).

1.3 Current work

Currently I am involved in modelling emotion for a task-independent robotic agent by integrating a module of needs. Based on theories which adhere to the idea that an affect system is influenced by a motivation system, we propose an emotion framework for a task-independent autonomous agent by integrating a module of needs. The need framework is based on Abraham Maslow's motivation theory (hierarchy of needs). What makes the approach different from other appraisal-based approaches is the incorporation of the need-layers as a module that functions as a decomposer of task-specific events according to their importance and urgency. The incorporation of this module has given two advantages: scalability in terms of tasks and needs, whereby the agent's tasks and needs can be added or appropriately changed to suit various application domains, and a priority mechanism on tasks in a multi-tasking environment.

One of my future plans and hopes is to integrate a **culture module** in this emotional framework that would personalise the agent according to the target user's socio-cultural background.

2 Future of Spoken Dialogue Research

Despite the strong evidence for cross-cultural consistency in emotion appraisal processes and perceptions, there are cultural differences as well. Affective applications that are developed with a general framework of emotion according to Western standards (individualistic culture) may not be suitable to be adopted in an environment belonging to collectivist culture (East). A subtle level of awareness of difference in spoken dialogue (and also other modalities) is vital to go beyond stereotypes of how a particular affective conversational agent might differ from the mainstream. I believe analysis focusing on socio-cultural grounding would help with modelling dialogue framework or other aspects of affective agents that increase cross-culture acceptance.

3 Suggestions for Discussion

- Culturally-sensitive dialogue design: what are the factors to be identified in designing it? Is there structured information available?
- Evaluation design: There is no gold standard for evaluating the believability aspects of emotive agents. How to design it?
- Virtual/robotic agents: Do people really prefer human-like agents to characters with simpler appearance such as line-drawn cartoon characters?

References

C. Nass and K. Lee. 2002. Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. In *Journal of Experimental Psychology: Applied*, 7 (3), pp. 171–181.

Roberto Barra-Chicote, Fernando Fernandez, Syaheerah Lutfi, Juan Manuel Lucas-Cuesta, Javier Macias-Guarasa, Juan Manuel Montero, Ruben San-Segundo, and Jose Manuel Pardo. 2009. Acoustic Emotion Recognition Using Dynamic Bayesian Networks and Multi-Space Distributions. In *Procs. of Interspeech 2009*, Brighton UK.

P. J. Silzer. 2001. Miffed, Upset, Angry or Furious? Translating Emotion Words. In *Procs. of ATA 42nd Annual Conference*, pp. 1–6, Lost Angeles, CA, Oct 31-Nov 3.

Syaheerah L. Lutfi, Raja N. Ainon, and Zuraidah M. Don. 2006. Expressive Text Reader Automation Layer (Extra): Template-driven 'Emotion Layer' in Malay Concatenated Synthesized Speech. In *Proc. Oriental COCOSDA'2006*, Penang, Malaysia.

Syaheerah L. Lutfi, J. M. Montero, J. M. Lucas, and R. Barra-Chicote. 2009. Expressive Speech Identifications Based on Hidden Markov Model. In *Procs. of HEALTHINF'2009*, pp 488-493, Porto, Portugal, Jan 14-17.

Syaheerah L. Lutfi, J. M. Montero, Raja N. Ainon, and Zuraidah M. Don. 2008. Extra: A Culturally Enriched Malay Text to Speech System. In *Procs. of AISB'2008*, pp. 77–83, Aberdeen, Scotland, Dec 9-11.

Biographical Sketch

Syaheerah Lutfi is a PhD candidate at Universidad Politécnica de Madrid (UPM), Spain under MOHE/USM scholarship. Previously she obtained a BSc. Computing from Bolton University, UK, and a Masters in Software Engineering from University Malaya (UM), Malaysia under the (USM) scholarship. She enjoys travelling and meeting people as cultures of the world continue to amaze her. She is blessed with a loving husband and two beautiful children, Waheedah and Mash'al.

François Mairesse

Department of Engineering
University of Cambridge
Trumpington street
Cambridge CB2 1PZ

francois@mairesse.co.uk
mi.eng.cam.ac.uk/~farm2

1 Research Interests

My research interests lie generally in models of individual differences in language, with a special focus on **stylistic natural language generation**, **expressive speech synthesis** and **natural language understanding**.

1.1 Past work

I've recently completed my Ph.D. thesis on statistical models for detecting and conveying linguistic variation in dialogue systems. Although there are many ways to express any given content, most dialogue systems do not take linguistic variation into account in both the understanding and generation phases, i.e. the user's linguistic style is typically ignored, and the style conveyed by the system is chosen once for all interactions at development time.

Over the past few years, psychologists have identified the main dimensions of individual differences in human behaviour: the Big Five personality traits (Norman, 1963). The Big Five traits are hypothesised to provide a useful computational framework for modelling important aspects of linguistic variation. My thesis first explores the possibility of recognising the user's personality using data-driven models trained on essays and conversational data (Mairesse et al., 2007). I then tested whether it is possible to generate language varying consistently along each personality dimension in the information presentation domain. I implemented PERSONAGE: a language generator modelling findings from psychological studies to project various personality traits (Mairesse and Walker, 2007). PERSONAGE was used to compare various generation paradigms: (1) rule-based generation, (2) overgenerate and select and (3) generation using parameter estimation models that learn to produce recognisable variation without the computational cost incurred by overgeneration techniques. These generation methods were evaluated based on human judgements, showing that human judges can detect the personality conveyed by the system's utterances, even if multiple traits are projected simultaneously (Mairesse and Walker, 2008).

As far as natural language understanding is concerned, I recently presented a semantic decoding method that can learn to predict semantic trees from unaligned data, i.e. without any word-level semantic annotation, by recursively combining support vector machine classifiers predicting semantic concepts from n-gram features computed over the whole utterance. As unaligned data is cheaper to annotate, this methods reduces dialogue system development time compared with techniques requiring Treebank-style annotations. This technique was shown to perform as well as state-of-the-art semantic parsers (Mairesse et al., 2009).

1.2 Current and future work

I recently joined Cambridge's Dialogue Systems group in order to develop statistical models for semantic parsing and language generation, and to test these components within a fully trainable dialogue system. This work is part of the CLASSiC project funded by the European Union. In the future, I am planning to combine stylistic generation together with expressive text-to-speech synthesis, in order to produce coherent and convincing linguistic variation in dialogue systems.

2 Future of Spoken Dialogue Research

Spoken dialogue systems are still a long way from being widely adopted by the population, mostly because the current level of performance of the understanding components is not matching the user's expectations, and because the generation side is producing repetitive, limited outputs. Whereas a lot of research currently focuses on the understanding side, modelling linguistic variation can greatly improve the interaction in dialogue systems, such as in intelligent tutoring systems, video games, or information retrieval systems, which all require specific linguistic styles.

I thus believe that the future of dialogue system research is to produce systems that can control that variation in a scalable way, both for understanding all the pragmatic nuances of the user input, as well as conveying realistic outputs. This scalability will require moving towards data-driven approaches at all levels of the output generation, i.e. controlling the level of initiative, the language generation process and the speech synthesis module in a consistent way. While previous research has produced highly trainable components (e.g. speech recogniser, text-to-speech engine), young researchers will need to focus on how to make the language understanding, dialogue management and language generation components re-usable across domains.

3 Suggestions for Discussion

- Given an infinite amount of data and computational power, could dialogue systems be undistinguishable from human beings? And should they?
- Dialogue system personalisation: is it important to model the system's output style, and if so what kind of style or personality should the system convey for different dialogue applications (e.g. intelligent tutoring systems, tourist information presentation, financial information retrieval)?
- In the near-future, will data-driven methods remove the need for rule-based knowledge in all dialogue system components?

References

François Mairesse and Marilyn A. Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–503.

François Mairesse and Marilyn A. Walker. 2008. Trainable generation of Big-Five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL).*

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young. 2009. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of ICASSP.*

W. T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583.

Biographical Sketch

François Mairesse is a research associate at the University of Cambridge, working in the Machine Intelligence Lab as part of the EU CLASSiC project. He recently completed his Ph.D. thesis at the University of Sheffield supervised by Prof. Marilyn Walker, focusing on models of individual differences in dialogue systems. The thesis investigates techniques for the recognition of the personality of the user as well as the control of the personality conveyed by the system. In 2006, he did an internship at AT&T Labs in Florham Park working on paraphrase acquisition from web reviews. In 2004, he obtained a Master's degree in Computer Science and Engineering from the Université Catholique de Louvain in Belgium.

Matthew Marge

Carnegie Mellon University
Language Technologies Institute
School of Computer Science
Pittsburgh, PA 15213

mrmarge@cs.cmu.edu
www.cs.cmu.edu/~mrmarge

1 Research Interests

My research interests lie in the domain of **spoken dialogue systems** for **human-robot interaction**. As part of the TeamTalk project, I work on improving how multimodal robots can better interpret spatial language and their environment (Marge et al., 2009). Our project's domain requires that humans and robots work together on a "treasure hunting" task. I rely on the principle of **grounding** for my research efforts. Currently, I am exploring the capabilities of spatial perspective-taking in **human-robot dialogue**.

1.1 Spoken Dialogue Systems for Surveys

My undergraduate research included building a spoken dialogue system for course-based surveys with the assistance of colleagues in the Computer Science Department at Stony Brook University. My chief responsibilities were to design, implement, and evaluate an automated course evaluation system called the "Rate-A-Course System" (Stent et al., 2006). This system allowed people to set their own goals for the conversation, such as how many topics to discuss about a course and in what order to discuss them. For my honors thesis, I designed and conducted experiments with human participants to study how speech-driven dialogues can adapt to users. Our project goal was to understand what dialogue designs were most effective when interacting with users that have specific goals in conversation.

1.2 Adaptive Human-Robot Dialogue

In past research, I investigated the cognitive and social aspects of robotics at the Carnegie Mellon University Human-Computer Interaction Institute. I studied how robots might adapt to the existing knowledge of novices and experts via dialogue interaction. Pearl, an interactive robot from the CMU Robotics Institute, was used in our experiments. We investigated whether, given a topic of conversation such as cooking, Pearl should use technical terms with experts but longer explanations of those terms with novices. One of my responsibilities was to enhance Pearl's existing adaptive dialogue system. I did this by increasing the number of appropriate responses Pearl could give to participants' questions about cooking tools. When we compared responses of novices and experts, we found that novice cooks appreciated Pearl and performed best when it gave detailed explanations of the tools rather than the tool names alone (Torrey et al., 2006). By contrast, expert cooks found Pearl patronising when it gave them detailed explanations of the tools.

As an intern at the Naval Research Laboratory, I developed a scenario for disambiguating dialogue in a human-robot team (Fransen et al., 2007). In this scenario, two people are directing a mobile robot to retrieve an item from a dangerous area. In order to improve the ability of the robot to disambiguate dialogue spoken by the team members, I worked with graduate students to integrate several functionalities into the robot, including gesture recognition, sound localisation, and natural language understanding.

1.3 Spatial Human-Robot Dialogue

Currently, I am working with other members of the TeamTalk project and the larger Boeing TreasureHunt project under the supervision of Dr. Alex Rudnicky. This project investigates how robots can better collaborate with humans using speech and dialogue with the goal of finding "treasures" in a real-world location. I helped prepare the current version of the virtual TeamTalk system, along with a research associate and a summer intern. This simulation, built using the USARSim system, provides us a vehicle to test our spoken dialogue system without the need to manage actual robots (Balakirsky et al., 2006). The system is built using the RavenClaw/Olympus Dialogue Architecture (Bohus et al., 2007).

I am currently exploring the capabilities of spatial perspective-taking in human-robot dialogue. We want to find out how spatial perspective-taking, both in reference to members of a human-robot team and to objects in the environment, can be incorporated into the current TeamTalk platform. This included developing a small scenario that we could begin to study. I am also learning more about the dialogue concept of grounding for this work.

Our exploration of spatial perspective-taking in human-robot dialogue has led us to design an experiment to assess how humans give simple dialogue commands in reference to members of a human-robot team. We have developed this experiment in order to conduct it formally with human participants. I have also conducted a literature review of spatial reasoning, spatial language, and its applications in human-robot interaction.

We are currently expanding the experimental stimuli from snapshots of scenarios to real-time interaction with the TeamTalk virtual system. Once we are established on a spatial reasoning component that can refer to team members in a scenario, I plan to extend the component to reference objects in the environment of the human-robot dialogue scenario. This will require developing a basic ontology of the environment that will be shared by humans and robots in the scenario. We will focus on using TeamTalk's virtual system for developing and testing this component.

2 Future of Spoken Dialogue Research

Interactive spoken dialogue applications will steadily increase in popularity over the next decade. This is dependent upon public acceptance of these systems, which have often frustrated the general public due to poor design and implementation. We should see a decrease in the number of call centers answering customer service phone lines, for example. I also expect that as the field of robotics expands, dialogue will become an increasingly greater need given the wide range of tasks that robots can perform.

Our generation of spoken dialogue researchers should be able to develop formalised, accepted methods for evaluating spoken dialogue systems. Also, our generation should be able to analyse and improve the fine-grained aspects of human-computer dialogue to improve everyday interactions with public users. I think that grounding will become an even more ever-present concept that dialogue researchers must rely on when designing effective spoken dialogue systems. Attaining these goals requires that formal user studies be performed with publicly accessible spoken dialogue systems. In addition, these goals will require our generation of dialogue researchers to educate incoming students in our departments about the importance of spoken dialogue applications and the need for more researchers in the field.

3 Suggestions for Discussion

- Cognitive plausibility: What is the cognitive plausibility of existing spoken dialogue systems? Do they manage dialogue in a way that the cognitive science community would find acceptable?

- Environment interaction: How can we have spoken dialogue systems better process changing information about the environments of their applications?

- Appealing to the next generation: What kind of applications should we discuss with new members of our departments to excite them about dialogue research? Should we be discussing new "killer apps"?

References

S. Balakirsky, C. Scrapper, S. Carpin, and M. Lewis. 2006. Usarsim: providing a framework for multirobot performance evaluation. In *Performance Metrics for Intelligent Systems Workshop (PerMIS)*, pages 98–102.

Dan Bohus, Antoine Raux, Thomas K. Harris, Maxine Eskenazi, and Alex Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *HLT-NAACL 2007 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*.

B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. 2007. Using vision, acoustics, and natural language for disambiguation. In *2nd ACM/IEEE Conference on Human-Robot Interaction*, pages 73–80.

M. Marge, A. Pappu, B. Frisch, T. K. Harris, and A. Rudnicky. 2009. Exploring spoken dialog interaction in human-robot teams. In *Robots, Games, and Research: Success stories in USARSim Workshop*.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: The rate-a-course system. *Spoken Language Technology Workshop, 2006. IEEE*, pages 210–213.

Cristen Torrey, Aaron Powers, Matthew Marge, Susan Fussell, and Sara Kiesler. 2006. Effects of adaptive robot dialogue on information exchange and social relations. In *1st ACM SIGCHI/SIGART Human-Robot Interaction Conference*, pages 126–133.

Biographical Sketch

Matthew Marge is a PhD student in the Language Technologies Institute at Carnegie Mellon University. He received the MSc degree in Artificial Intelligence specialising in Natural Language Engineering from the University of Edinburgh in 2007. Prior to graduate school, he received the BSc degree in Computer Science and Applied Mathematics from Stony Brook University in 2006. He is currently funded by the Boeing Treasure Hunt Project and a National Science Foundation Graduate Research Fellowship. Matthew is a former recipient of a St. Andrew's Society of the State of New York Scholarship and an Edinburgh-Stanford Link studentship. Some of his hobbies include squash, bowling, and travelling.

Lluís Mas Manchón

Universitat Autonoma de Barcelona
Edifici I Facultat de CC de la Comunicació
Bellaterra 08193
Barcelona, Spain

Lluis.mas.manchon@gmail.com
laicom.uab.cat

1 Research Interests

My research interests are not in the field of engineering itself, but how to optimise the use of engineering from a **communication-centric perspective**. My goal is thus to become part of an interdisciplinary team in which I can give important insights about how agents in the communication process use the variables of the message to successful communication. Specifically, my interest is focused in the **prosody of TV News Broadcast** (Mas, 2008), which could, through the use of **keyword spotting**, segment them in pieces of news. However, my interest is also more general and **theoretical**: "taking the communication process as whole, the objectives of the emitter are always represented in some variables of the message, in which the attention of the receiver also focuses; this can be very valuable for engineers to exclusively deal with those variables" (Rodriguez Bravo, 1989, 2004).

This could make the processing of every message much more accurate and less resource-wasting. Even the research in human-computer communication would benefit of an approach that prioritises the functions of variables in the process of communication: phoneme, word or phrase are to be considered as what functions they serve in the discourse. My interest would be to become part of interdisciplinary teams working in a wide range of topics: **emotion synthesis**, word spotting, **segmentation** and **artificial intelligence**.

1.1 Current Research

My thesis (to be presented in April) focuses on searching parameters of prosody belonging to the end and beginning of pieces of news, so they can be used in an algorithm for automatic news segmentation. I did an internship in the Department of Applied Mathematics (in the Universidad Politécnica de Cartagena, Spain) to where my task was to design a processing algorithm to represent and analyse those parameters. For this pur-pose, Labview was used.

The time domain signal was broken into 512 data points with no overlapping between them, as the periodic spotting of F0 was the result of averaging approximate fragments. Then, the amplitude and phase spectrum (VRMS in Labview) was applied twice to get the cepstrum. The process was taking volts and seconds, and getting the Root Mean Square amplitude in the frequency domain, so we could visualise the power spectra in each moment. The value of F0 was finally isolated by multiplying by 0 all the range of power between 25 and 200 points in the spectra, as the F0 we were targeting (correctly articulated vowels) was expected at a maximum level in the little range left. This was proved iteratively with different ranges of signals for the VRMS and for the maximum power spectrum search.

Meanwhile, intensity was easily obtained by integrating the square power spectra, and both the intensity and F0 values were placed along a time axis. From here on, the search of variables was a matter of using different Labview tools: levels of F0 and energy in the different phases of the "locution".

We had three models of inspiration to design the acoustical system:

1. Intonational Model of Speech (Cantero, 2002): considers every data of pitch in Hz as a variation in percentages of the prior one, always starting in the level 100.

2. MOMEL (Espesser, 1993): depicts parabolas out of the pitch data, searching for the turning points: maxima and minima.

3. Autosegmental Model (Pierrehumbert, 1980): transcribes the tone movements in a symbolic language in which phonological features are related to phonetic ones.

Then, an algorithm for spotting pauses was designed to make use of the intensity: if the intensity fell to some quantity during a half second, the system would use the prosodic algorithm that examines if the intonation belongs to a typical end and beginning of pieces of news. Despite some damaging factors like reporters, noise, signature tune and other news formats, results shows a 60% of success in spotting the cuts between pieces of news. This has been achieved independently of any other factor, such as channels of the program, season, gender of presenters and even language (the corpus is in Catalan, Spanish and Portuguese). Furthermore, when a set of 1000 keywords (taken from a project on automatic news topic classification) was utilised, the success increased to 80% in spotting news cut. This implies that we have successfully found the typical intonation of ending and beginning of news locution, because the most part of errors are about using this intonation for emphasising purpose in long and tedious pieces of news.

2 Future of Spoken Dialogue Research

According to the Penta Model (Xu, 2004), every spoken signal should be analysed as the result of the functions to which it serves. As stated before, articulation possibilities, objectives of emitter and effects over the receiver should all be related to the acoustical variations of the signal, so engineers can design accurate algorithms to exclusively process the part of the signal that is relevant to some specific communication purposes.

So, in 5 to 10 years time, interdisciplinary teams could accomplish a big improvement in applied spoken dialogue algorithms if they work together with the research knowledge already at hand in disciplines as well as in others, in particular linguistics.

Taking the example of prosody, the knowledge of the different fields on how every language modulates, could be used for automatic language learning systems. Another level would imply the use of emotional intonation (already quite developed), plus the influence of prosody over the context and goal of the discourse. So in the end, different levels of prosody variables could be put together in an algorithm for intelligent recognition and synthesis systems; as, for example, they could process irony and sarcasm in a language and under certain conditions.

3 Suggestions for Discussion

We propose the following topics for discussion:

- Paralinguistic phenomena in the phases of discourse: prosody on beginnings, developments and endings.
- Paralinguistic phenomena on discourse finality.
- Discourse analysis for human-computer systems (Swerts and Ostendorf, 1995; Taboada and Mann, 2006).
- Systemic and functional Communication Model.

References

Francisco José Cantero. 2002. Teoría y análisis de la entonación. Barcelona: Ed. Publicacions i edicions UB.

Robert Espesser and Daniel Hirst. 1993. Automatic Modelling of Fundamental Frequency using a Quadratic Spline Function. Aix en Provence: Ed. IPA, Travaux de l'Institut de Phonétique d'Aix, 15.

Lluís Mas Manchón. 2008. Estructura de la Prosodia de la noticia. In *Actas y Memoria Final. Congreso Internacional Fundacional AEIC. Santiago de Compostela*. Asociación Española de Investigación de la Comunicación.

Pierrehumbert, Janet B.. 1980. The phonology and phonetics of English intonation. PhD Thesis, MIT, Cambridge, MA, USA.

Àngel Rodríguez Bravo. 2004. La investigación aplicada: una nueva perspectiva para los estudios en comunicación. In *Anàlisi Quaderns de Comunicació i Cultura* 30, pp.17–36. Barcelona.

Marc Swerts and M. Ostendorf. 1995. Discourse prosody in human machine interactions. In *Procs. of ESCA Workshop*, Vigso, Denmark.

Maite Taboada and William C. Mann. 2006. Applications of Rethorical Structure Theory. In *Discourse Studies*, 8:4, pp.567–588. DOI: 10.1177/1461445606064836.

Yi Xu. 2004. The Penta Model of Speech Melody: transmitting multiple communicative functions in parallel. Sound to Sence, June 11-13 at MIT.

Biographical Sketch

Lluís Mas is graduated in Advertising and Public Relation in the University of Alicante (Spain) and is doing his PHD in the Universidad Autónoma de Barcelona. He is part of Laboratorio de Análisis Instrumental de la Comunicación, which deals with tools of objective analysis of image, text and sound, for experimental tests and/or content analysis, to create new knowledge with immediate applications in different fields.

Some of the projects in which he has worked are about synthesis of emotions, keywords spotting for news classification, quality of TV programs, and structure analysis of news discourses and prosody of discourse.

Toyomi Meguro

NTT Communication Science Laboratories
NTT Corporation
2-4 Hikaridai, Seika-cho
Soraku-gun, Kyoto 619-0237, Japan

meguro@cslab.kecl.ntt.co.jp
www.kecl.ntt.co.jp/icl/lirg/members/meguro/index.html

1 Research Interests

My interests lie in building **listening agents**. Here, listening agents mean systems that can attentively listen to the user and satisfy his/her desire to speak and have himself/herself heard and understood. Such agents would lead the user's state of mind for the better as in a therapy session or in senior peer counselling, although I want the listening agents to help users mentally in everyday conversation. I think everyone wants to talk to someone and this desire can be partly satisfied by computational means.

There has been little research on listening agents. One exception is (Maatman et al., 2005), which showed that systems can make the user have the sense of being heard by using gestures, such as nodding and shaking of the head. Although my work is similar to theirs, the difference is that I focus more on verbal communication instead of non-verbal one.

As a first step for the study, I analysed the characteristics of listening-oriented dialogues by comparing them with casual conversation. Here, casual conversation means a dialogue where conversational participants have no predefined roles (i.e., listeners and speakers). The analysis using Hidden Markov Models (HMMs) found that it is important for listeners to self-disclose before asking questions and that it is necessary to utter more questions and acknowledgment than in casual conversation to be good listeners (Meguro et al., 2009).

I would like to further analyse the characteristics of listening-oriented dialogues using a larger amount of data. Then, I would like to incorporate the results of the analysis to build a workable listening agent. I am currently concerned with labelling schemes of dialogue acts for listening-oriented dialogues, how agents should introduce appropriate topics, what statistical models to adopt other than HMMs, and the evaluation of listening agents.

I am also interested in recommendation systems using dialogue. I think dialogue has the potential to act on one's mental attitude towards things (Fogg, 2002). I would like to create a system that finds and recommends users' best suited articles through dialogue.

2 Future of Spoken Dialogue Research

Dialogue offers an intuitive and natural way of communication between users and systems. I would like this feature to be exploited more in the next 5 to 10 years. Currently, there are many sensors and mobile devices around us to make our communication easier. This trend would continue for many years to come. In the future, dialogue technology would come in between such devices and humans to activate human communication so that the quality of life (QoL) of the human being can be improved.

For example, under time and physical constraints, it is sometimes difficult for one to communicate with the person that he/she wants to talk to. If such a situation continues for a long time, there would be no shared topics between them, and they might lose their will/desire to communicate. A dialogue system that resides in one's side, listens to one's episodes and feelings, and delivers them to other people for sharing would contribute to enhancing human communication. Twitter may be useful for sharing one's feelings with others but the messages are generally directed at anonymous people on the web. I would like dialogue technology be used to strengthen the personal bond between certain people.

3 Suggestions for Discussion

- Integration of verbal and non-verbal communication: Currently, there is little work that focuses on the integration of modalities. What kind of architecture is appropriate for modality fusion? Can we learn anything from human science?

- Social aspects of dialogue: Humans perform dialogue for social reasons. Can we make dialogue systems recognise their social situation? What kind of information is necessary to make them produce appropriate social behavior?

- Personality of dialogue systems: Although there is some emerging work on assigning a personality to dialogue systems (Mairesse and Walker, 2008), current dialogue systems generally show little personality. Is a personality really necessary for a dialogue system to be our conversational partner? How can we build a model of personality for dialogue systems?

References

B. J. Fogg. 2002. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann.

R. M. Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA)*, pages 25–36.

François Mairesse and Marilyn Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 165–173.

Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, Yasuhiro Minami, and Hideki Isozaki. 2009. Analysis of listening-oriented dialogue for building listening agents. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL)*. (to appear).

Biographical Sketch

Toyomi Meguro is a researcher 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 and recommendation systems. She enjoys travelling, playing tennis, cooking, and finding good restaurants.

Gregory Mills

Interaction, Media and Communication Group
Department of Computer Science
Queen Mary, University of London.

gj@dcs.qmul.ac.uk

1 Research Interests

My research is primarily concerned with the empirical investigation of **human-human communication**, focusing on the interactive mechanisms deployed by interlocutors in **semantic co-ordination**. I am particularly interested in the mechanisms used for dealing with problematic understanding, e.g. **repair, clarification requests, corrections**, and their role in **language change** that occurs during the development of co-ordination. I am currently focusing on how **alignment** is exploited by interlocutors and also on the boundary between communication/**miscommunication**.

1.1 Past and Current Research

Perhaps one of the least contentious statements concerning language is that it is intrinsically underdetermined, dynamic, and adaptable to novel dialogue contexts. Despite this insight, research on dialogue systems has primarily focused on the information-exchange aspects of language use, pre-supposing that both interlocutor and dialogue manager already "know how to talk" about the particular domain and have already coordinated their linguistic resources to suit the communicative situation. Consequently, dialogue system implementations have traditionally relegated the importance of this co-ordination of linguistic resources by selective incorporation of highly domain-specific vocabularies and ontologies (Larsson, 2006).

The point of departure of my research is that this static treatment of language presupposes semantic transparency between interlocutors and is inadequate for describing how co-ordination is achieved. The inherently dynamic and adaptable nature of language, and the fact that interlocutors necessarily have different interaction histories suggests the importance of an account that explains how interlocutors manage to resolve these differences and converge on a semantic model during dialogue.

Although existing models of dialogue agree that this is achieved within interaction, they disagree on which particular interactive mechanisms are implicated in this process. The collaborative model of Clark et al (Clark, 1996) characterises this as occurring through iterative cycles of "positive evidence of understanding", while the interactive alignment model (Garrod et al, 1994, 2004) prioritises the role of priming in the development of co-ordination.

To address these differences and investigate in detail the negotiation of semantic co-ordination, my research involved a series of experiments using a novel text-based chat tool. In contrast to existing experimental approaches which involve relatively coarse modification of the communicative context, e.g. confederates or wizard-of-oz scenarios, the chat tool allows fine-grained manipulation of the unfolding dialogue by selectively interfering with individual co-ordination mechanisms (Healey and Mills, 2007).

Key findings from the experiments carried out with this chat tool are: interfering with sequential coherence of the dialogue leads participants to align more (Mills, 2007); interlocutors index the level of semantic co-ordination of partners with different levels of participation (Healey and Mills, 2006); and on encountering problematic utterances interlocutors resort to less specific and "vaguer" descriptions (Mills and Healey, 2006).

The findings from these experiments present co-ordination phenomena that are difficult to reconcile with both models, due to their "semantic neutrality" (Healey and Mills, 2006). In particular, the findings demonstrate how interlocutors are sensitive to semantic differences between different kinds of referring expression, and exploit these differences in order to develop and sustain mutual-intelligibility. Importantly, this is achieved tacitly by interlocutors, in particular when dealing with problematic understanding (Mills, 2007).

These findings also work against semantic transparency underwriting co-ordination. Interlocutors frequently introduce and use terms without having full understanding of their applicability to the dialogue situation. Instead, terms are introduced opportunistically, their meaning fleshed out through iterative cycles of repair (Healey and Mills, 2006).

My research has yielded rich patterns of co-ordination in clarification subdialogues that exhibit semantic change which are not strictly reducible to the exchange of propositionally encoded information, yet still have the effect of resolving problematic understanding. Importantly, they provide compelling evidence supporting the thesis that alignment is not simply an outcome of successful interaction, but is also exploited by interlocutors when dealing with problematic understanding (Mills, 2007; Mills and Healey 2008).

1.2 Current and future work

I am currently working on refining the experimental chat tool methodology, developing it as a general experimental platform for dialogue researchers, in order to facilitate the experimental investigation of dialogue phenomena[1].

I am also conducting experiments using this chat tool to investigate how interlocutors exploit alignment in dialogue, in particular on the role of figure and ground. I am also using the chat tool to investigate how these findings scale up to multilogue.

2 Future of Spoken Dialogue Research

It is essential that the development of more naturalistic dialogue systems be guided by "what interlocutors actually do". All too frequently, the natural spontaneity, and flexibility of language and the interactive mechanisms used to deal with problems of mutual-intelligibility that arise when language is adapted to novel contexts of use is treated as inferior to the idealisation of well-formed speech acts. This of prime importance, as experimental evidence demonstrates that repair facilitates comprehension (Brennan and Schober, 2001).

Further, this flexibility introduces the notion of semantic change occurring during the course of individual conversations, bringing the problem of semantic opacity to the foreground. Dialogue research could benefit from a critical assessment of the insights from the philosophy of language concerning intentional and also semantic transparency, in particular Wittgenstein's consideration of language as practice and his arguments concerning the limitations of a strictly informational view of language. This would address the issue that using a term necessarily introduces change into its meaning for an agent and the language community as a whole. Creating a typology of these changes, and how they are achieved through coordination mechanisms would decrease the gulf that exists between dialogues with dialogue managers and actual human-human conversation.

3 Suggestions for Discussion

- Methodologies used to determine which particular mechanisms to include in dialogue systems: empirical approaches, e.g. wizard-of-oz, confederates, introduce different biases from user-simulations. Discussion of the relative merits and limitations of both would be of great benefit to dialogue system designers.
- Semantic and intentional transparency: what steps can be taken to make models of agency and language production/comprehension more similar to actual human-human conversation.
- Language change: how can dialogue managers be designed to allow for novel uses of terms by both interlocutors and dialogue system?

[1] DiET: Dialogue Experimentation Toolkit. http://www.dcs.qmul.ac.uk/research/imc/diet/

References

Brennan, S.E., and Schober M. F. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44: 274–296.

Clark, H. H. *Using Language*. Cambridge University Press, Cambridge.

Garrod, S. and Doherty, G. 1994. Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53: 181–215.

Healey, P.G.T. and Mills, G.J. 2006. Participation, Precedence and Co-ordination in Dialogue. In *Proceedings of the 28th Conference of the Cognitive Science Society*: 1470–1475, Vancouver, Canada.

Larsson, S. 2006. Semantic plasticity. Paper presented at *LCM 2006. (Language, Culture and Mind)*, July 2006, Paris, France.

Mills, G. J. 2007. *The development of semantic coordination in dialogue: the role of direct interaction.* Unpublished PhD Thesis.

Mills, G.J. and Healey, P.G.T. 2006. Clarifying Spatial Descriptions: Local and Global Effects on Semantic Co-ordination. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*: 122–129 University of Potsdam, Germany.

Biographical Sketch

The author completed his PhD in 2007 and is currently working as a Postdoc at Queen Mary University, focusing on the interactive mechanisms involved in semantic co-ordination. He is interested in the philosophy of language and mind, phenomenology of (mis)communication, the evolution of language, and also language use in therapeutic settings. Outside of academia, he flies gliders and plays electronic music and the harmonica.

Teruhisa Misu

National Institute of Information and
Communications Technology (NICT)
2-2-2 Hikaridai, Keihanna Science City,
Kyoto, Japan

teruhisa.misu@nict.go.jp
mastarpj.nict.go.jp/~xtmisu/

1 Research Interests

My current research interest is in making **proactive dialogue systems**. Although most current spoken dialogue systems (SDSs) handle simple database retrieval or transactions driven by users' explicit requests, it is desirable that SDSs should make proactive interactions such as clarifications, recommendations and advice like a hotel concierge or an experienced operator. More specifically, I am interested in **dialogue management in non-database retrieval tasks** and **adaptive response generation**.

1.1 Previous work

My previous work has focused on SDSs based on **information retrieval** and **question-answering** (QA) using **a set of documents as a knowledge base** (Misu, 2008).

1.1.1 Confirmation and Clarification in Document Retrieval Tasks

It is indispensable for SDSs to interpret user's intention robustly in the presence of speech recognition errors and extraneous expressions characteristic of spontaneous speech. In speech input, moreover, users' queries tend to be vague, and they may need to be clarified through dialogue in order to extract sufficient information to get meaningful retrieval results. In conventional database query tasks, it is easy to cope with these problems by extracting and confirming keywords based on semantic slots. However, it is not straightforward to apply such a methodology to general document retrieval tasks.

To solve these problems, we proposed a confirmation method based on two statistical measures that are not based on the confidence measure of ASR, but on the impact on retrieval as well as the degree of matching with the backend knowledge base (Misu and Kawahara, 2006b; Misu and Kawahara, 2008). We also proposed a method to make clarification questions by dynamically selecting from a pool of possible candidate questions. As the criterion for the selection the information gain is defined based on the reduction in the number of matched items (Misu and Kawahara, 2005; Misu and Kawahara, 2006b).

1.1.2 Interactive Navigation based on QA and Information Recommendation

We proposed an interactive dialogue framework. In conventional audio guidance systems, such as those deployed in museums, the information flow is one-way and the content is fixed. We prepare two modes, a user-initiative retrieval/QA mode (pull-mode) and a system-initiative recommendation mode (push-mode), and switch between them according to the user's state. In the user-initiative retrieval/QA mode, the user can ask questions about specific facts in the documents in addition to general queries. In the system-initiative recommendation mode, the system actively provides the information the user would be interested in. The system utterances are generated by retrieving from and summarising Wikipedia documents. We implemented a navigation system containing Kyoto city information. The effectiveness of the proposed techniques was confirmed through a field trial by a number of real novice users (Misu and Kawahara, 2007b; Misu and Kawahara, 2007a).

1.1.3 Efficient Language Model Construction for SDSs

We proposed a bootstrapping method of constructing statistical language models for new SDSs by collecting and selecting sentences from the World Wide Web. Out of the collected texts, we select "matched" sentences both in terms of the domain and in utterance style, thus appropriate for training data of the language model (Misu and Kawahara, 2006a).

1.2 Current and Future Work

Currently, we are developing consulting dialogue systems that help make decisions through spontaneous interactions.

Most previous studies assumed a definite and consistent user goal. The dialogue strategies were usually designed to minimise the cost of information access. However, this assumption fails in various real world situations. Specifically, we are developing a dialogue system that handles tourist guidance. Thus far, we have collected itinerary planning dialogues in Japanese, in which users plan a one-day visit to Kyoto City (Misu et al., 2009).

It contains various exchanges, such as clarifications and reasonings. The user may explain his/her vague preferences by listing examples. The server would sense the users preference from his/her utterances and then request a decision.

In order to construct a consulting dialogue system from the corpus, we are annotating dialogue acts (Misu et al., 2009). Moreover, we proposed an efficient dialogue management scheme based on WFSTs using N-gram DA tag sequences (Hori et al., 2009). Of course, some sort of dialogue state modelling (e.g. (Thomson et al., 2008)) would be needed, and re-scoring of system actions based on the state is also our future work.

2 Future of Spoken Dialogue Research

I am afraid that the more multifunctional cell phones become (e.g. iPhone, Android), the less users use speech interfaces for simple query tasks such as train search, hotel reservation, etc. Thus we will have to propose applications that can make use of something that such smartphones do not have. It may be a large (human-sized) touchscreen or a motion detection sensor. Another direction would be expansion of tasks that a speech interface can handle (but a small touch screen cannot).

3 Suggestion for discussion

- **How to evaluate dialogue systems**
  What is a good evaluation measure of dialogue systems? Can we define a evaluation measure like "BLEU/NIST score" for machine translation. (especially in non-goal-oriented dialogue systems)

- **How to make users behave naturally as a human operator can.**
  User's talk to the systems in different utterance style from the style they talk to human operators. What are the causes of this phenomenon?

- **Cost reduction in developing new SDSs using the WWW**
  How can we make use of web resources for construction of SDSs.

References

C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura. 2009. Recent Advances in WFST-based Dialog System. In *Proc. Interspeech*, page accepted for presentation.

T. Misu and T. Kawahara. 2005. Speech-based information retrieval system with clarification dialogue strategy. In *Proc. Human Language Technology Conf. (HLT/EMNLP)*.

T. Misu and T. Kawahara. 2006a. A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proc. Interspeech*, pages 9–12.

T. Misu and T. Kawahara. 2006b. Dialogue Strategy to Clarify User's Queries for Document Retrieval System with Speech Interface. *Speech Communication*, 48(9):1137–1150.

T. Misu and T. Kawahara. 2007a. An interactive framework for document retrieval and presentation with question-answering function in restricted domain. In *Proc. Int'l Conf. Industrial, Engineering & Other Applications of Artificial Intelligent Systems (IEA/AIE)*, pages 126–134.

T. Misu and T. Kawahara. 2007b. Speech-based Interactive Information Guidance System using Question-Answering Technique. In *Proc. ICASSP*.

T. Misu and T. Kawahara. 2008. Bayes risk-based dialogue management for document retrieval system with speech interface. In *Proc. COLING, Vol.Posters & Demo*, pages 59–62.

T. Misu, K. Ohtake, C. Hori, H. Kashioka, and S. Nakamura. 2009. Annotating Communicative Function and Semantic Content in Dialogue Act for Construction of Consulting Dialogue Systems. In *Proc. Interspeech*, page accepted for presentation.

T. Misu. 2008. *Speech-based Navigation Systems based on Information Retrieval and Question-Answering with Optimal Dialogue Strategies*. Ph.D. thesis, School of Informatics, Kyoto University.

B. Thomson, J. Schatzmann, and S. Young. 2008. Bayesian Update of Dialogue State for Robust Dialogue Systems. In *Proc. ICASSP*, pages 4937–4940.

Biographical Sketch

Teruhisa Misu is a researcher in the Spoken Language Communication Group, MASTAR Project at National Institute of Information and Communications Technology (NICT), Japan. He 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.

Aasish Pappu

Carnegie Mellon University
Language Technologies Institute
School of Computer Science
Pittsburgh, PA 15213

aasish@cs.cmu.edu
www.cs.cmu.edu/~apappu

1 Research Interests

My research interests span task oriented dialogue systems to spoken dialogue interfaces for knowledge acquisition. Currently, I am interested in investigating ontology driven dialogue task management. I am study-ing the usefulness of ontology for preparing a common ground of knowledge for humans and robots while they perform collaborative work.

1.1 Past Research

During my undergrad research I worked on a natural language interface for programming in Java. We developed a system that could respond to simple English sentences with primitive Java constructs. Sentences such as "*Select third positive element from array*", "*Add a to b*", "*Select every element which is greater than two from array*", etc.

Primarily, the motivation behind this project is to create an environment for people who are planning to learn new programming languages with little programming background. Since, the syntax is taken care of one should know little about the new language. Al-so, system actively learns about the typecasting of the variables from the initialised value. Like, "initialise g with 9.8", would inform the system about typecasting scope for the variable g. Through conversation, user was able to ask the system about existing variables and system could alert the user while rewriting existing variables. Altogether, we could setup a very basic programming environment through dialogue.

This work is inspired from M'PAL (Rosenbloom, 1985) a programming language project based on definitions written in English and another interesting work that took case-based reasoning approach for natural language interface to Unix shell scripting (Won Il Lee and Geunbae Lee, 1995)

1.2 Current Research

Currently, I am associated with TeamTalk project group under the supervision of Dr. Alex Rudnicky. This project investigates different aspects of human-robot teams participating in collaborative task. My present work focuses on designing ontology to drive robot navigation in an environment. Ontology driven planning and learning has been popular with systems related to personal assistants (Niekrasz et al., 2004), that keep track of user's actions and learn about surroundings to keep the user informed about alerts in the environment. Also, goal oriented systems urban search and rescue systems (Schlenoff and Messina, 2005) employed ontology based approach for robot navigation.

Primary purpose of having ontology is to generate spatial representations that allow a human robot team to refer to objects using common sense in an environment. It is necessary to have a symbolic representation of the objects and relationships with other objects. Conceptualisation of these objects avails the opportunity for the robot to infer complex queries such as "*Face the door and walk until you find a window*".

Furthermore, we could assign attributes to each object keep track of their status and update the sensory map i.e., a robot's view of the world perceived through its sensory inputs. Additionally, sub-teams working on a common plan could share their '*view*' of the world with their teammates by updating the common knowledge. Furthermore, queries like "*which robot walked through the second door*". This kind of queries de-mand good understanding of answer-types conditioned with a context. These answer types are nothing concepts in the ontology. Additionally, events that take place in the environment and actions performed by robots are tracked and stored in the ontology, thereby helping a robot to answer queries like, "*when was the last time we talked about the blue ball*".

1.3 Queuing models for dialogue management

Dialogue systems that are severely exposed to diverse users exhibit interesting challenges pertaining to behavioral changes due to changing environment. Not only, dialogue systems but also traditional service provider systems like network traffic management and disaster evacuation systems (Gross, 2008) have adopted wide range of simulation models that approximate real environment.

We could observe analogies between queuing models and dialogue systems. Every user utterance could be treated as a Poisson arrival, whereas decoding and prompting to the user as a service. In M/M/1 queuing model where the Kendall's notation says that both the arrival rate and service rate are Poisson processes with one service provider. This model is very much similar to single server dialogue system. On the other hand, we could have M/M/k multiple servers with single queue system that are highly efficient and perform better than each having their own queue. It is intuitive to under-stand a single large set of servers perform better than two or more smaller sets.

Similarly, we could have several dialogue systems working together, each maintaining a dialogue state chosen from n-best possible states for a given user utterance. This kind of setup may help in multi-participant conversation, where the conversation towards an addressee could be shared with another addressee or re-main independent from other addressees. Therefore, more than one participant may apparently share the state, each one with a different dialogue manager.

2 Future of Spoken Dialogue Research

In my opinion spoken dialogue research will have to take assistance from other modalities to infer from tractable user actions and anticipate what user might say. Additionally, speech recognition mistakes should be complemented by inputs from other modalities, thereby how could we learning from word errors.

2.1 Learn from humans

It will be interesting to investigate effectiveness of dialogue systems in active learning and knowledge acquisition tasks through which humans can share their knowledge with machines.

2.2 Spoken dialogue forum

An interesting application of spoken dialogue re-search could be in the area of product review systems. A user may voice out his opinion about a product and the system could voice out similar opinions put forth by other people. Additionally, the purpose of this system is to accelerate the process of data collection for speech processing research.

3 Suggestions for Discussion

- Investigating effective (and inexpensive) ways of collecting lots of "real" data and simulating real conversations.

- How to identify domain independent sub-conversations from a collection of data and adapt models trained on them to other domains?

References

Michael H. Rosenbloom. 1985. M'PAL: a programming language based on natural English definitions. *ACM SIGPLAN Notices*, Volume 20, New York, NY.

Won Il Lee and Geunbae Lee. 1995. From natural language to shell script: a case-based reasoning system for automatic UNIX programming. In *Expert Systems with Applications*, 9:1, 1995, pp.71–79.

J. Niekrasz, M. Purver, J. Dowding, and S. Peters. 2005. Ontology-based discourse understanding for a persistent meeting assistant. In *Proceedings of the AAAI Spring Symposium.*

Craig Schlenoff and Elena Messina. 2005. A robot ontology for urban search and rescue. In *Proc. of the 2005 ACM workshop on KRAS*. ACM, New York, NY.

D. Gross. 2008. Fundamentals of Queuing Theory. Wiley-India.

Biographical Sketch

Aasish Pappu is currently a 2^{nd} year Master's student and Research Assistant at the Language Technologies Institute, CMU at Pittsburgh under the supervision of Dr. Alex Rudnicky. He obtained his BTech degree in Information Technology from Indian Institute of Information Technology, Allahabad, India. Besides research, his interests include photography, philosophy, languages and poetry.

Joana Paulo Pardal

Dept. Computer Science and Engineering
IST, Technical University of Lisbon
Spoken Language Systems Laboratory (L^{2}F)
Lisboa, Portugal

joana.paulo.pardal@l2f.inesc-id.pt
www.l2f.inesc-id.pt/~joana

1 Research Interests

My research interests lie generally in the area of **spoken dialogue systems** with particular interest in **software engineering** techniques to **dynamically integrate structured knowledge sources**, like databases and **ontologies** (Paulo Pardal, 2007), and in **evaluation frameworks** to measure the improvements. The challenges of creating **coaching, tutorial, and educational systems** that can be used by the **general public** at their homes, schools, or museums are also part of my research.

1.1 Past research

My MSc thesis was on "automatic terms acquisition" (Paulo et al., 2004). I've taught object-oriented programming and design patterns, knowledge representation, artificial intelligence (AI), autonomous agents and multi-agent systems (Melo et al., 2006), and distributed systems (Pardal et al., 2008). When starting my PhD I've moved to spoken dialogue systems. Being a CS engineer from the field of AI I was interested on how to use ontologies to ease the extension of a system to new tasks, similarly to what is done with databases. Databases' structure was used to extract domain knowledge and it allows the generic use of that kind of information reducing the coding time and adaptations needed to build new dialogue systems. With this task in mind, I worked on OntoChef, a cooking ontology (Ribeiro et al., 2006). The ontology was later enriched with the use of a natural language specific tool (Machado, 2007) through information extraction techniques. A collection of nearly 9000 recipes where extracted from Portuguese websites with a specially designed tool. They are to be converted into ontology-based format soon. To better understand how humans coach each others while cooking, a human-human cooking corpus was collected, where a person helped another one while s/he cooks a recipe (currently there are approximately 3 hours with 6 different participants, in 3 teams doing a 'chocolate mousse'). This corpus is to be annotated with the requests made by the person executing the task and the relative answers, tips, and comments from the coach. This follows my previous work at Rochester (Gomez-Gallo et al., 2007).

1.2 Current research

Most practical dialogue systems are designed for a specific task, and even if the authors were concerned with possible future extensions, integrating new tasks is always a challenge. Dynamic integration of new tasks according to some kind of structured knowledge is an interesting research topic. The main goal of my thesis (Paulo Pardal, 2007) is to study how different levels of knowledge stored in ontologies can be used to facilitate the creation of new coaching dialogue systems capable of domain reasoning. I'm taking McGuinness' ontologies spectrum (McGuinness, 2003) to split the ontology into increasingly complex knowledge levels.

The hypothesis being studied is whether ontologies can be used to enrich a coaching spoken dialogue system and be used in it in such way that the system can abstract the source of domain-specific knowledge – related to the tasks being coached – focusing only on the dialogue phenomena. The integration of ontological knowledge should be done with few architecture adaptions to the dialogue system so that when adding a new domain – a new class of tasks – minor changes in special modules are sufficient.

1.2.1 Case Study: Cooking Coach

Cooking is something that everybody ends up doing. Some have been cooking for a lifetime, some are just beginning. Most of the time the user's hands are busy and dirty. In such a scenario, manually handling a recipe book is to be avoided. A system that helps the user by dictating the steps while hands and eyes are occupied with the cooking tasks is much desired. My goal is to develop a spoken dialogue system that provides assistance in reading the procedure and detailing all steps that may be unclear to a user lacking expertise. Based in our cooking experience, we built a prototype system that helps the user while cooking (Martins et al., 2008c). The system adapts itself to the users' needs and expertise. It was built over DIGA (Martins et al., 2008b; Martins et al., 2008a), a spoken dialogue systems framework. Some experiments with CMU's Olympus (Bohus et al., 2007) are currently being done.

1.3 Future research

The use of an ontology delivers another interesting result: the system could reason about the tasks and plans at hand. After saying '*Separate egg whites and egg yolks.*' to the user, the system needs to know that the existing 'eggs' will disappear and give place to 'egg whites' and 'egg yolks'.

It should also make it easier to include different languages by translating (Graça et al., 2008) the cooking ontology and adapting the necessary linguistic resources (understanding and generation).

The integration of additional sensors in the kitchen could bring interesting enhancements like in DFKI's 'SmartKitchen'(Schneider, 2007) or the MIT's Media Lab 'intelligent counter' (`http://www.media.mit.edu/ci/`)

2 Future of Spoken Dialogue Research

Currently spoken dialogue systems are proposed only when no other input modalities are available (like when there is no access to a keyboard or when the user has some kind of special need – blindness, reduced accessibility, etc.) However, we should consider the use of speech whenever it is natural. That would be easier if interaction with these systems was more natural (more similar to human-human interaction).

When human-computer interaction approaches human-human interaction, people will feel comfortable on delegating some tasks to a digital helper while they will concern themselves with some other tasks. Managing priorities and knowing the right times to interrupt are important. The "uncanny valley", however, must be avoided.

Exploring the new emerging technologies and mobile devices (like iPhone or Android) is another path that is worth considering: integrating speech with new information from sensors can further leverage the current state-of-the-art, making the available interfaces easier to use and much more natural.

3 Suggestions for Discussion

- Teaching SDS (methods, frameworks, evaluation)
- SDS for Coaching (assist with task) and Teaching (tutorial and educational applications).
- Evaluation: universal metrics for comparing disparate systems, tasks, languages and modalities; expert systems against rapid development frameworks.

References

D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. I. Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Proc. of Bridging the Gap: Academic and Industrial Research in Dialog Technology, workshop at HLT/NAACL.*

C. Gomez-Gallo, G. Aist, J. Allen, W. de Beaumont, S. Coria, W. Gegg-Harrison, J. Paulo Pardal, and M. Swift. 2007. Annotating continuous understanding in a multimodal dialogue corpus. In *SemDial – DECALOG.*

J. Graça, J. Paulo Pardal, L. Coheur, and D. Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. In *LREC*.

T. Machado. 2007. Extracção de informação – introdução automática de receitas de acordo com ontologia. Master's thesis, IST, UTL.

F. Martins, A. Mendes, J. Paulo Pardal, N. Mamede, and J. P. Neto. 2008a. Using system expectations to manage user interactions. In *PROPOR*, LNCS. Springer.

F. Martins, A. Mendes, M. Viveiros, J. Paulo Pardal, P. Arez, N. Mamede, and J. P. Neto. 2008b. Reengineering a domain-independent framework for spoken dialogue systems. In *SET&QA4NLP*. ACL.

F. Martins, J. Paulo Pardal, L. Franqueira, P. Arez, and N. Mamede. 2008c. Starting to cook a tutoring dialogue system. In *SLT*. IEEE/ACL.

D. McGuinness. 2003. Ontologies come of age. In *The Semantic Web: Why, What, and How*. MIT Press.

C. Melo, R. Prada, G. Raimundo, J. Paulo Pardal, H. S. Pinto, and A. Paiva. 2006. Mainstream games in the multi-agent classroom. In *IAT*. IEEE Computer Society.

M. Pardal, S. Fernandes, J. Martins, and J. Paulo Pardal. 2008. Customizing web services with extensions in the STEP Framework. *International Journal of Web Services Practices - IJWSP*, 3(1).

J. L. Paulo, D. Martins de Matos, and N. Mamede, 2004. *Terminology Mining with ATA and Galinha*, chapter 2. Edições Colibri, Lisbon, Portugal.

J. Paulo Pardal. 2007. Dynamic use of ontologies in dialogue systems. In *NAACL-HLT Doctoral Consortium.*

R. Ribeiro, F. Batista, J. Paulo Pardal, N. Mamede, and H. S. Pinto. 2006. Cooking an ontology. In *AIMSA*, LNCS. Springer.

M. Schneider. 2007. The semantic cookbook: sharing cooking experiences in the smartkitchen. *3rd Interntl. Conf. on Intelligent Environments.*

Biographical Sketch

Joana Paulo Pardal is a 4^{th} year Ph.D. student in Computer Science Engineering at IST, Technical University of Lisbon, under the supervision of Nuno J. Mamede (IST), H. Sofia Pinto (IST) and James F. Allen (U. Rochester). She holds a fellowship from FCT (Portuguese Nat. Science Foundation). Joana received a *licenciatura* in 2001, and an M.Sc. in 2004 both in CS Eng. from IST. Joana is a researcher at L^2F INESC-ID since 2001. She is a Lecturer at IST since 2002. She is a student member of ISCA, ACL and AAAI. and participates on CMU's *DoD* reading group. She has worked abroad thrice: in Clermont-Ferrand, France (Summer 2004); Rochester, NY, USA (Fall 2006), Cambridge, MA, USA (Fall 2008, Winter 2009). She enjoys traveling, cooking, gardening, playing Nintendo Wii, and eating at good restaurants.

David Díaz Pardo de Vera

Universidad Politécnica de Madrid
Cuidad Universitaria s/n
28040 Madrid
Spain

dpardo@gaps.ssr.upm.es

1 Research Interests

My main research interests lie in **spoken dialogue systems** (SDSs) incorporating **nonverbal channels of communication**, focussing particularly on **user-centred quality evaluation**.

1.1 Main research goal/focus

At GAPS (the Signal Processing Applications Group at UPM) our main goal is to improve the robustness of human-machine communication while making the experience for the users as efficient and enjoyable as possible. This can only be achieved by combining work in the design of multimodal communication acts, developing dialogue strategies that provide flexibility and robustness, and then learning how to properly evaluate the users' experience.

We have paid special attention to studying the effects that incorporating an embodied conversational agent (ECA) might have on the usability, efficiency and user acceptance of spoken dialogue systems.

1.2 User-centred evaluation

We may identify two main goals of user-centred evaluation: the first is to be able to establish the users' overall opinion of the "quality" of a dialogue system; the second is to develop models to predict the users' perception of quality from a set of parameters.

No standard procedure exists to evaluate dialogue systems in specific contexts, such as enrolment and verification dialogues for speaker authentication systems, which is an area we have explored at GAPS. Nor have we found well defined and tested guidelines to evaluate multimodal dialogue systems that include a humanlike figure in the visual communication channel (an ECA), or standardised approaches to take user emotions into account. We have followed the ITU P.851 recommendation (ITU-T P.851, 1999) on questionnaire design for the evaluation of spoken dialogue systems for general telephone services, and expanded it to include dimensions to evaluate user perceptions related with secure access and with the ECA. Inspired by Möller's taxonomy of quality factors (Möller et al., 2007), we combine user questionnaire responses with interaction data registered automatically.

We have defined a quality-parameter class structure, or frame, that provides conceptual clarity both to develop questionnaires and future experiments and to analyse the data obtained in them. The class structure arises from considering two orthogonal sets of dimensions. First, from the literature on usability and ergonomics (among which Angela Sasse's work is especially relevant; see, e.g., Sasse (2004)) we have extrapolated the notion that three major classes of parameters are generally (and tacitly) considered to be related to user acceptance: likeability factors (those that have to do with the experience of using the system), rejection factors (those that can only have a negative valence) and perception of usefulness. Secondly, the class structure is broken down into various levels that may independently affect or be subject to the users' perception. We may distinguish an overall system-assessment level, task and goal-related levels, and an interaction/interface level.

1.3 Case studies

We have performed experiments to study the effects of an ECA in a SLDS. Responses to our questionnaires do not show a clear difference in scores between users with and without the ECA in their interface, regarding their subjective experience, despite differences observed in important objective parameters, such as number of turns, time-outs and barge-in attempts. ECA users were slightly more positive about the experience, but they were no more inclined to use a similar system in the future (this is commonly taken as an important indicator of user acceptance). Factor analyses on our questionnaires have allowed us to gain extra insight into the structure of user acceptance-related factors and the dynamics of the users' experience, revealing similar factor structures to those identified by other authors (especially Hone and Graham (2001)). We identified a new factor that appeared persistently in all questionnaires: rejection of the system caused by privacy and security concerns. Furthermore, privacy concerns were strongly coupled with perception of usefulness and inclination to use the system. We also observed a slight advantage in the perception of performance quality for the ECA system in the beginning, perhaps due to false expectations generated by the stronger social presence of the system with the ECA. Differences disappear in subsequent stages. Finally, the ECA seems to help users understand what is going on and what they are supposed to do (say) throughout the dialogue.

1.4 Multimodal communication act generation

We have conceived a multimodal interaction platform for the multimodal expression of communication goals. The central module of the system we are working on (at the moment it is just a concept) we call the interaction manager, which produces a communication intention base (CIB). The CIB defines the ECA's response on a preverbal level. It has three elements: the interaction control element defines turn management and theme structure; the open discourse element deals with the literal meaning the system wishes to communicate. The non-declared communication element describes an intention behind the literal meaning of the message. Note that hidden intentions may differ greatly from what is openly said!

2 Future of Spoken Dialogue Research

I believe in the near future we will see an intensification of efforts to standardise the specification of communication acts for system dialogue output. The generation process has to be studied, first of all to determine what stages it should be broken down into. We need to know how the system should react to user input in particular interaction contexts, how to specify multimodal, multi-layered semantic output, from concept to output. Efforts from the AI camp centred on cognition are already taking off (e.g., the SAIBA framework). An HMI counterpart focusing on the interaction and the user's experience is also needed. Multi-layered semantic appraisal of the user's messages and behaviour may take longer, but is nonetheless also necessary achieve sophisticated human-machine conversation.

3 Suggestions for Discussion

My suggestions for discussion are the following:

- What elements of quality should we focus on when evaluating multimodal dialogue systems?
- Methods of qualitative and quantitative analysis of the users' experience.
- Formalisation of multimodal communication act generation.

References

A. Hernández, B. López, D. Díaz, R. Fernández, L. Hernández, and J. Caminero. 2007. A person in the interface: effects on user perceptions of multibiometrics, In *Procs. of Workshop on Embodied Language Processing*, in the 45th Annual Meeting of the Association for Computational Linguistics, ACL, pp.33–40, Prague.

ITU-T P.851. 1999. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union (ITU), Geneva.

B. López Mencia, D. Díaz Pardo de Vera, A. Hernández Trapote, L. Hernández Gómez, M. C. Rodríguez and J. Relaño Gil. 2008. Evaluation of ECA gesture strategies for robust human-machine interaction. In *International Journal of Semantic Computing Special Issue on Gesture in Multimodal Systems*.

S. Möller, P. Smeele, H. Boland, and J. Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: a case study. In *Computer Speech & Language* 21, pp.26–53.

SAIBA: `wiki.mindmakers.org/projects:saiba:main/`

M. A. Sasse. 2004. Usability and trust in information systems, Cyber Trust & Crime Prevention Project. University College London.

Biographical Sketch

David Díaz Pardo de Vera has an MEng in telecommunications engineering from Universidad Politécnica de Madrid. He is currently a 2^{nd} year PhD student and holds a research position in the same university, under the supervision of Luis Hernández. David is also (still!) writing (or, rather, not...) a masters' thesis on the sociology and ethics of human-animal relations in the context of tourism in Lapland and in Spain. David (long ago) used to run competitively (though not too competently), and now wishes he hadn't wasted so much time in his youth instead of focusing on music so he could now (maybe) play jazz (or anything at all!), which he loves.

Florian Pinault

LIA - University of Avignon,
BP 1228, 84911 Avignon,
France

florian.pinault@univ-avignon.fr

1 Research Interests

My research interests about automatic language processing lie generally in the area of **dialogue management**, to which I want to apply principled mathematical models and linguistic theories.

The PhD that I am currently completing explores the field of **human-machine dialogue** systems from a stochastic approach using **reinforcement learning**. Statistical modelling relies on using **Partially Observable Markov Decision Process (POMDP)**.

1.1 Stochastic dialogue management modelling (POMDP)

Classical dialogue systems model the *dialogue state* using logical rules, grammar or planning. A major issue is that they are usually very sensitive to speech recognition error. Therefore complex mechanisms to repair, to confirm or to correct these mistakes have been designed on a meta-dialogue level. Nevertheless, triggering these mechanisms at the appropriate time has proved to be difficult. A potential solution lies in modelling the uncertainty of the system about the dialogue. The POMDP framework statistically models the *belief* of the system about the dialogue state.

In a POMDP, a Hidden Markov Process models the unobserved dialogue state s_t and observation o_t. Belief about s is $b_t(.) = P(s_t = .|\text{ all past history})$ and can be computed using Bayes rules. At each time step an action $a_t = \pi(b_t)$ is taken (decision) by the system from the belief b_t according a policy π. The state s_{t+1} depends also on a_t. An introduction can be found in (Young et al., 2007).

1.2 Parameters estimation:

Algorithms finding the optimal policy π can be model-based or model-free. In the model-based approach the POMDP parameters are fully estimated. I used a corpus annotated with hierarchical frame semantics (Meurs et al., 2009), based on the FrameNet paradigm (Baker et al., 1998). The model-free approach usually learn their parameters through user-simulation, I am currently adapting an agenda based user simulator (Schatzmann et al., 2007) in this purpose.

1.3 Computational complexity issue:

POMDP policy optimisation is known to be intractable for real-world tasks, though some algorithms finding sub-optimal policies have been proposed recently (Pineau, 2004). To reduce the complexity of the search space, the **dialogue state s and action a are summarised** into a small number of descriptors where it is assumed that summary machine actions can be decided from the information still available in the summary state. The resulting POMDP becomes more tractable due to the lower search space size.

1.4 Evaluation:

Evaluation has been performed at a summary level, reusing the same probability model. This auto-evaluation paradigm where the same model is used to fulfill the task and evaluate the performance is a major issue in dialogue management research.

These work are described more precisely in (Pinault et al., 2009).

2 Future of Spoken Dialogue Research

Present research focuses rightly to task-specific systems. I believe that dialogue involving complex reasoning, analogy, poetry or humour will stay out of reach for long.

In my view, the state of dialogue research in the next decade will show different characteristics:

- **Multi-modal based :** Future human-machine dialogues will occur through a mobile device, including diverse input/output device. A strong demand to develop efficient dialogue systems will come from this part of the web industry. (e.g. VoiceXML)
- **Rich annotated corpus availability :** As multimedia classification and semantic web annotation will progress, data collected about users (e.g. google) will generate huge corpora.
- Transcriptions provided by ASR (Automatic Speech Recognition) will keep being noisy and corrupted by

errors. The next big improvement in **speech recognition** will come from a better use of the dialogue context including semantic modelling.

3 Suggestions for Discussion

- Dialogue state modelling: Should dialogue state include more than an estimation of the user goal?
- Academic and industry applications: How to cast a POMDP in VoiceXML?
- Cooperative user adaptation: How to use natural ability of the user to adapt to the system?
- Rigid turn by turn sequence: Dialogue are always modeled as a succession of speaker turns. How to naturally include synchronous events as mixed turns, barge in, etc.?

References

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of the COLING-ACL*, 98.

M.J. Meurs, F. Lefèvre, and R. De Mori. 2009. Spoken language interpretation: On the use of dynamic bayesian networks for semantic composition. In *ICASSP*.

F. Pinault, F. Lefèvre, and R. De Mori. 2009. Feature-based summary spaces for stochastic dialogue modeling with hierarchical semantic frames. *INTERSPEECH*.

J. Pineau. 2004. *Tractable Planning Under Uncertainty: Exploiting Structure*. Ph.D. thesis.

J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. *ACL*.

S. Young, J. Schatzmann, K. Weilhammer, and H. Ye. 2007. The Hidden Information State Approach to Dialog Management. *Proc. of ICASSP*.

Biographical Sketch

Florian Pinault is currently completing a Ph. D. in Stochastic Dialogue Systems at the University of Avignon (France). He received a Master in Probability and Statistics from University of Orsay (Paris XI) and a *magistere* of Mathematics and Computer Science from the Ecole Normale Superieure of Ulm (Paris).

His main research interests lie in applications of stochastic models to language processing. His thesis work focuses on stochastic modelling of dialogue state and finding optimal dialogue policies using reinforcement learning.

Mara Reis

Federal University of Santa Catarina
Dept. of Applied Linguistics
Second Language Phonetics and Phonology
Florianópolis, SC, Brazil

marasreis@hotmail.com

1 Research Interests

My research interests lies in practical uses of spoken dialogue systems, towards studies that investigate **foreign language speech perception and production**, **pronunciation training**, and **speech intelligibility improvement**.

1.1 Past Research

Given that accurate foreign language production is believed to be closely related to accurate perception (Flege, 1995; Kuhl and Iverson, 1995; Best and Tyler, 2007), I have investigated the relationship between perception and production among Brazilian, Dutch, and French speakers of English as a foreign language. If such a relationship exists, perceptual phonetic training may be able to alter production. In order to examine this hypothesis, I have conducted perceptual training with Brazilian learners of English using the consonants /p, t, k/ that, although present in Portuguese and in English, are phonetically different in their realisation (Reis, 2007a; Reis et al., 2007). The results showed significant improvement towards a more English-like pronunciation, corroborating the claim that perception and production are related.

1.2 Current and Future Research

Bearing in mind that foreign language perception and production may be related for some speech sounds, my colleagues and I are currently working with pronunciation training and, thus, the improvement of foreign language intelligibility.

In a globalised world in which English is undeniably the contemporary lingua franca, successful cross-language communication requires a minimal level of speech intelligibility, particularly for business reasons (Celce-Murcia, 1987; Morley, 1991; Reis, 2007b).

As far as foreign-accented speech and intelligibility are concerned, some studies indicate, some studies indicate that accented-speech does not necessarily favor intelligibility by non-native speakers (e.g., Mackay et al., 2006; Munro et al., 2006), whereas other studies show that foreign language users benefit from speech produced with their own accent (e.g., Smith and Bisazza, 1982; Gass and Varonis, 1984).

We have recently investigated whether accented-speech would be differently evaluated by Brazilian and Dutch speakers of English (Reis and Kluge, 2008; Kluge et al., 2008). Due to different phonological representations of the consonants /m/ and /n/ in word-final position in their own language, Brazilian Portuguese and Dutch speakers tend to have different pronunciation of English words in this context — Dutch present an English-like production, whereas Brazilians' output is remarkably accented. When asked to identify and rate samples of Brazilian-accented English, both groups showed that accurate pronunciation led to more word recognition. However, Dutch were consistently more aware of the accented-speech than Brazilians. These somehow contradictory results were interpreted as indication that further studies are needed in order to understand to what extent accented-speech is more or less intelligible for foreign language speakers.

Bearing in mind such need, our current work aims at analysing whether speakers of Brazilian Portuguese and French recognise and judge their own pronunciation of English words that contain TH spelling, as in 'theater', more intelligible than native and the other non-native production.

With regard to the application and impact of the results of this study on research on spoken language systems, if the findings show that foreign speakers tend to evaluate their own accent more intelligible than native or other non-native pronunciations, adaptations of the systems to certain cultural or political contexts is a possibility to be considered. Multicultural environments in which foreign accent is maintained as an identity and cultural bond could require an adjustment of the systems according to these environments' needs.

2 Future of Spoken Dialogue Research

With the increasing use of speech technology, I like to think that further research will deepen knowledge of dialogue in human-machine interaction, as well as the knowledge of language learning.

As for the improvement in service oriented by human-machine interaction, I believe that not only should machines be more able to recognise indirect speech acts that may lead to misinterpretation, but also offer an output that is equally more easily recognised, identified, and interpreted by humans, as it would be the case of machines providing accented-speech for members of certain communities in multicultural environments.

With regard to dialogue for language learning, studies will shed more light into the possibilities of non-experts developing their own dialogue systems according to their needs. Therefore, teachers and students would benefit more if they could customise their own dialogue models, fitted according to their own reality and necessities.

3 Suggestions for Discussion

- To what extent can dialogue systems be user-friendly? How can they be endowed with 'social competence'?
- How can dialogue systems benefit from a even wider multidisciplinary interaction?
- Is the multimodal approach applicable and suitable to any dialogue system?

References

C. T. Best and M. D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Bohn, Ocke-Schwen and Murray J. Munro (Eds.), *Second language speech learning: The role of language experience in speech perception and production*, pp.13—34. Amsterdam: John Benjamins.

M. Celce-Murcia. 1987. Teaching pronunciation as communication. In J. Morley (Ed.), *Current perspectives on pronunciation*, pp.5–12. Washington, D.C.: TESOL.

J. E. Flege. 1995. Second Speech Learning: Theory, Findings, and Problems. In W. Shange (Ed.), *Speech Perception and Linguistic Experience – Issues Cross-Language Research*, pp.233–277. Timonium: York Press.

S. Gass and E. Varonis. 1984. The effect of familiarity on the comprehensibility of nonnative speech. In *Language Learning* 34, pp.65-–89.

D. C. Kluge, M. S. Reis, D. Nobre-Oliveira, and A. S. Rauber. 2008. Intelligibility of accented speech: the perception of word-final nasals by Dutch and Brazilians. In *Procs. of V Jornadas en Tecnología del Habla*, pp. 199–202, Bilbao.

P. K. Kuhl and P. Iverson. 1995. Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, pp.121–154. Timonium, MD: York Press.

I. R. A. Mackay, J. E. Flege, and S. Imai. 2006. Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. In *Applied Psycholinguistics* 27, pp.157–183

J. Morley. 1991. The Pronunciation Component in Teaching English to Speakers of Other Languages. In *TESOL Quarterly* 25/1, pp.51–74.

M. J. Munro, T. M. Derwing, and S. L. Morton. 2006. The mutual intelligibility of foreign accents. In *Studies in Second Language Acquisition* 28, pp.111–131.

M. S. Reis, D. Nobre-Oliveira, and A. Rauber. 2007. Effects of perceptual training on the identification and production of the English voiceless plosives by Brazilian EFL learners. In *Procs. of Fifth International Symposium on the Acquisition of Second Language Speech*, Florianopolis.

M. S. Reis. 2007a. O efeito de treinamento perceptual na percepção e produção das plosivas não-vozeadas do inglês. In *III CELLI Colóquio de Estudos Linguisticos e Literários*, Maringá.

M. S. Reis. 2007b. Guia de Pronúncia do Inglês para Brasileiros: Book review and experimentation. In José Marcelo Freitas de Luna. (Org.). *Educação e Linguistica: ensino de línguas*, pp.137–149. Itajaí: Univali Editora.

M. S. Reis and D. C. Kluge. 2008. Intelligibility of Brazilian Portuguese-Accented English Realization of Nasals in Word-Final Position by Brazilian and Dutch EFL Learners. In *Revista Crop* 13, pp.215–229.

L. Smith and J. Bisazza. 1982. The comprehensibility of three varieties of English for college students in seven countries. In *Language Learning* 32, pp.259-–269.

Biographical Sketch

Mara Reis is a 4th year PhD student in Second Language Acquisition, particularly interested in English Phonetics and Phonology learning. She received her MA in the same field in 2006, her BA in Visual Arts in 2003, and her BA in Dentistry in 1995. Last summer she completed part of her PhD studies at University College London under the supervision of Prof. Valerie Hazan.

Sylvie Saget

LLI-IRISA
6 rue de Kerampont
BP 80518 - 22305 Lannion Cedex
France

saget.sylvie@gmail.com
www.sylvie-saget.cabanova.com

1 Research Interests

My research interests lie generally in the area of **dialogue management**, with a special focus on the **grounding process** and **common ground modelling**.

1.1 Grounding and Common Ground

Considering dialogue has a collaborative activity while designing spoken dialogue systems aims at enhancing a system's robustness. Indeed, such a modelling is a way of handling non-understandings by giving a dialogue system and its users the capacity to interactively refine their understanding until a point of intelligibility is reached.

The interaction between the different fields concerned by collaborative activities may be a source of richness and of significant overhangs. In particular, in philosophical studies, the necessity of distinguishing the context-dependant (pragmatic) mental attitude which is acceptance, from the context-free mental attitude which is belief has been brought (again) to the forefront by a lot of philosophers.

We integrate this distinction within a formal model of dialogue management which leads to architectural proposal on information modelling within dialogue systems. The fundament of this approach has been presented in [Saget 2006; Saget and Guyomard, 2007], the collaborative model of dialogue and of reference in [Saget and Guyomard, 2006; Saget and Guyomard, 2007], and a first step for linking this model to reference treatment in [Saget, 2007].

1.2 Robust and adaptive/adaptable interface dedicated to UVs Systems

The next generation of Unmanned Vehicle (UV) System will include several vehicles with high autonomous capabilities. As a consequence, such systems will require new forms of Human-System interaction.

LUSSI department of Telecom Bretagne has developed a prototype multi-Uvs ground station control (SMAART) that allow an operator to supervise the surveillance of a simulated strategic airbase by a swarm of a rotary-wing UVs.

In this project, we adapt our approach of the grounding process to the specific design of operator's interface of UV Systems [Coppin; Legras and Saget, 2009; Saget, Legras and Coppin, 2008]. Dialogue management is also an adaptable and adaptive process in order to avoid negative side effects of automation.

2 Future of Spoken Dialogue Research

One of the main limitations one may encounter while designing a dialogue system is to find the proper toolkit, dialogue System designer, as well in a commercial as well as in a research perspective, may need to use a simple toolkit allowing him to develop a specific, adaptive and possibly multi-tasking interface. Then, effort has to be done in order to develop a wide range of such toolkits corresponding to the major kind of support, task complexity, etc.

Another important aspect is to enhance interface interactivity by improving user trust in spoken dialogue capacity. Integrating alignment processes in spoken dialogue systems is a way to highly enhance trust as well as decreasing system complexity. Moreover, users of dialogue systems need to have a sufficient knowledge of system's capacity (vocabulary, level of interaction flexibility, etc.) to have a suitable trust level. Then, how do we train future users of a dialogue system such that they will have an accurate trust level? How do we verify an online user's trust and how correcting it if necessary?

3 Suggestions for Discussion

I would suggest the following subjects of discussion:

- Interactive alignment in spoken dialogue interaction: usefulness in HCI, what level has to be aligned, are alignment process useful for user modelling and learning?
- User's training process: What is the main difference between professional and public large applications? When and how accurately does online help assist dialogue system users?

References

Gilles Coppin, François Legras, and Sylvie Saget. 2009. Supervision of Autonomous Vehicles: Mutual Modeling and Interaction Management. *HCI International 2009*, San Diego, CA, USA.

Sylvie Saget, François Legras, and Gilles Coppin. 2008. Collaborative model of interaction and Unmanned Vehicle Systems' interface. *HCP workshop on "Supervisory Control in Critical Systems Management", 3rd International Conference on Human Centered Processes (HCP-2008)*, Delft, The Netherlands.

Sylvie Saget and Marc Guyomard. 2007. Doit-on dire la vérité pour se comprendre ? Principes d'un modèle collaboratif du dialogue basé sur la notion d'acceptation. *Quatrième Journées Francophones des Modèles Formels de l'Interaction (MFI'07), Annales du Lamsade*: 239–248, Paris, France.

Sylvie Saget. 2007. Using Collective Acceptance for modelling the Conversational Common Ground: Consequences on referent representation and on reference treatment. *Proceedings of the 5th IJCAI's Workshop on Knowledge and Reasoning in Practical Dialog Systems*: 55–58, Hyderabad, India.

Sylvie Saget and Marc Guyomard. 2006. Goal-oriented Dialog as a Collaborative Subordinated Activity involving Collaborative Acceptance. *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial 2006)*: 131–138, University of Potsdam, Germany.

Sylvie Saget. 2006. In favor of collective acceptance: Studies on goal-oriented dialogues. *The Fifth International Conference on Collective Intentionality (CollInt V)*, Helsinki, Finland.

Biographical Sketch

Sylvie Saget is preparing a PhD in Computer Science at the University of Rennes 1 since 2003. She is also a board member of ISCA SAC as the General Coordinator. In 2008, she worked at Telecom Bretagne in the LUSSI Department for a one-year project.

She obtained the M. Sc. and B. Sc. in Computer Science, respectively in 2003 and 2002, with a major in Artificial Intelligence from the IFSIC Institute of the University of Rennes 1.

Nur-Hana Samsudin

School of Computer Science
University of Birmingham
B15 2TT Edgbaston
United Kingdom

n.h.samsudin@cs.bham.ac.uk
www.cs.bham.ac.uk/~nhs/

1 Research Interests

My research interest revolves around **Speech Processing** and specifically **Text-to-Speech (TTS) Synthesis**. The focus of my current work is on **multilingual/polyglot TTS** and **linguistic modelling for cross languages**. I am also interested in implementing related **Speech Recognition** models as well as issues of **Speaker Adaptation** into my research.

1.1 Polyglot Speech Synthesis

Automatic Speech Synthesis is an emerging technology with applications in many different domains such as Human-Computer Interaction, Assistive Technology and Speech-to-Speech Systems. However Speech Synthesis or Text-to-Speech (TTS) Systems require extensive natural language resources. Unfortunately such resources are not easily available for all languages. Such requirements make it even more difficult to be achieved by multilingual TTS application. The issues are not only because of the resources; it is also about how different system needs to be developed separately for different languages (Samsudin, 2009). Information reuse for polyglot TTS may be a solution. Information reuse will make it possible to adopt existing data and rules within a polyglot TTS system which makes it possible to develop just one speech engine for multiple languages (Latorre, 2006).

1.2 Linguistic Modelling for Polyglot Speech Synthesis

There are two parts in TTS architecture; the front end; which involves text processing, natural language processing and prosody assignment; and the back end; which involve all signal processing aspects in speech generation. Currently, I am working on the front end of a TTS system and doing research to develop a model to represent the linguistic aspect of the speech. There are two components in this model; the phonetic processing component and phonological processing component.

In phonetic data preparation, the goal is to come out with a finite list of phonemes which include the phonemes used in languages available in the resources. The processing should be able to provide phoneme mapping where a pronunciation dictionary is not available. Currently, phonemes from 31 languages have been extracted, analysed and indexed. These phonemes come from MBROLA database. I have selected MBROLA because that is the most comprehensively available phonemes set available. There are 6 language families involve: Indo-European (20), Afro-Asiatic (2), Uralic (2), Austronesian (3), Altaic (3) and Dravidian (1) language.

In phonological processing, the process will be focussed on creating the effect of speech production as close to the native speaker's speech as possible. To achieve that target, a few rules need to be identified. The rules which are being considered are: general phonological rules, language specific phonological rules, language pronunciation clusters and language intonation templates. These rules need to be constructed to complete phonological processing component.

The implementation is targeted to fit into the Festival framework. Festival is selected due to do its framework flexibility. Festival already comprises the core architecture of a TTS system (CSTR, 2004). Theoretically, if we can create a Malay TTS in Festival (which is not available in the available text and speech corpus in the Festival framework), then we can create an X TTS using Festival where the X is actually a combination of all languages in question. A lot of other issues will then arise. How is it possible to represent all phonemes in the languages in question? Since it is possible to represent each sound unit based on the IPA, how is it possible to represent the phonological rules of different languages using the identified list - there might be rules conflict when we are discussing for one language as compared to another. One way to solve this is by using set of training data of language in question. One possible advantage is of course to get the phonological rules right for each language and another is to get the spectral value correct in order to generate the speech for the language in question. This however will make the speech's voice restricted to the trainer's timbre only.

It is hoped that by developing this model we will be able to make the process of creating a multilingual/polyglot TTS system less time consuming and achievable to build with less resource languages, or minority languages or even with less linguistic information or technical expert.

2 Future of Spoken Dialogue Research

In the next 5 to 10 years, I predict that research will focus more on the standardisation of resources in spoken language processing so that resources are applicable for other languages and other domain. I also hope that tools to aid system development will be more accessible and resources will have a standard format to increase usability.

The future of spoken dialogue research is progressing toward assistive usage not only for disable but also for everyday need. The current systems like speech-to-speech translation, speech-to-speech system, question-answering system and edutainment system are still in the process of perfection. In the near future, these components will be able to be integrated into a fully working robotic system and other domain specific spoken dialogue applications to create a generation of ubiquitous computing.

3 Suggestions for Discussion

Some topics which maybe of interest are:

- Creating a standard format for corpora
  Lexicon, speech and text corpora may be easily available. But applying different corpora requires modifications in the particular application that want to make use of those resources.

- Resource Sharing - What are the limitations
  MBROLA, Festival, FestVox, Sphinx and Dragon are among the applications that have successfully created an engine which allows a single platform to be applied for multilingual applications. Discussion on resource sharing for spoken dialogue application components, module and tools may be able to help new scientist not to make the expensive mistakes in data collection.

- From Generic to Domain Specific Applications
  Some tools like HTK, HTS and FestVox have been developed to aid speech applications development. Addressing the development issues faced by the developer of the tools and conversion issues faced by the users of the tools will anticipate future developed tools with the 'compulsory' requirements.

References

Nur Hana Samsudin. 2009. *Reusing Multilingual Resources for Polyglot Speech Synthesis*, Poster Presentation, University of Birmingham, UK.

Javier Latorre and Koji Iwano and Sadaoki Furui. 2006. New Approach to the Polyglot Speech Generation by means of an HMM-based Speaker Adaptable Synthesizer. *Speech Communication*, 48(10):1227–1242.

The Centre for Speech Technology Research. 2004. *Festival*. http://www.cstr.ed.ac.uk/. University of Edinburgh, UK.

Biographical Sketch

Nur-Hana Samsudin is a PhD student at the School of Computer Science, University of Birmingham, United Kingdom under Dr Mark Lee's supervision. Previously she obtained Bachelor of Computer Science and Master of Science from Universiti Sains Malaysia. Her MSc work was on Malay Speech Synthesis. Her previous work revolved around unit selection speech synthesis, adjacency analysis and prosody analysis of Malay. After submitting her MSc dissertation, she worked as a researcher at MIMOS Bhd (Malaysia) under the Artificial Intelligence Centre in Speech Processing Group. She then resigned from the company to further her studies and currently she is sponsored by the Ministry of Higher Education, Malaysia.

Stefan Schaffer

Technische Universität Berlin
Graduiertenkolleg Prometei
Franklinstraße 28-29
10587 Berlin

stefan.schaffer@zmms.tu-berlin.de
www.tu-berlin.de/?id=sschaffer

1 Research Interests

My research interest is the development of **model-based methods** for **simulation** and **automated usability prediction** of **multimodal interfaces**. In particular I want to investigate how users' modality choices can be predicted and simulated by a computational model estimating the quality of a multimodal interface. Therefore I am also interested in exploring rules and cognitive processes which impact multi-modal interaction.

1.1 Spoken Language System related Work

At Deutsche Telekom Laboratories (T-Labs) I worked on the integration of a janus speech recognition module (Finke et al., 1997) in so called Attentive Displays. The Attentive Displays are an interactive wall mounted information system for employee and room search in an office environment. Originally the system was controlled with a touch screen only. To enhance the input facility speech recognition with acoustic models for distant speech was embedded. Thereby the system input was altered from unimodal to multi-modal.

The fusion module of the dialogue manager was based on a simple first-come, first-served strategy, because users can interact with the system via touch screen or speech in each dialogue state. Furthermore the system was represented as a finite state machine whereas the grammar of the speech recogniser was updated at each state change.

1.2 Research Work

Within my research group I conducted usability studies with unimodal and multimodal system versions of the Attentive Displays to investigate how ratings of single modalities relate to the ratings of their combination (Wechsung et al., 2009a). To collect subjective user judgements the AttrakDiff questionnaire was used (Hassenzahl et al., 2003). We figured out that for overall and global judgments ratings of the single modalities are very good predictors for the ratings of the multimodal system. Additionally the results indicate that the modality used more frequent in multimodal interaction has a higher influence on the judgment of the multimodal version than the less frequent used modality. However, for separate usability aspects (e.g. hedonic qualities) the prediction was less accurate. The findings of this study were limited to the tested system and test design. The multimodal version was always the system tested last. We conducted a follow-up study with an enhanced version of the system (Wechsung et al., 2009b). Thereby we changed the order with the multimodal system version tested first. The results of the first experiment could partially be supported. The prediction of the ratings of the multimodal system via overall and global judgments was still possible. The hypothesis that lower prediction performance is partially a consequence of the participants' effort to rate consistently was supported.

Moreover we examined influences of user characteristics (training) on direct an indirect measures for the evaluation of multimodal systems (Seebode et al., 2009). We found some differences in the interaction patterns of trained and untrained users, as experts needed less time and fewer trials to solve the tasks and used speech more frequently as input modality than novice users did. We could observe different groups of users with different interaction patterns that have an influence on their ratings.

1.3 Future Work

My previous research concerning usability evaluation via direct and indirect measures indicated that the use of multimodal human-machine interfaces is influenced by certain user interaction patterns. I want to investigate if these patterns can be characterised by rules that explain multimodal interaction. My further experiments will focus on how users of a multimodal system choose the interaction modality. Rules should be derived from the observed interaction patterns.

In a next step the rules should be used for model-based evaluation. Therefore the MeMo workbench for automated usability testing will be extended (Engel-brecht et al., 2008). Test scenarios and prototypes will be modelled and simulated. The data gathered through simulation will be compared to empirical data of previously conducted experiments. An additional verification of the findings can be provided through an ACT-R model that implements the rules and test scenarios in a simulation

(Anderson, 1996).

2 Future of Spoken Dialogue Research

My impression is that recently only small improvements on the part of speech recognition technology are made. The integration of specialised acoustic models and language models seems to have only little impact on a further enhancement of the reliability of dialogue systems. However an improvement of the recognition capabilities is necessary to build systems of high acceptance. Hence researchers should also investigate how a system can profit from additional information like e.g. knowledge about the user. By means of this a system could adapt components like dialogue strategies, grammars or acoustic models to the actual user. This may beneficially affect the reliability of the recognition.

3 Suggestions for Discussion

- Model driven evaluation and modelling of user behaviour.
- Additional knowledge bases for future dialogue systems.
- How to define rules for the simulation of interaction.

References

J. R. Anderson. 1996. ACT: A simple theory of complex cognition. *American Psychologist*, 51: 355–365.

K.-P. Engelbrecht, M. Kruppa, S. Möller and M. Quade. 2008. MeMo Workbench for Semi-automated Usability Testing. In: *Proc. Interspeech 2008*

M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. 1997. The Karlsruhe-Verbmobil Speech Recognition Engine. In: *Proc. ICASSP '97*, Munich, Germany.

M. Hassenzahl, M. Burmester, , and F. Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Ziegler J., Szwillus G. (eds.) *Mensch & Computer 2003. Interaktion in Bewegung*: 187–196, B.G. Teubner, Stuttgart.

I. Wechsung, K.-P. Engelbrecht, S. Schaffer, J. Seebode, F. Metze, and S. Möller. 2009a. Usability Evaluation of Multimodal Interfaces: Is the whole the sum of its parts?. accepted for: *13th International Conference on Human-Computer Interaction*.

I. Wechsung, K.-P. Engelbrecht, A. B. Naumann, S. Schaffer, J. Seebode, F. Metze, and S. Möller. 2009b. Predicting the quality of multimodal systems based on judgments of single modalities. accepted for: *Interspeech 2009*.

J. Seebode, S. Schaffer, I. Wechsung, and F. Metze. 2009. Influence of Training on Direct and Indirect Measures for the Evaluation of Multimodal Systems. accepted for: *Interspeech 2009*.

Biographical Sketch

Stefan Schaffer studied Communication Science and Computer Science at the Technical University Berlin. After receiving his Magister Degree in 2009, he has joined the research training group prometei. Currently he is working towards his PhD thesis in the domain of automated usability evaluation for multimodal systems.

Alexander Schmitt

Dialogue Systems Research Group
University of Ulm
Albert-Einstein-Allee 43
89081 Ulm

alexander.schmitt@uni-ulm.de
www.uni-ulm.de/in/it/staff/ds/
alexander-schmitt.html

1 Research Interests

My research topic is centered around **stochastically-based detection of problematic dialogue situations** in telephone-based speech applications. I apply **statistical classifiers** trained on log-data for detecting such situations. Apart from this dialogue feature-based detection, my work includes the **robust detection of the user's emotional state (especially anger)** and the incorporation of speaker knowledge such as age and gender into the recognition process. Also, I am interested in the field of speech recognition on mobile devices, where I have done research during my Master thesis project (Zaykovskiy and Schmitt, 2007; Zaykovskiy et al., 2007; Zaykovskiy and Schmitt, 2008).

1.1 Problematic Dialogue Situations

In the telephone application context, by "problematic" we mean situations where the caller is about to hang up without completing the task. Reasons for that might be that he is annoyed by automated agents in general or stuck and helpless by faulty system behavior. Especially in larger conversations such as with automated technical support agents as described in (Acomb et al., 2007), where dialogues not rarely consist of 50 system- and user turns, the ability to repair a problematic situation would lower the drop-out rate.

Given my past work, my goal is to render SDS more "intelligent" than they are today. The central question is: Is it possible to detect problems based on previously observed calls and if so, how robust is this approach?

1.2 Dialogue Features for Problematic Dialogue Prediction

My first Problematic Dialogue Predictor was based on turn-wise dialogue features that have been logged during previous calls such as ASR accuracy, ASR rejection, current system state, system prompt, user answer etc. based on a confidence-rated rule learner, similar to (Walker et al., 2002). It turned out, that it worked quite well in early points in the dialogue, but had flaws when the dialogues became longer (Schmitt and Liscombe, 2008).

Additional features had to be developed. Into our classification, I included acoustic information (by early fusion), assuming that angry callers have a lower task completion rate. It did not improve the results since the acoustic features performed too poor in contrast to the dialogue features (Herm et al., 2008).

1.3 Anger Recognition

Late fusion of acoustic information with dialogue features seems to be more promising than early fusion. In recent work, I have thus developed an anger detection system in order to provide additional information to the Problematic Dialogue Prediction task which will be lately fused with the log-data classifier. It is based on a combination of a Support Vector Machine, a Neural Network and a Decision Tree (Schmitt et al., 2009).

My current goal is to make this anger detection system as robust as possible by including all available acoustic, lexical and dialogue features into an anger/non-anger classifier. I further consider the dialogue context and previous emotions of the caller.

In parallel I am developing additional lexical-, dialogue- and context features for the Problematic Dialogue Predictor, such as bag-of-words, ratios of appropriate user answers and inappropriate ones, probabilities of task success in the current system state.

Another central question that I want to answer is the impact of age and gender in the ability to cope with an automated agent. Might there be a difference between senior users and younger users, as well between male and female users? Beside the labelling of the corpus, I plan to develop an age and gender recognition system whose output is incorporated into the Problematic Dialogue Predictor.

2 Future of Spoken Dialogue Research

In my estimation, telephone-based applications are *the* platform for proving sense and nonsense of SLD technology. Here, the acceptance of the users has to be won. While most people have already had "conversations" with those Interactive Voice Response Systems (IVRs), they often have a distaste for them. The telecommunication provider O2 even just started an ad campaign in Germany proudly announcing to send the "robots" (i.e. IVRs) into retirement.

One of our most important tasks in my view is to increase the acceptance of SDS by endowing them with "intelligence" and empathy. SDS have to appear more on equal footing with the user and have to look more competent and sensitive than they are today.

Future interfaces should be able to adapt to the user's peculiarities, intentions and emotions, i.e., the interface should be capable of identifying and understanding the feelings, ideas, and (personal) circumstances of its users.

One of the decisive points here is the further progress of speech-based emotion recognition. Here, additional real-life corpora have to be collected that are different from most acted corpora that are currently available to the community. Further, user-modelling in general will be a central research topic in the next 10 years: who is the user, what are his or her preferences, is the current situation suited for pro-active behaviour from the part of the system, etc.

Since SDS will gain further importance in the near future (intelligent homes, automotive applications, multimodal information systems), the field of SDS will further grow. However, we have to ensure to deliver pragmatic research results that can be contemporarily realised in the field.

3 Suggestions for Discussion

- Emotion Recognition: Are we at the limit? No more classification accuracy to gain?
- Telephone-based SDS: are Interactive Voice Response systems doomed to death... and what comes next?
- Distributed Speech Recognition: a technology with future or superfluous invention?
- Problems in SDS: how to detect and how to prevent them?

References

Kate Acomb, Jonathan Bloom, Krishna Dayanidhi, Phillip Hunter, Peter Krogh, Esther Levin, and Roberto Pieraccini. 2007. Technical support dialog systems:issues, problems, and solutions. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 25–31, Rochester, NY, April. Association for Computational Linguistics.

Ota Herm, Alexander Schmitt, and Jackson Liscombe. 2008. When calls go wrong: How to detect problematic calls based on log-files and emotions? In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, September.

Alexander Schmitt and Jackson Liscombe. 2008. Detecting Problematic Calls With Automated Agents. In *4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems*, Irsee (Germany), June.

Alexander Schmitt, Tobias Heinroth, and Jackson Liscombe. 2009. On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection? In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, SigDial Conference 2009*, London, UK, September. Association for Computational Linguistics.

M A Walker, I Langkilde-Geary, H W Hastie, J Wright, and A Gorin. 2002. Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, (16):293–319.

Dmitry Zaykovskiy and Alexander Schmitt. 2007. Java to Micro Edition Front-End for Distributed Speech Recognition Systems. In *The 2007 IEEE International Symposium on Ubiquitous Computing and Intelligence (UCI'07)*, Niagara Falls (Canada), May.

Dmitry Zaykovskiy and Alexander Schmitt. 2008. Java vs. Symbian: A Comparison of Software-based DSR Implementations on Mobile Phones. In *4th IET International Conference on Intelligent Environments*, Seattle (USA), July.

Dmitry Zaykovskiy, Alexander Schmitt, and M. Lutz. 2007. New Use of Mobile Phones: Towards Multimodal Information Access Systems. In *3rd IET International Conference on Intelligent Environments*, Ulm (Germany), September.

Biographical Sketch

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 and is currently carrying out his PhD research project (3rd year) at Ulm University. His topic is centered around the 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 Modelling and regularly gives lectures on the development of Voice User Interfaces in several places, e.g. the German University in Cairo, Egypt.

Niels Schütte

Dublin Institute of Technology
School of Computing, DIT Dublin
Kevin Street
Dublin 6, Ireland

niels.schutte@student.dit.ie

1 Research Interests

My research interest is in the area of **Multimodal Dialogue Systems**, especially in **situated dialogues**.

1.1 Past and planned work

As part of my Diplom degree I worked in the SmartChair project. In this project we developed a multimodal dialogue interface for the Bremen Rolland autonomous wheelchair platform.

My current research is part of the Lok8 project at DIT Dublin, which has started in early 2009. Goal of this project is to develop a multimodal dialogue system that allows users to access location based services using mobile devices. In particular, we are investigating the use of mobile devices such as iPhones or Google Android Phones to track the position of the user in an environment as well as the use of the sensors of those devices to recognise certain classes of gestures such as directed pointing. This is the subject of the “Tracker” strand of the project.

An animated avatar, that can either be displayed directly on the device, or on displays that are available in the environment, will be utilised for interaction with the user. This will be developed in the “Avatar” strand of the project.

Apart from conventional ideas of dialogic interaction we are also looking to integrate methods of sonification, such as audio based navigation, in the “Vocate” strand. One goal will be to research how those different ideas can complementarise each other.

My work in this project (in the “Contact” strand) is to develop and implement the dialogue framework of the system. This will among other things entail developing models to integrate the linguistic interaction with the user with information supplied by the mobile device. Information about the current location of the user and direction of the device in conjunction with spatial models of the environment will enable us to interpret pointing gestures as exophoric references to object in the environment. Information about general movements of the device may be used to interpret movement gestures like dismissive hand waving.

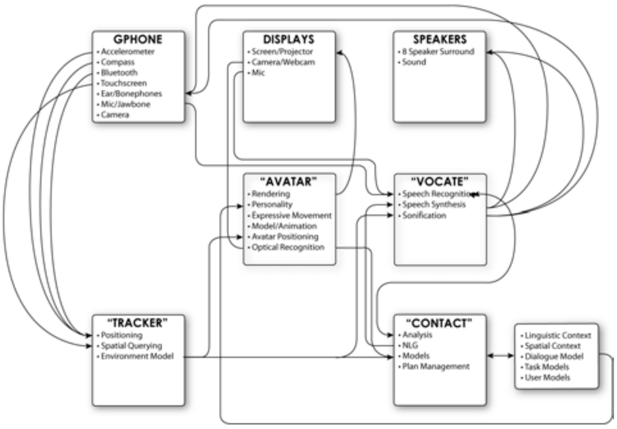


Figure 1: Lok8 system overview

Apart from that we are going to investigate the advantages of different types of interaction, when which type of interaction may be preferable and how we can smoothly transition between them.

The use of an avatar as a personification of the system poses several interesting research questions. For example we are considering the use of personalised avatars that, to a certain extend, may have their own personality traits, such as playfulness or seriousness. It will have to be researched how to integrate such individual behaviors with the general behaviors required by the system task.

Each strand is handled by a PhD Student from a different background, ranging from Computer Science to Design.

Since the project, and my part of it is quite broad in scope, and my work has only recently begun, I am still developing my research questions.

2 Future of Spoken Dialogue Research

We believe that an important topic in the future development of dialogue systems will be the use of dialogic interaction with mobile devices. This kind of dialogue will be situated and will have to take the modalities other than speech into account. The mobility of such a system may also open up questions of scalability of such systems. Some situations may make it necessary to reduce the activity of the system while maintaining a minimal functionality. For example it may be awkward to communicate with a system in a crowd of people. In such a situation it might be appropriate to reduce the activity of the system to audio cues that highlight points of interest or guide navigation. If both hands are occupied, it may be impossible to operate buttons or a touch screen. In such a situation the interaction should shift towards speech.

We think that this kind of scalability is important and should be researched. This question could also be generalised into researching general flexible models for varying availability of modalities, that would allow deployment on different hardware platforms.

3 Suggestions for Discussion

I think the following areas might be worth looking at

- How can multimodality be used to enhance the use of mobile devices for dialogue? How can weaknesses of mobile devices such as small screens be balanced out using other modalities?
- How can environments be efficiently modeled or annotated to interact with mobile dialogue systems?
- How can we model linguistic context and attention in a mobile situated dialogue system i.e. when context changes depending on the movement of the system?

Biographical Sketch

The author is currently a first year PhD student at Dublin Institute of Technology. He is supervised by Dr John Kelleher and Dr Brian Mac Namee. Previously the author studied Computer Science at the University of Bremen in Germany. The title of his Diplom thesis is “Automatic Generation of Concept Descriptions” and deals with generating natural language descriptions of ontology contents.

Álvaro Sigüenza Izquierdo

Universidad Politécnica de Madrid
Avda. Complutense 30
28040 Madrid
Spain

alvaro.siguenza@gaps.ssr.upm.es

1 Research Interests

My main research interests lie in the area **human-machine interaction**, with a special focus on **spoken dialogue systems** while the user is engaged in a **driving multitask environment**.

1.1 Past and ongoing research

Driving is a complex task that requires the interaction and coordination of physical, sensorial and mental driver skills. Consequently, it is necessary that a great part of the user attention will be oriented to the driving task with the purpose of accomplishing a high level of performance. Traditionally, the driver attention could be diverted to other tasks that might arise inside the vehicle, which are frequently interrelated with the driving task. Nevertheless, the proliferation of the in-vehicle information and communication systems has imposed new tasks which accumulation has led drivers to assimilate a lot of information in a brief period of time. When a demanding driving situation appears (i.e. an intersection), the driver might be too engaged in the in-vehicle system messages processing, putting drivers at risk of an accident.

In the driving context, the interaction between drivers and systems can not be considered as primary task, so that driver must pay their attention to driving safely. Traditional in-vehicle interfaces require the driver taking their hands from the wheel to press a button or looking away from the road to look at a display. However, speech interfaces allow drivers to keep their hands on the wheel and eyes on the road, considering itself as a non-distract interaction mode.

At GAPS (the Signal Processing Applications Group at UPM) we have been researching on speech human-vehicle interaction, collaborating on several projects that they deal with this thematic. Our main aim in this field of research is to adapt the in-vehicle spoken dialogue systems to the circumstances of the driving, focusing on the state of the driver.

1.2 Adapting in-vehicle spoken dialogue systems

Several studies have showed that, in spite of the benefits provided by vocal interfaces, there are situations where the interaction with the in-vehicle can entail distraction. This distraction might be caused by the **workload** related to the concurrence of driving and a task involved in a spoken dialogue (secondary task). To cope with this additional information, the driver needs to percept the aural message of the system, to process the information and to react to it, with the possibility of interfering with the driving task. According to the multiple resources model, the resources of the driver are limited and when the driving situation is too demanding and the same resources are needed for driving and secondary task, interference between tasks can impair driving if the driver is overloaded (Vollrath 2007). In addition, driving is a special context where recognition errors are likely to appear caused by environmental factors (such as noise), intrinsic user factors (mood state, fatigue…), task variables or even the design of the interaction flow, increasing the workload experienced by the driver (Kun et al. 2007).

With the aim of avoiding situations where the safety of driver can be affected by the interference between the driving and spoken dialogue systems, is necessary to carry out an evaluation of the vehicle context to take the opportune measures. From the driving context can be estimated an assessment of the workload experienced by the user, allowing to adapt the spoken dialogue to the workload. As the priority in this environment is given to the processing of the driving task, the driver shouldn't engage in a spoken dialogue so long as the junction of driving workload and dialogue workload don't exceed the driver capacity.

Accordingly, to carry out the control of the dialogue flow is necessary the presence of a **workload manager** that attempts to determine if a driver is overloaded or distracted, and if they are, alters the operation of in-vehicle systems (Green 2004). This workload manager has to consider what the right mode to interact with the driver is. If the driver is overloaded, might be necessary stopping or interrupting the dialogue flow to solve this situation. When this situation finished, the dialogue should be resumed just in the point where it was before its interruption. However, the control of the dialogue might make use of additional information (such us driving context, user preferences…), providing with intelligence to the dialogue in order to focus it and to decrease the dialogue turns and consequently the workload. For example, if an application knows that the driver like vegetarian restaurants, it might suggest that he stop to have lunch without the necessity of asking him about his preferences. In the same way, depending on the driving situation might be better presenting the information of the system in other modality of interaction (visual display, haptic actuator).

Since, there are few studies focusing on the workload management of a spoken dialogue system when appearing concurrently with the driving task, GAPS have developed a bench simulator to research on all of these presented issues (Blanco et al. 2009).

2 Future of Spoken Dialogue Research

I believe in the near future we will see a proliferation of spoken dialogue systems in a wide range of context such as the vehicle. This will cause a great concern for analysing the effects that the concurrent performance with other tasks produces on the curse of the dialogue. Due to in some contexts the spoken interaction with a system can not be the priority task (i.e. driving), a study about the workload generated by the dialogues is needed to ensure the right performance of the primary task. We need to incorporate technologies for representation of context in order to adapt the dialogue flow to the workload situation. In the same way, will be necessary a research on methods that allow to detect overloaded users a-priori.

In the driving context, to carry out these aims will be important a simulation bench where we can evaluate different strategies of workload adaptation without putting drivers at risk. Anyway, a modelling of the structure of human attention resources is needed to can consider the spare resources when we analyse what is the best way of interaction in a specific instant.

3 Suggestions for Discussion

My suggestions for discussion are the following:

- Incorporation of contextual information and inference technologies to the spoken dialogue considering the time restrictions in this kind of systems.

- How it should be resumed the spoken dialogue system when it is interrupted due to the user is overloaded or caused by adverse driving situations?

- Strategies to reduce the workload in dual task environments.

- Methods to detect overloaded users using only speech information of the dialogue.

References

Kun A., Paek T., and Medenica Z. 2007. The Effect of Speech Interface Accuracy on Driving Performance. *Interspeech 2007*.

Vollrath M. 2007. Speech and driving-solution or problem? *Intelligent Transport Systems*, IET 1: 89–94.

Green P. 2004. Driver Distraction, Telematics Design, and Workload Managers: Safety Issues and Solutions. *SAE paper* Number 2004-21-0022.

Nishimoto T., Shioya M., Takahashi J. and Daigo H. 2005. A study of dialogue management principles corresponding to the driver's workload. *Biennal on Digital Signal Processing for In-Vehicle and Mobile Systems*.

Blanco J.L., Sigüenza, A., Díaz D., Sendra M. and Hernández L.A. 2009. Reworking spoken dialogue systems with context awareness and information prioritisation to reduce driver workload. In: *Proceedings of the NAG-DAGA 2009, International Conference in Acoustics*.

Biographical Sketch

Álvaro Sigüenza Izquierdo has a MEng in telecommunications from Universidad Politécnica de Madrid. He is currently a PhD student and holds a research position in this university under the supervision of Luis Hérnandez.

Vasile Topac

Department of Computer Science and
Information Technology
Politehnica University of Timişoara
Bd. Vasile Pârvan nr. 2
RO-300223 Timişoara - Romania

vasile.topac@aut.upt.ro

1 Research Interests

My research lies on **improving the accessibility to text-based information** (TBI). By text-based information I mean text having different media representations (e.g. speech, digital text, printed text). For this I am focused on **designing software applications** which are **integrating technologies** like **speech recognition**, **spoken dialogue systems**, **machine translation**, text processing, **TTS**, OCR and others for improving the access to TBI. Also my research extends to the topic of **text encoding**.

1.1 Past Work

While working for my diploma project I have developed an application for improving access to printed documents. So the application allowed the user to place a document under a high resolution webcam or into a scanner and listen the text spoken in any language ("any" language stands for a limited set of languages, of course...). Even if the application was designed for users with visual disabilities, it has proven to be useful in many other scenarios. Since then I have identified more limitations to TBI accessibility and decided to address some of them.

1.2 Accessibility Limitations

When trying to access TBI, depending on how it is represented a user may encounter several limitations. I have identified and addressed four main limitations:

- Physical (user has hearing or vision problems);
- Language (the information is presented in a language which is unknown by the user);
- Specialised Text (the information is specialised on a specific domain, like medicine, being hard to understand for normal users)
- Time (the user has to access a big quantity of information in a limited time).

1.3 Current Work

My current work is focused on creating software architectures and developing solutions which are trying to overstep the limitations described above.

Starting from the applications I have previously described I continued with an application that improves the access to TBI, and addresses the four limitations. Currently I work on text adaptation methods and on adding text summarisation features, so that the user can listen to a resume of a printed page or of an audio file. I'm currently researching how this kind of applications can be controlled by using spoken dialogue systems. Using this type of applications leads to the problem of storing TBI so that it can be presented in any media representation and any language without changing meaning of the information. To do this the text has to be encoded with enough metadata. The existing xml-based formats like VoiceXML or TEI have to be extended in order to store enough information about the text.

1.4 Spoken Dialogue Systems and TBI

Communication between spoken dialogue systems and applications like the one described above can completely change the way users interact with TBI. While most of the applications using spoken dialogue systems the focuses on the communication, here the focus will be on the text and on increasing TBI accessibility. So imagine applications that allow the users to query audio files or printed pages containing TBI in any language, allowing the usage of interpretation or summarisation features.

2 Future of Spoken Dialogue Research

I believe that the spoken dialogue systems will get closer to text-based information accessibility. I'm thinking at desktop applications for text visualisation or editing having integrated spoken dialogue systems and allowing "intelligent" document interrogation in a multilingual fashion.

I can imagine scanners or smart phones with high resolution cameras having embedded spoken dialogue systems for querying a paper in a multilingual fashion; also mp3 players or smart phones having embedded spoken dialogue systems not only for navigation, but also for interaction with the content of an audio file.

I believe that the technology is already here and is mature enough to do these steps.

3 Suggestions for Discussion

- How can spoken dialogue modules become easier to integrate into applications developed in different environments and running on different operating system?
- How to define metrics for measuring the quality of the dialogue?
- What about adding spoken dialogue system in devices not only for menu navigation but also for content interrogation? For example a smart phone having the ability to interrogate a mp3 file and to respond in a multilingual way?
- What about consuming spoken dialogue web services in smart devices?

References

Bohus, D. and Rudnicky, A. 2008. The RavenClaw dialog management framework: architecture and systems. In *Computer Speech and Language*, DOI:10.1016/j.csl.2008.10.001

Gruenstein A., J. Orszulak, S. Liu, S. Roberts, J. Zabel, B. Reimer, B. Mehler, S. Seneff, J. Glass, J. Coughlin. 2009. City Browser: Developing a Conversational Automotive HMI. In *Proc. CHI*: 4291–4296, Boston, April.

Ergun Biçici and Marc Dymetman. 2008. Dynamic Translation Memory: Using Statistical Machine Translation to Improve Translation Memory Fuzzy Matches, SpringerLink.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL).*

Kevin Knight and Philipp Koehn. 2003. *Statistical Machine Translation Tutorial.*

Topac, V. and Stoicu-Tivadar, V. 2009. Software Architecture for Better Text-Based Infromation Accessibility. *AICT '09. Fifth Advanced International Conference on Telecomunication*: 198–202. Digital Object Identifier 10.1109/AICT.2009.41

Biographical Sketch

Vasile Topac is a PhD student in the 1'st year of study at “Politehnica” University of Timişoara, Romania. He graduated at “Aurel Vlaicu” University of Arad, and has study one semester in Portugal at “Fernando Pessoa” University. He has a passion for programming, and he has 2 years of working experience as software engineer. He's a common friend of Java and C# while he worked with the first for two years, and played with the second in the Imagine Cup competitions, and also in research projects. When he is around science, he likes Einstein's words: “Imagination is more important than knowledge”.

Bogdan Vlasenko

Cognitive Systems IESK
Otto-von-Guericke University Magdeburg
D-39106 Magdeburg
Germany

Bogdan.Vlasenko@e-technik.uni-magdeburg.de
www.iesk.uni-magdeburg.de/Bogdan_Vlasenko.html

1 Research Interests

My research interests lie generally in the area **multimodal human-machine dialogue systems**, with a special focus on **human behavior modelling by multimodal emotion and intention recognition** in such systems.

At the first stage of my research I investigate opportunity of emotion recognition within speech. All evaluations are carried out on the public and commercial databases containing acted and spontaneous emotional colored speech. From public databases I single out EMO-DB (Berlin Database of Emotional Speech) and SUSAS (Speech Under Simulated and Actual Stress). From a list of commercial databases I deal with SmartKom and ABC (Airplane Behavior Corpus). In (Vlasenko, 2007 a,b), I describe a robust method for emotion classification within speech. Emotion recognition for simplest behavior models and deep acoustic analysis showed good results. Now I am modelling emotion recognition methods depending on prepotent user behavior model.

The next stage of my research assumes multimodal human behavior modelling. I want to combine Acoustic, Video (mimic, gestures) and Linguistic analysis for robust emotion and intention classification. I will realise early and late fusion of Spech and Video based Emotional Prosody classifiers. Interesting ideas about multimodal emotion classification based on acted audiovisual database can be found in (Schuller, 2007b).

In the framework of Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems (NIMITEK) project wdok.cs.uni-magdeburg.de/nimitek. I am realising speech recognition with intention and emotion classification in human-machine dialogue system. Foremost emotion classification is supposed to indicate problematic situations in real human-machine dialogue evaluation. A workgroup of NIMITEK project realised a Wizard-of-Oz experiment and collected an auxiliary audiovisual database of simulated human-machine dialogues with substantial amount of spontaneous emotions.

2 Future of Spoken Dialogue Research

Taking into consideration experience of Verbmobil, SmartKom (Streit, 2006) and SmartWeb (Hacker, 2006) projects, prosody analysis plays a very important role for dialogue system modelling. Automatic dialogue systems get easily confused if the recognised speech cannot be matched with the dialogue turn. Besides noise or other peoples conversations, even the users utterance can cause difficulties when he is talking to someone else or to himself (“Off-Talk”). For this reason prosody analysis combined with video user gaze detection can be used for automatic classification of the users focus of attention. Integration of On-Talk and Focus of attention recognition can increase the level of reliability and flexibility of human-machine dialogue system for Intelligent House solution. Such a system can identify objects of user attention and a set of possible user interests.

At the present moment the community of spoken dialogue system researchers became aware of the fact that emotions, mood, intentions and other attitudes play an important role in natural communication. Emotions may work as a self-contained dialogue move. They show complex relation to explicit communication. This insight in human behavior resulted in research on affective interfaces, development of user behavior models in context of human-machine interaction. The main aim of this is supporting the recognition of problematic dialogue situations by analysing the affective state of the user. On the other hand, for human-machine dialogue system for Intelligent House solution it will be good to realise classification of the users mood and condition. It will be possible to realise this with complex multimodal human behavior modelling. Such a system will have opportunity to cope with user bad mood or tiredness.

Speech recognition confidence can be used in conjunction with the semantic model of natural language understanding to promote correct machine response. Semantic models can help the system to detect possible thematic domains by spontaneous user request. Self-trained user dependent or universal language models for detection of thematic domains can advance us to a natural language human-machine communication era. In practice it will make possible to the dialogue system to automatically scoop and generate language models for detected thematic domain using World-Wide Web information resources and the current user dialogue history. In case of Intelligent House solutions, it will make possible for the system to collect user pleas-ing (according to user detected thematic domains) films, TV show, current cultural events, sport sights and others to make the users home life more comfortable.

In the near feature human-machine dialogue systems will consist of: user individual/universal multimodal behavior learning machine, robust Prosody analysis (Off-Talk, speaking style, accents and phrase boundaries detection), intention and focus of attention determination module, semantic thematic domain detection methods and self-trained language model for predefined thematic domains. Dialogue dependent semantic systems will be able to operate with the full spectra of human communication interfaces and multitude of user interests.

3 Suggestions for Discussion

- Collecting of realistic human-machine dialogue databases including sufficient amount of Prosody Events. Analysis of current data-bases. How we can collect real dialogues data-base in a better way?
- Role of human behavior modelling module in a human-machine dialogue system. How can human behavior modelling suppress problematic situations in user-system communication? Which kinds of behavior have to be providential?
- Multimodality in human-machine communication. How we can generate human natural communication interfaces? How we can use them in natural human machine interaction?
- How can we use Prosody in dialogue systems? How it is used in known projects (Zeissler, 2006)? Which new features can we add?

References

C. Hacker, A. Batliner, and E. Nth. 2006. Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In *Proc. Text, Speech and Dialogue. 9th International Conference*, Brno, Czech Republic.

B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. 2007. Towards more Reality in the Recognition of Emotional Speech. In *Proc. ICASSP 2007*, Honolulu, Hawaii, USA.

B. Schuller, D. Arcis, G. Rigoll, M. Wimmer, and B. Radig. 2007. Audiovisual behavior modeling by combined feature spaces. In *Proc. ICASSP 2007*, Honolulu, Hawaii, USA.

M. Streit, A. Batliner, and T. Portele. 2006. Emotions Analysis and Emotion-Handling Subdialogues. In Wahlster and Wolfgang (Eds.), *SmartKom: Foundations of Multimodal Dialogue Systems*: 317–332, Berlin : Springer.

B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. 2007. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. In *Proc. of InterSpeech 2007*, Antwerpen, Belgium.

B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. 2007. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing In *Proc. of ACII 2007*, Lisbon, Portugal.

V. Zeissler, J. Adelhardt, A. Batliner, C. Frank, E. Nth, R. P. Shi, and H. Niemann. 2006. The Prosody Module. In Wahlstera and Wolfgang (Eds.), *SmartKom: Foundations of Multimodal Dialogue Systems*: 139–152, Berlin : Springer.

Biographical Sketch

Bogdan Vlasenko is currently a 2nd year Ph.D. student at the Chair of Cognitive Systems, Institute for Electronics, Signal Processing and Communications, Otto-von-Guericke University, Magdeburg, Germany. He works under supervision of Prof. Andreas Wendemuth. He was born in Ukraine and received his B.Sc. and M.Sc. in Computer Science in National Technical University of Ukraine (KPI). His current research areas include multimodal human-machine dialogue modelling (human behavior model processing, emotion and intention recognition). His other interests include travelling, cinematography, swimming and communication.

Ina Wechsung

Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7
10589 Berlin

ina.wechsung@telekom.de

1 Research Interests

My research interest is the development of **subject-based** usability and user experience **evaluation methods** for **multimodal interfaces**. Furthermore I am interested in the cognitive foundations of multimodal interaction to explain users modality preferences, choices and ratings.

1.1 Evaluating multimodal systems

Although a lot of subject-based usability evaluation methods are available only few are designed for multimodal or spoken dialogue systems. The probably most common technique applied in subject-based evaluations are questionnaires, but a standardised and validated questionnaire addressing the evaluation of multimodal systems is still not available. Even for speech-based systems the probably most common questionnaire, the Subjective Assessment of Speech Interfaces questionnaire (SASSI) (Hone & Graham, 2000), still lacks final psychometric validation.

Thus I used several established questionnaires to assess the users opinion about multimodal systems. It was analysed in which aspects different established questionnaires lead to the same result and where there are inconsistencies across different questionnaires. The general objective was to investigate to which extent well known and widely used scales for the evaluation of graphical user interfaces and speech-based systems are appropriate for the evaluation of multimodal systems. It was shown that questionnaires designed for unimodal systems are not very applicable for usability evaluation of multimodal systems, since they seem to measure different constructs (Wechsung et al. 2008). The questionnaires with the most concordance were the AttrakDiff (Hassenzahl et al., 2003) and the SASSI (Hone & Graham, 2000). A possible explanation for that could be that the kind of rating scale, the semantic differential, used in the AttrakDiff is applicable to all systems. The semantic differential uses no direct ques-tions but pairs of bipolar adjectives, which are not linked to special functions of a system. In a next step these subjective ratings were validated with more objective and continuous parameters like log data. It was shown that the subjective data (questionnaire ratings) and the interaction data (task duration) showed concordances only to a limited extend (Naumann et al., 2008). In line with the earlier findings the questionnaire ratings most consistent to interaction data was the AttrakDiff questionnaire. Thus this questionnaire measures the construct it was developed for.

1.2 Relationship between ratings of single modalities and multimodal systems

Since the AttrakDiff showed the highest validity and reliability I continued to use the AttrakDiff questionnaire to examine how judgments of unimodal system versions relate to the judgments of the multimodal version (Wechsung et al., 2009). It was shown that accurate predictions of overall and global ratings of multimodal systems are possible on the basis of the respective ratings of the unimodal systems. Ratings of the multimodal systems matched the weighted sum of the ratings of the unimodal systems with the modality used more often having the stronger influence. However regarding specific measures (in this case three of the AttrakDiff subscales) prediction performance was lower. Only for one subscale (*hedonic-quality-stimulation*) equally good prediction performance as for the overall and global scales was shown. But these findings were limited to the test design: The multimodal version was always the system tested last. Therefore it is possible that the participants tried to rate consistently, adding up their single-modality judgments in their minds. Consequently, the judgments of the multimodal version would not represent the actual quality of that system. So, in a follow-up study the order was changed with the multimodal system version tested first. As in the previous study (Wechsung et al., 2009) prediction was best for overall and general judgments and judgments on the scale *hedonic quality-stimulation*. The poor prediction performance for the scale *hedonic quality-identity* might be explained by the underlying construct measured via this scale. The theoretical base of the AttrakDiff questionnaire is Hassenzahls model of user experience (Hassenzahl et al., 2003). This model postulates that a products attributes can be divided in hedonic and pragmatic attributes. Hedonic refer to the products ability to evoke pleasure and emphasise the psychological well-being of the user. Hedonic attributes can be differentiated into attributes providing stimulation, attributes communicating identity and attributes provoking memory. The AttrakDiff aims to measure two of these hedonic attributes: stimulation and identity. A product can provide stimulation when it pre-sents new interaction styles, like for example new modalities. Thus multimodality should be measurable with the scale *hedonic quality-stimulation*. Furthermore it should, and this was the case in all experiments, enhance a products ability to provide stimulation. The scale *hedonic-quality identification* measures a products ability to express the owners self. The system (a room management and information system) we tested in the experiments was not developed (and is not available) for personal use. Moreover it was custom tailored for Deutsche Telekom Laboratories (T-Labs). None of the participants was employed at T-Labs. Thus this system was in none of its versions de-signed to promote self expression or identification by communicating personal values. Furthermore the test was very task-oriented and goal-oriented. All tasks referred to the systems functionality for daily business of T-Labs employees and had no actual relevance for the participants. Data confirmed that users rated neu-tral on this scale. Thus this scale might have been an inappropriate measure. However regarding the scale *pragmatic quality* measuring to a large extent the concept of usability no such explanation can be found and further research how single modalities influence ratings of multimodal systems should be conducted at this point.

2 Future of Spoken Dialogue Research

In my opinion todays spoken dialogue systems are only under very specific circumstances preferable over a graphical user interface. Eyes and hands busy situation like the in-car scenario are few examples were I think speech control is actually useful. With the possibility to offer different modalities in multimodal systems speech will be one option and the actual usage will depend on the tasks, the situation and individual preferences. To offer situation and task adequate modalities more research needs to be done in investigating the underlying cognitive processes of multimodal interaction.

3 Suggestions for Discussion

- How to get away from this self-made questionnaire lacking psychometric validation?
- Influence of test situation on subjective ratings?
- Importance of user experience?

References

Hassenzahl, M., Burmester, M., and Koller, F. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualitt. In Ziegler J., Szwillus G. (eds.) Mensch & Computer 2003. *Interaktion in Bewegung*: 187–196. B.G. Teubner, Stuttgart.

Hone, K.S., and Graham, R. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3/4), 287–305.

Naumann, A. B., and Wechsung, I. 2008. Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures. In E. L.-C. Law, N. Bevan, G. Christou, M. Springett & M. Larusdottir (Eds.), *Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement*: 8–12.

Wechsung, I., and Naumann, A. B. 2008. Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. In E. Andr, L. Dybkjr, W. Minker, H. Neumann, R. Pieraccini & M. Weber (Eds.), *Perception in Multimodal Dia-logue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*: 276–284. Heidelberg: Springer.

Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F., and Mller, S. 2009. Usability Evaluation of Multimodal Interfaces: Is the whole the sum of its parts?, accepted for: *13th International Conference on Human-Computer Interaction*

Biographical Sketch

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.

Charlotte Wollermann

Institute of Communication Sciences
University of Bonn
Poppelsdorfer Allee 47
53115 Bonn
Germany

cwo@ifk.uni-bonn.de
www.ikp.uni-bonn.de/Members/cwo/homepage

1 Research Interests

My research interests lie in the experimental investigation of the role of **audiovisual prosody** for **focus production** and also **interpretation.** For these purposes natural and synthetic speech are used. In my project questions, methods and techniques from multiple disciplines are combined: **Speech production, perception** and **interpretation**, **audiovisual prosody**, **experimental pragmatics** and **phonetics**. However, basic research in human-human dialogue is important in order to develop believable spoken dialogue systems with deep linguistic performance, i.e. syntactic, semantic and pragmatic analysis.

1.1 Past, Current and Future Work

The purpose of my PhD research is to find out experimentally which role acoustic and/or visual cues play in the production, perception and interpretation of audio and/or visual speech.

In one experiment the role of emotion and attitude in audio speech perception was investigated. In Wollermann and Lasarcyk (2007) we measured the influence of (un)certainty on the perception of articulatory speech synthesis. For these purposes we varied acoustic cues which were identified in human-human dialogue as conveying uncertainty (e.g. Smith and Clark 1993; Swerts and Krahmer 2005). Results show that subjects are generally able to identify intended different levels of uncertainty.

Further, we tested to what extent audiovisual prosody effects the interpretation of focus produced by a Talking Head: We showed that the intensity of accent and eyebrow movement can influence the interpretation of a pragmatic focus (Fisseni et al. 2006). “Focus” refers to the concept as it is defined in Formal Semantics, e.g. Rooth (1992).

Moreover, the goal of a series of interpretation studies was to find out the contribution of uncertainty as paralinguistic expression to pragmatic focus interpretation. Here, natural stimuli were used. The results of Wollermann and Schröder (2008a, 2008b) suggest that prosodic cues effect the exhaustive interpretation of answers which follows pragmatic focus detection, but contextual factors (micro and macro context) do have a stronger effect.

Present and future work tests which audiovisual cues are relevant for pragmatic focus production. Our recent study (Wollermann, Schröder to appear) suggests that speakers use audiovisual cues of (un)certainty when uttering (non-)exhaustive answers. We also find empirical evidence for the co-expressivity of the audio and visual signal.

2 Future of Spoken Dialogue Research

I assume that current and future dialogue research deals increasingly with the modelling of emotion, attitude and personality. The modelling of emotion and attitude in synthetic speech has gained considerable importance in recent years as one aims to generate synthetic speech, which is natural and human-like as possible. Such systems were for instance developed by Burkhardt (2001) and Schröder (2004). But not only modelling of emotion in the acoustical channel has become more and more popular, also for the visual modality for creating virtual agents which are “able” to express their emotional state trough different channels. There has been much progress in developing Embodied Conversational Agents which are able to express personality and emotional states (e.g. Becker and Wachsmuth 2006). Nowadays, applications for these programs are usually restricted to special domains: REA (Cassell et al. 1999), e.g., functions as a real estate agent and MAX (Kopp et al. 2005) as a museum guide. A challenge for researchers in this field is the adoption of these programs to a larger variety of domains. For instance it would be necessary to investigate and evaluate in a very fine-grained manner which domains do actually benefit from a virtual agent.

However, in order to implement emotional cues it is firstly necessary to know how these cues are produced and perceived in natural speech. Most studies in the field of emotion research use acted emotional speech. In order to develop believable dialogue systems, it would be useful to use methods for exploring real emotional speech.

3 Suggestions for Discussion

I propose the following subjects for discussion:

- Speakers and hearers use different cues in communication for signalling and detecting uncertainty. Several studies show that uncertainty can be expressed by audio and/or visual synthetic speech. In this context the following question arises: Does the user of such a dialogue system expect that the machine has it's own meta-cognitive state?

- It has been shown that the McGurk effect (McGurk and MacDonald, 1976) is a universal phenomenon, but its strength depends on different factors, e.g. cultural origin. Which cultural aspects need to be considered for the development of talking heads given the fact that some facial expressions and gestures differ cross-culturally in their meaning?

- In natural language different phenomena of focus occur, e.g. semantic focus, pragmatic focus, contrastive focus, broad vs. narrow focus. How can we account on the complexity of the phenomenon and prosodic implications for the development of spoken dialogue systems?

References

C. Becker and I. Wachsmuth. 2006. Modeling Primary and Secondary Emotions for a Believable Communication Agent. In *Proc. of the First Workshop on Emotion and Computing*, pp. 31–34. University of Bremen.

F. Burkhardt. 2001. Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren. *PhD Thesis*, University of Berlin, Shaker Verlag.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjlmsson, and H. Yan. 1999. Embodiment in conversational interfaces: REA. In *Proc. of ACM CHI 99*, pp. 520–527.

B. Fisseni, F. Hülsken, B. Schröder, and C. Wollermann. 2006. Eyebrows, Accent and Focus detection. Paper presented at the *7th Szklarska Poreba Workshop*. March 03, 2006. Poland.

S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth. 2005. A conversational agent as museum guide – design and evaluation of a real-world application. In Panayiotopoulos et al. (eds.): *Intelligent Virtual Agents*, LNAI 3661, 329-343, Berlin: Springer-Verlag.

H. McGurk and J.W. MacDonald. 1976. Hearing lips and seeing voices. In *Nature*, vol. 264, pp. 746–748.

M. Rooth. 1992. A theory of focus interpretation. In*Natural Language Semantics*, 1, pp. 76–116.

M. Schröder. 2004. Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. *PhD thesis*, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

V. Smith and H. Clark. 1993. On the course of answering questions. In *Journal of Memory and Language*, vol. 32, pp. 25–38.

M. Swerts and E. Krahmer. 2005. Audiovisual prosody and feeling of knowing. In *Journal of Memory and Language*, vol. 53 (1), pp. 81–94.

C. Wollermann and E. Lasarcyk. 2007. Modeling and perceiving of (un)certainty in articulatory speech synthesis. In *Proc. of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 40–45. Bonn, Germany.

C. Wollermann and B. Schröder. 2008a. Does Uncertainty Effect the Case of Exhaustive Interpretation? In *Proc. of ISCA Tutorial and Research Workshop on Experimental Linguistics*, pp. 233–236. Athens, Greece.

C. Wollermann and B. Schröder. 2008b. Certainty, Context and Exhaustivity of Answers. Paper presented at *Speech and Face to Face communication*, October 27-29, 2008. Grenoble, France.

C. Wollermann. 2009. Effects of Exhaustivity and Uncertainty on Audiovisual Focus Production. To appear in *Proc. of AVSP'2009*. Norwich, UK.

Biographical Sketch

Charlotte Wollermann is a research assistant at the Institute of Communication Sciences at the University of Bonn. Her PhD thesis is supervised by Prof. Dr. Bernhard Schröder and apl. Prof. Dr. Ulrich Schade. She holds a M.A. in communication research and phonetics from the same university; she also time abroad at the *Dublin City University*. As a student assistant she developed and tested commercial question-answering-systems at *Q-Go online Marketing & Self Service; at Philips Laboratories* she tested multimodal home dialogue systems. The author also took a course in advanced voice user interface design at *VoiceObjects*.

YOUNG RESEARCHER'S ROUNDTABLE on SPOKEN DIALOGUE SYSTEMS 2009

September 13th-14th

Queen Mary
University of London