YOUNG RESEARCHERS' ROUNDTABLE

on Spoken Dialog Systems

1 September 2005
Lisbon, Portugal

www.yrrsds.org

Sponsors List:

International Speech Communication Association (ISCA)
Microsoft Research

Endorsed by:

International Speech Communication Association (ISCA)
Special Interest Group on Discourse and Dialogue (SIGdial)
Dialogs on Dialogs Student Reading Group

Contact Information:

Web page:	www.yrrsds.org
Email:	yrr-organizers@yrrsds.org
Fax:	+1-412-268-6298

Preface

The design and study of spoken dialog systems is a relatively young research field compared to other speech technologies such as recognition and synthesis. In recent years however, as these core technologies have improved, the field of spoken dialog systems has been generating increased interest both in the research community and in the industry. While most of the early work originated from the artificial intelligence community and addressed high-level issues such as discourse planning, the development and deployment of actual usable systems has led to the emergence of a wide range of new issues such as error handling in dialog, multimodal integration, or rapid system development. At the same time, researchers from a variety of disciplines including speech and language technologies, robotics, and human-computer interaction have started to bring their unique skills and backgrounds to bear on these issues.

Unfortunately, while this richness and variety of interests constitute a definite strength, they can also be a source of isolation and discouragement, particularly for newcomers to the field. Many young researchers in spoken dialog systems work within small research groups and find it difficult to share their ideas with peers having similar or complementary interests. While annual conferences such as SIGdial and Interspeech provide excellent opportunities for young researchers to present their own work and hear about work that is done in similar areas, there have been few opportunities to date for more intensive discussion and thought about interesting and challenging questions in the field today.

We believe that both young researchers and the field itself would benefit greatly from a better communication across institutions and disciplines. By working together, getting peer-level feedback on their research, and engaging in brainstorming sessions, researchers could identify the questions that are most relevant to the overall problem of spoken human-machine communication, and come up with fresh ideas to answer these questions.

With these goals in mind, in 2002 we started Dialogs on Dialogs (www.cs.cmu.edu/~dod), an international student reading group focused on the area of Spoken Dialog Systems/Conversational Agents. The group is based at Carnegie Mellon University and involves participants from other universities through teleconferencing. Our bi-weekly meetings provide a setting in which we can present our own research and obtain feedback from others who are at our level and who are working on similar problems. The Young Researchers' Roundtable on Spoken Dialog Systems workshop was conceived as an extension of these activities. The two main objectives of the proposed workshop are to foster creative and actionable thinking about current issues in spoken dialog systems research, and to create a network of young researchers working in spoken dialog systems.

For the inaugural Young Researchers' Roundtable on Spoken Dialog Systems workshop, we are pleased to have received 26 submissions from young researchers in ten countries on four continents. The quality and diversity of the submissions, with topics ranging from statistical methods for dialog management to rapid development, evaluation, robust speech recognition and language understanding, etc. promises thought-provoking and useful discussions. We hope that the workshop participants enjoy this event and that the workshop can continue to be held on a regular basis.

We wish to express our sincerest thanks to our sponsors: the International Speech Communication Association (ISCA) and Microsoft Research for their generous support of this event. We would also like to thank SIGdial for endorsing this workshop. We also wish to thank Alex Rudnicky, Alan Black, David Traum, Rolf Carlson, Diamantino Caseiro, Antonio Serralheiro and Isabel Trancoso, as well as the other members of the advisory committee for their advice, help, and ideas, without which we would have had a much more difficult time organizing the workshop.

The YRRSDS-2005 organizing committee,
1 September 2005, Lisbon, Portugal.

Organizing Committee:

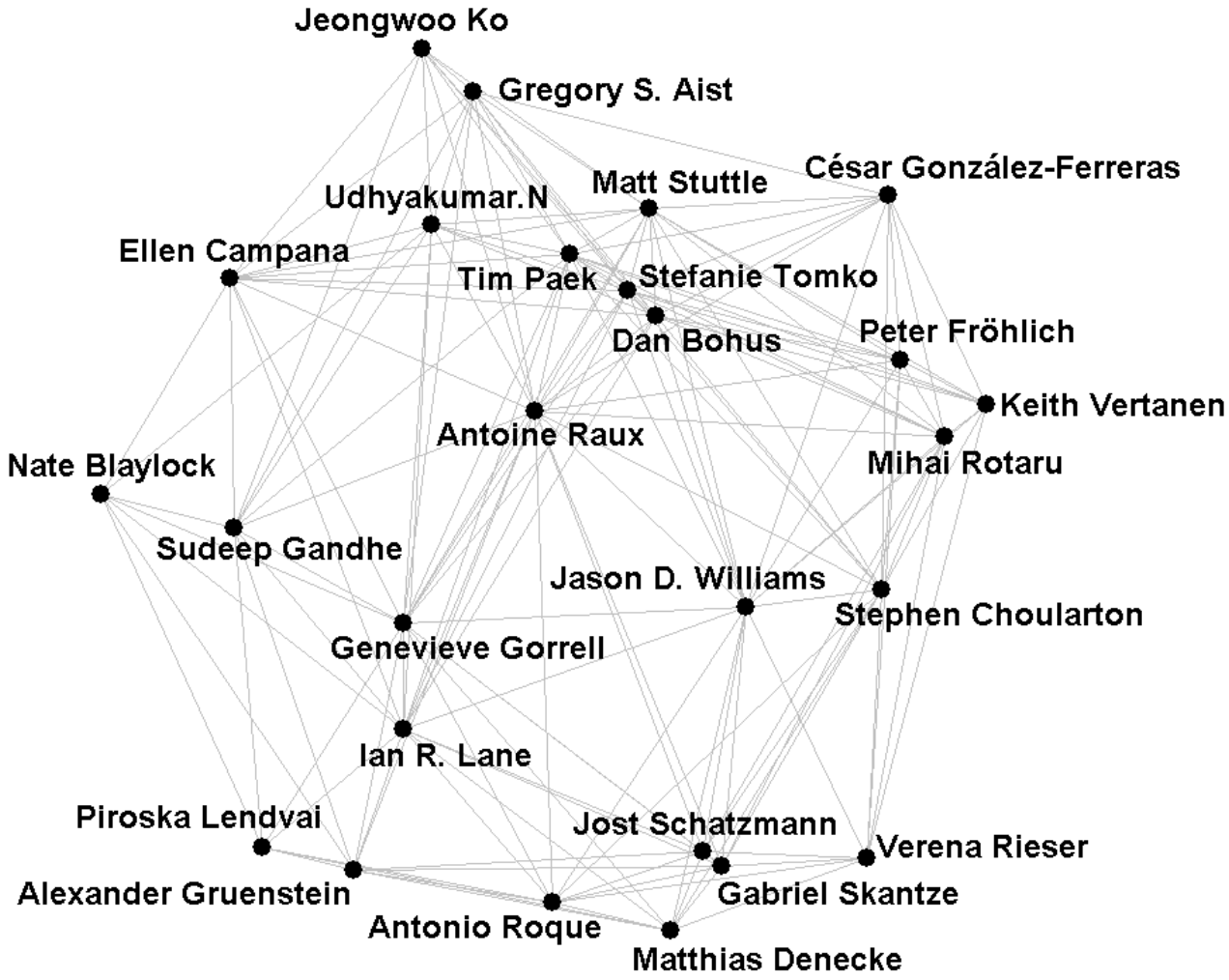
Satanjeev Banerjee, Carnegie Mellon University
Dan Bohus, Carnegie Mellon University
Ellen Campana, University of Rochester
Stephen Choularton, Macquarie University, Australia
Antoine Raux, Carnegie Mellon University
Stefanie Tomko, Carnegie Mellon University
Jason Williams, Cambridge University, UK

Advisory Committee:

Gregory Aist, University of Rochester, USA
Alan Black, Carnegie Mellon University, USA
Chih-yu Chao, Carnegie Mellon University, USA
Grace Chung, CNRI, USA
Robert Dale, Macquarie University, Australia
Matthias Denecke, NTT, Japan
Genevieve Gorrell, Linköping University, Sweden
Alex Gruenstein, MIT, USA
Thomas Harris, Carnegie Mellon University, USA
Kristiina Jokinen, University of Helsinki, Finland
Michael Kipp, DFKI, Germany
Jeongwoo Ko, Carnegie Mellon University, USA
Kazunori Komatani, Kyoto University, Japan
Staffan Larsson, Göteborg University, Sweden
Akinobu Lee, Nara Institute of Science and Technology, Japan
Oliver Lemon, Edinburgh University, UK
Masafumi Nishida, Chiba University, Japan
Tim Paek, Microsoft Research, USA
Alexander Rudnicky, Carnegie Mellon University, USA
Gabriel Skantze, KTH, Sweden
Kishore Sunkeshwari Prahallad, International Institute of Information Technology, India
Mark Swerts, Tilburg University, Netherlands
Steve Young, Cambridge University, UK

Workshop Program

8:00 -	**9:00**	**Check-in and on-site registration**
9:00 -	**10:00**	**Welcome and introductions**
10:00 -	10:30	Morning coffee break
10:30 -	**12:00**	**Morning roundtable sessions**
		Room (Sala) A: Evaluation
		Room (Sala) C: Empirical / Statistical Methods
12:00 -	**12:30**	**Morning sessions' summary**
12:30 -	2:00	Lunch
2:00 -	**3:30**	**Afternoon roundtable sessions**
		Room (Sala) A: Applications
		Room (Sala) C: Automatic Speech Recognition and Language Understanding
3:30 -	**4:00**	**Afternoon sessions' summary**
4:00 -	**4:30**	**Wrap-up**
4:30 -		Afternoon coffee social


Jeongwoo Ko
Gregory S. Aist
César González-Ferreras
Matt Stuttle
Udhyakumar.N
Ellen Campana
Tim Paek
Stefanie Tomko
Dan Bohus
Peter Fröhlich
Keith Vertanen
Antoine Raux
Mihai Rotaru
Nate Blaylock
Sudeep Gandhe
Jason D. Williams
Stephen Choularton
Genevieve Gorrell
Ian R. Lane
Piroska Lendvai
Jost Schatzmann
Verena Rieser
Alexander Gruenstein
Gabriel Skantze
Antonio Roque
Matthias Denecke

Contents

Position Papers

Gregory S. Aist

University of Rochester
Computer Science
RC 270226
Rochester NY 14607 USA

gaist@cs.rochester.edu
www.gregoryaist.com

1 Research Interests

My interests are in language, computation, and learning. The first main thread is **interactive language learning,** particularly people learning language skills from human-computer dialogs and machines improving language processing abilities from interactions with users. Next is **multimedia documents,** with text, speech, graphics, animation, photos, video, and/or physical objects & acts. The third is **empirical methods for dialog systems** R&D, especially experimental methods using direct interaction with a dialog system.

2 Past, Current and Future Work

My work has been in the context of four primary dialog systems: Project LISTEN's Reading Tutor, the CLARISSA astronaut assistant, the Purchasing Assistant from the CALO Project, and a continuous understanding testbed tentatively titled CAFE/FruitCarts.

2.1 Project LISTEN's Reading Tutor: A dialog system to help children learn to read

Project LISTEN's goal is to develop a Reading Tutor that uses automated speech recognition to help children learn to read (J. Mostow, director). In five years as the principal graduate student on Project LISTEN, I worked on nearly every aspect of developing, evaluating and improving an intelligent tutoring system that uses spoken dialog to interact with its users. One of the first areas I worked on was extending dialogue system turn-taking – I developed a novel architecture that allowed not only alternating turns and user barge-in, but also computer-generated backchanneling and content-driven interruption (Aist 1998). Subsequently I worked on mixed-initiative task choice (which story to read, in this case; see Aist and Mostow in press), and on augmentation of the document being read with an automatically generated glossary (Aist 2001). We compared the resulting version of the Reading Tutor to classroom instruction and to one-on-one human-guided oral reading. The result: Second graders did about the same on word comprehension in all three conditions. However, third graders who read with the 1999 Reading Tutor, modified as described, performed statistically significantly better than other third graders in a classroom control on word comprehension gains – and even comparably with other third graders who read one-on-one with human tutors.

2.2 CLARISSA: A dialog system to help astronauts perform tasks on the International Space Station

Space flight is a challenging and labor-intensive endeavour. Astronaut time while in orbit is especially valuable given costs such as training and orbital launches. Any scientific or technological advance that would help astronauts work more efficiently and effectively would be hugely beneficial. I began in December 2001 to analyze the day-to-day tasks of astronauts. I identified several major areas where spoken dialogue systems might help with training astronauts or helping them perform their tasks. In early 2002, I collaborated with other Ames researchers and Johnson Space Center (JSC) personnel and selected as an area of further development astronaut support for procedural tasks on the International Space Station. We next conducted initial data collection on human-human conversation during procedural tasks based on the Earth Observation program, a long-running NASA endeavor where astronauts take pictures of the Earth with a digital camera. We used these user data to develop an initial voice-in, voice-out prototype that walked the user through the steps of how to unpack and use a digital camera. We demonstrated this prototype at NASA sites and at international conferences (e.g. Aist, Dowding, et al. 2002). Concurrently, I led a team of developers to build a follow-on prototype of a procedural assistant for astronauts that ran over actual Space Station procedures represented as XML. By December 2002 we had completed the Checklist system version 2 (Aist et al. 2003). Five astronauts at JSC used the system to run through a water sampling procedure and their feedback was used to refine the system further. In early 2003 we sought a memorable name using "ISS", and subsequently renamed the system CLARISSA - the Checklist and Robotics Intelligent Space Station Assistant. During the spring of 2003, I led the development and integration effort for the third prototype of CLARISSA. Key functionality here included being able to tell when the user was talking to the system (vs. to another astronaut or to Mission Control); navigating steps of the procedure; handling spoken corrections; and re-

sponding to requests for extra information (Aist et al. 2003). CLARISSA continued to receive strong support from research and management at Ames, and from the astronaut corps and training personnel at JSC; this institutional buy-in was critical to its early survival and later success (Aist et al. 2004). CLARISSA was subsequently further refined, tested, and hardened; more procedures were added to its database and it was delivered in May 2004 to the Johnson Space Center for use in astronaut training. CLARISSA was launched to the International Space Station in December 2004. Initial use is anticipated during the spring of 2005. CLARISSA is thus the first dialog system in space.

2.3 Intelligent Purchasing Assistant: A CALO subproject to develop a spoken dialog system for helping users with purchasing, such as constructing a specification

The overall goal of the CALO project is to build an intelligent assistant that learns, over time, and improves itself to help you better. The Purchasing Assistant focuses on the subtask of helping the user formulate a specification, select a possible option, and make a purchase. The initial gomain was computer purchases (e.g. Laptops) Current plans include extending to projection equipment and online book sales (e.g. Amazon.com).

2.4 CAFE: Continuous understanding by machines; and a testbed domain of making and moving objects - "Fruit Carts"

One standard way of building dialog systems is to feed information forward from the speech recognizer, to the parser, to interpretation, and so forth. It is clear from human-human conversational behavior (such as acknowledgements, interruption, and so forth) that human conversational understanding transcends a strict pipeline. The Rochester CAFE architecture aims to build an asynchronous, agent-based framework for dialogue systems that allows multiple constraints to be brought to bear – in realtime – on spoken language as it unfolds (Blaylock, Allen, and Ferguson 2002). Within this context I have been contributing to the development of a testbed domain for continuous understanding - "Fruit Carts". In this setup, people interact with a conversational partner to select, modify, and place objects on a map. (Currently the objects are either (abstract) "cart" shapes or "fruit"; thus the name "fruit carts".) Since in this domain properties such as position and angle of rotation can vary continuously, the language used includes not only specifications such as "rotate it thirty degrees to the left" but also more interactively style language such as "rotate it left a bit... a bit more... oops back – that's good stop right there." In these situations the ability to interpret while the user is still speaking could help move along the dialog.

2.5 Empirical Methods

Dialog system learns from previous interactions – for example, using automatically transcribed and selected utterances from human-computer dialog to assist with acoustic model training. For example, we trained a speech recognizer on automatically collected, automatically transcribed children's speech. This improved its accuracy on a variety of different tasks on test data collected under various acoustic conditions (Aist et al. 1998).
Embedded experiments as described in Mostow et al. (2001), and **embedded training** as described in Hockey et al. (2003).
System-user-expert dialogs (SUE) help you elicit what added functionality a system needs to have (Aist 2004). For example, adding human-supplied emotional scaffolding to computer tutoring, with the goal of encouraging greater student persistance (Aist et al. 2002).
Other empirical methods are in more embryonic stages:
Extracting Similar Speech from Dialog helps to find alternate ways for the system to express a thought (Aist, Allen, and Galescu 2004).
Comprehensive Path Analysis of dialog systems helps to debug, test, and refine dialog systems.

3 Challenges in Spoken Dialog Systems Research

I would like to separate the challenges into, roughly speaking, old challenges with some new twists, and relatively new challenges.
One familiar challenge is speech recognition accuracy and speed. There has definitely been tremendous improvement in both accuracy and speed of automatic speech recognition. The new twist is conversational speech. Dialog systems that are to be used outside of domains which lend themselves to step-by-step actions (& turns) place new demands on both accuracy and speed.
Another familiar challenge is domain specificity. Many systems are designed with a domain-independent core, and a component that can be loaded at runtime – for example, knowledge about how to follow hyperlinks vs. loading a particular set of webpages. The new twist here is functional reuse (not just lexical items and a new knowledge base.) For example, if a dialogue system knows that given a plan you can (say) "break that down", it should be straightforward to also "break down" a diagram. However, such use may not carry over directly across domains. Perhaps it should.
Finally, a relatively new challenge is the issue of training – how can people best learn to interact with dialog systems? Should training precede use, be embedded in it, come as reviews; or should we as designers seek to eliminate it altogether? What about cases where a system that requires much training might be more efficient for experts? As more dialog systems are deployed in use, this issue will become increasingly important.

References

Aist, G.S. and Mostow, J. In press. Faster, Better Task Choice in a Reading Tutor that Listens. To appear in Philippe DeCloque and Melissa Holland (Editors), Speech Technology for Language Learning. The Netherlands: Swets & Zeitlinger Publishers.

Aist, G. S. 2004. Three-way system-user-expert interactions help you expand the capabilities of an existing spoken dialogue system. 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004.

Aist, G., Allen, J., and Galescu, L. 2004. Expanding the linguistic coverage of a spoken dialogue system by mining human-human dialogue for new sentences with familiar meanings. Member Abstract, 26th Annual Meeting of the Cognitive Science Society, Chicago, August 5-7, 2004.

Aist, G., Bohus, D., Boven, B., Campana, E., Early, S., and Phan, S. 2004. Initial development of a voice-activated astronaut assistant for procedural tasks: From need to concept to prototype. *Journal of Interactive Instruction Development* **16**(3): 32-36.

Aist, G., Dowding, J., Hockey, B.A., Rayner, M., Hieronymus, J., Bohus, D., Boven, B., Blaylock, N., Campana, E., Early, S., Gorrell, G., and Phan, S. Talking through procedures: An intelligent Space Station procedure assistant. 2003. European Association for Computational Linguistics (EACL) 2003 meeting, Software Demonstration, Budapest, Hungary, April 2003.

Aist, G., Dowding, J., Hockey, B.A., and Hieronymus, J. 2002. A Demonstration of a Spoken Dialogue Interface to an Intelligent Procedure Assistant for Astronaut Training and Support. Association for Computational Linguistics (ACL) 2002 meeting, Demo Session. Philadelphia, July 7-12, 2002.

Aist, G., Kort, B., Reilly, R., Mostow, J., and Picard, R. 2002. Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens Increases Student Persistence. Poster presented at Intelligent Tutoring Systems (ITS) 2002 conference, Biarritz, France, June 5-7, 2002.

Aist, G. 2002. Helping Children Learn Vocabulary during Computer-Assisted Oral Reading. *Educational Technology and Society* **5**(2). http://ifets.ieee.org/periodical/vol_2_2002/aist.html

Aist, G. 2001. Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education.* **12**, 212-231. http://cbl.leeds.ac.uk/ijaied/abstracts/Vol_12/aist.html

Aist, G. 1998. Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption. International Conference on Spoken Language Processing (ICSLP) 1998, Sydney, Australia, Nov. 30 - Dec. 4, 1998. Paper 928.

Aist, G., Chan, P., Huang, X.D., Jiang, L., Kennedy, R., Latimer, D., Mostow, J., and Yeung, C. 1998. How effective is unsupervised data collection for children's speech recognition? International Conference on Spoken Language Processing (ICSLP) 1998, Sydney, Australia, Nov. 30 - Dec. 4, 1998. Paper 929.

Blaylock, N., Allen, J., and Ferguson, G. Synchronization in an asynchronous agent-based architecture for dialogue systems. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialog*, Philadelphia, July 2002.

Hockey, B.A., Lemon, O., Campana, E., Hiatt, L., **Aist, G.**, Hieronymus, J., and Dowding, J. Targeted Help. 2003. European Association for Computational Linguistics (EACL) 2003, Budapest, Hungary, April 2003.

Mostow, J., **Aist, G.,**Bey, J., Burkhead, P., Cuneo, A., Rossbach, S., Tobin, B., Valeri, J., and Wilson, S. 2001. A hands-on demonstration of Project LISTEN's Reading Tutor and its embedded experiments. Demonstration at Language Technologies 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics. Pittsburgh, Pennsylvania, June 2-7, 2001.

Biographical Sketch

Gregory S. Aist is currently a Research Associate in the Computer Science Department at the University of Rochester. He previously held positions at the Research Institute for Advanced Computer Science at NASA Ames; the Media Lab at MIT; Carnegie Mellon University, and Microsoft Research. He holds a Ph.D. and an M.S. from Carnegie Mellon, where he was a National Science Foundation Graduate Fellow, and a B.A. from Messiah College. Outside of computer science and linguistics, his interests include card games, ballroom dancing, investment analysis, jewelry-making, rollercoasters, and sports.

Nate Blaylock

Saarland University
Dept. of Computational Linguistics
66125 Saarbrücken
Germany

blaylock@coli.uni-sb.de
www.coli.uni-sb.de/~blaylock

1 Research Interests

My research interests lie at the intersection of three research areas: **dialogue systems**, **autonomous agents**, and **planning**. I am interested in building **conversational agents**, which I see as the union of autonomous agents and a dialogue systems. Towards that end, I am most focused on work in **dialogue modeling and management** as well as **intention recognition** for language understanding.

2 Past, Current and Future Work

2.1 Past

Work for my dissertation has focused on taking steps to bridge the gap between autonomous agents and dialogue systems. It introduces an agent-based dialogue model which represents dialogue as *collaborative problem solving* (CPS) between agents (Allen et al., 2002; Blaylock et al., 2003). Whereas plan-based dialogue models typically represent dialogue as the creation (or execution) of a joint plan, the CPS dialogue model attempts to include the range of agent activities, including goal evaluation and selection, planning and execution (possibly interleaved), monitoring and replanning, and so forth. This allows the model to represent a much wider range of dialogue behavior.

At Rochester, I was also involved in development of the TRIPS MedAdvisor dialogue system (Ferguson et al., 2002), which utilized the CPS model. The MedAdvisor system also introduced a novel asynchronous dialogue architecture, and I helped in providing synchronization points in the architecture, which were closely related to grounding (Blaylock et al., 2002).

Additionally, my dissertation focused on supporting language understanding for agent-based dialogue models. Plan-based dialogue models are complicated, and require *intention recognition* in order to do language understanding. The wider range of behavior supported by agent-based models actually makes this a more difficult problem. A common complaint about intention recognizers is that they are not scalable, as they are based on intractable plan recognition algorithms. I have worked on applying statistical, corpus-based machine learning techniques from NLP to plan recognition in order to make plan recognizers more accurate and tractable (Blaylock and Allen, 2003; Blaylock and Allen, 2004; Blaylock and Allen, 2005).

2.2 Present

I am currently involved in the European project TALK[1], which is focused on flexible, portable and adaptive multimodal and multilingual dialogue systems for in-car and in-home use.

My current research focuses on integrating the CPS model from my thesis into an information state update (ISU) based dialogue model. The CPS model is not a full-fledged dialogue model in itself, as it does not cover linguistic phenomena such as turn taking, grounding, and discourse obligations. I am investigating possible overlaps with Traum and Hinkelman's 4-level dialogue model (Traum and Hinkelman, 1992), as well as compatibilities with Grosz and Sidner's tripartite model of dialogue (Grosz and Sidner, 1986), in order to produce a full dialogue model.

In addition, I am also implementing an agent-based dialogue manager which reasons with the CPS model within the ISU paradigm. This should make dialogue management more flexible and portable by allowing the dialogue manager to use domain-independent update rules which preclude the need for hand-coded dialogue plans for the domain. Instead, the dialogue manager will reason directly with the task models themselves. The dialogue manager is the key point where agent and dialogue technology come together. It needs not only to reason about beliefs, intentions, goals, plans and actions (like an autonomous agent), but also about how to collaborate (communicate) with another agent to accomplish those goals.

Lastly, I am involved in work on multimodal generation, especially in looking at how the very linguistically poor communicative intentions the CPS dialogue manager produces can be generated into multimodal content, especially into language.

[1] http://www.talk-project.org

2.3 Future

In the future, I plan to continue working on agent-based dialogue. One particularly large question remaining is that of language understanding in the CPS model. As mentioned above, this question goes back to work on intention recognition, which needs to be extended to cover the range of problem solving behavior, and not just plan creation or execution. I believe this could also provide a platform for tying together work on intention recognition with more traditional work on pragmatics.

3 Challenges in Spoken Dialog Systems Research

I believe there are a number of interesting challenges still to be met for dialogue systems. I mention only a few:

Handling a broad range of dialogue types Much recent work in dialogue systems has targeted a small set of dialogue types, including database search (e.g., flight reservations, movie information) and simple task control (e.g., in-home device control, procedure reading), and has seen a lot of success in these areas. There are, however, many dialogue types which we aren't yet as successful in, including planning (e.g., kitchen design, task scheduling). In addition, I believe that mixed-initiative dialogue still remains a challenge.

Portability of dialogue systems For dialogue systems to be commercially viable, we need systems which can be easily and cheaply ported to new domains. This includes issues like domain-independent components and the coding of linguistic and task models for new domains. Although there is a lot of work happening in this area, I believe it still remains a challenge.

Incremental language processing Perhaps the biggest challenge we have is that processing language incrementally, or at a finer granularity. Most dialogue systems (and underlying technology — from parsing to dialogue management to generation) process language more or less at an utterance-level granularity. In order to make dialogue more natural, we want to support (in a general way) phenomena like turn taking and backchanneling. For this, we need to have information at much more fine-grained level, like at a single word or morpheme level. Although these issues are heavily studied in psycholinguistics, not enough attention has probably been paid to them in NLP.

References

James Allen, Nate Blaylock, and George Ferguson. 2002. A problem-solving model for collaborative agents. In *First International Joint Conference on Autonomous Agents and Multiagent Systems*.

Nate Blaylock and James Allen. 2003. Corpus-based, statistical goal recognition. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*.

Nate Blaylock and James Allen. 2004. Statistical goal parameter recognition. In *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS'04)*.

Nate Blaylock and James Allen. 2005. Generating artificial corpora for plan recognition. In *International Conference on User Modeling (UM'05)*, Edinburgh, July. To appear.

Nate Blaylock, James Allen, and George Ferguson. 2002. Synchronization in an asynchronous agent-based architecture for dialogue systems. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialog*.

Nate Blaylock, James Allen, and George Ferguson. 2003. Managing communicative intentions with collaborative problem solving. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*. Kluwer, Dordrecht.

G. Ferguson, J. Allen, N. Blaylock, D. Byron, N. Chambers, M. Dzikovska, L. Galescu, X. Shen, R. Swier, and M. Swift. 2002. The Medication Advisor project: Preliminary report. Technical Report 776, University of Rochester, Department of Computer Science.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

David R. Traum and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599.

Biographical Sketch

Nate Blaylock was born and raised in Utah. In 1999, he received a BS in Computer Science and a BA in Linguistics from Brigham Young University. He has since been at the University of Rochester working with advisor James Allen, where he received an MS in Computer Science in 2001 and expects to complete his PhD in Summer 2005. Since Spring 2004, he has been employed as a research associate at Saarland University in Saarbrücken, Germany, working on the European project TALK.

Dan Bohus

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15217

dbohus+@cs.cmu.edu
http://www.cs.cmu.edu/~dbohus/

1 Research Interests

Currently, my research interests are focused on **dialog management** and **error detection and recovery** in spoken dialog systems. More generally, I am also interested in **multi-modal systems**, **embodied conversational agents**, and **spoken language interaction in intelligent environments**.

2 Past, Current and Future Work

A persistent and important problem in today's spoken language interfaces is their brittleness when faced with understanding errors. This problem appears across all domains and interaction types, and stems primarily from the inherent unreliability of the speech recognition process. The recognition difficulties are further exacerbated by the conditions under which these systems typically operate: spontaneous speech, large vocabularies and user populations, and large variability in input line quality. In these settings, average word-error-rates of 20-30% (and up to 50% for non-native speakers) are quite common. Two pathways towards increased robustness can be easily envisioned. One is to improve the accuracy of the speech recognition process. The second is to assume that inputs will be in fact noisy and create mechanisms for detecting and gracefully handling potential errors at the conversation level.

My thesis research (Bohus, 2004) aims to address this problem and is centered on the second approach. The high-level goal is to develop an adaptive, task-independent and scalable framework for error handling in task-oriented spoken dialog systems. I believe that three competencies are required for increased robustness: (1) the ability to accurately detect and diagnose errors; (2) a rich repertoire of error recovery strategies; (3) the ability to select the most appropriate error recovery strategy in any given situation.

2.1 Dialog Management

As a prerequisite, over the last few years, my efforts have been focused on the development of RavenClaw (Bohus and Rudnicky, 2003), a dialog management framework for complex, task-oriented domains. The framework enforces a clear separation between the domain-specific and the domain-independent aspects of dialog control. The domain-specific aspects are captured by a Dialog Task Specification, essentially a hierarchical plan for the interaction, provided by the system author. A fixed, domain-independent Dialog Engine manages the conversation by executing the given Dialog Task Specification. In the process, the Dialog Engine automatically contributes a set of domain-independent conversational behaviors such as error recovery, timing and turn-taking, and support for "universal" commands like help, repeat, cancel, suspend, quit, etc.

To date, several systems spanning different domains and interaction types (e.g. information access, browsing and guidance through procedural tasks, command-and-control, etc) have been successfully built and deployed using this framework. Together with these systems, RavenClaw provides the necessary infrastructure for my current work in error handling. More generally, RavenClaw provides a robust basis for research in various aspects of dialog management such as multi-participant dialog, timing and turn-taking, dynamic generation of dialog plans, learning at the task level etc.

2.2 Error Handling

My work in error handling gravitates around the three issues outlined at the beginning of this section: (1) error detection, (2) error recovery strategies, and (3) error handling policy.

With respect to **error detection**, I have previously worked on developing a semantic confidence annotation scheme which integrates information from multiple knowledge sources in a spoken dialog system (Carpenter et al, 2001). Furthermore, I have proposed a data-driven methodology for assessing the costs of confidence annotation errors in a spoken dialog system (Bohus and Rudnicky, 2001), and showed how these costs can be used to further tune the utterance rejection process (Bohus and Rudnicky, 2005c). While confidence scores can provide an initial assessment for the reliability of the information obtained from the user, ideally a system should leverage information available in subsequent user turns to update its beliefs and improve their accuracy. Currently, I am developing data-driven models for this **belief updating** problem (Bohus and Rudnicky, 2005a). The proposed approach bridges

previous work on confidence annotation and correction detection into a unified framework which allows spoken dialog systems to more accurately track their beliefs through time. In a related project (in collaboration with Antoine Raux), I am investigating the **transferability of confidence annotators across domains**. The classical data-driven approach for building semantic confidence annotators requires a corpus of manually labeled in-domain data, which is costly and hard to obtain in early system development stages. To address this issue, we are developing unsupervised techniques for migrating existing confidence annotators to new domains.

With respect to **error recovery strategies**, I have recently performed an empirical investigation of non-understanding errors and 10 corresponding recovery strategies (e.g. asking the user to repeat, to rephrase, notifying the user that a non-understanding has occurred, providing various levels of help, etc). More precisely, the questions under scrutiny were: what are the main causes of non-understandings? What is their impact on overall performance? What is the relative performance of various recovery strategies? Can that performance be improved by making smarter choices about which strategy to use at runtime? If so, can we learn how to make these smarter choices? The results of this study are presented in detail in (Bohus and Rudnicky, 2005b).

Based on the lessons learned in this user study, I am currently refining the set of error handling strategies. In the near future, I plan to perform a new study focused this time on the third aspect of error handling: **learning a policy** for engaging the error recovery strategies.

3 Challenges in Spoken Dialog Systems Research

I believe that one of the major problems in today's spoken language interfaces is their brittleness when faced with recognition errors. As recognition accuracy improves, interests also shift towards more complex systems, and the demands are not likely to be met in the near future. I think the solution lies in developing systems which can gracefully handle communication errors through interaction. Some of the important questions I see in this area are: how does a system "know that it doesn't know?" How can we build more accurate system beliefs? What set of strategies can be used to set a conversation back on track? What techniques can be used to learn optimal error recovery behaviors on-line, from detected error segments, and how can we build systems that adapt and improve over time?

Another important current challenge lies in finding ways to reduce the amount of human effort and expertise, and the amount of fine-tuning that is required with the development of each new spoken dialog system. I believe that aspects such as reusability, adaptability and unsupervised learning are all part of the game, but more work is required in each of these areas before a satisfactory solution is reached.

Last but not least, I think that a shift towards more complex domains (e.g. personal assistants, tutoring, unstructured information access, etc) will bring forth a set of new challenges, such as "deeper" yet robust language understanding, more sophisticated dialog models, integration with more sophisticated reasoning / back-end abilities, better interaction skills such as timing and turn-taking, multi-participant dialog, etc.

References

Dan Bohus and Alex Rudnicky. 2005a. *Constructing Accurate Beliefs in Spoken Dialog Systems*, submitted to ASRU-2005, Cancun, Mexico.

Dan Bohus and Alex Rudnicky. 2005b. *Sorry I didn't Catch That: An Investigation of Non-understanding Errors and Recovery Strategies*, SIGdial-2005, Lisbon, Portugal.

Dan Bohus and Alex Rudnicky. 2005c. *A Principled Approach for Rejection Threshold Optimization in Spoken Dialog Systems*, Interspeech-2005, Lisbon, Portugal.

Dan Bohus. 2004. *Error Awareness and Recovery in Task-Oriented Spoken Dialog Systems*, Ph.D. Thesis Proposal, Carnegie Mellon University.

Dan Bohus and Alex Rudnicky. 2003. *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda*, Proc. of Eurospeech'2003, Geneva, Switzerland; www.cs.cmu.edu/~dbohus/RavenClaw/

Dan Bohus and Alex Rudnicky. 2001. *Modeling the Cost of Misunderstanding in the CMU Communicator Dialog System*, Proc. of ASRU-2001, Madonna di Campiglio, Italy.

Carpenter, P., Jin, C., Wilson, D., Zhang, R., Bohus, D., Rudnicky, A., 2001 – *Is This Conversation on Track?*, in Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001.

Gabriel Skantze. 2003. *Exploring Human Error Handling Strategies*, Proceedings ISCA Workshop on Error Handling in Spoken Dialogue Systems, Chateau D'Oex, Switzerland.

Biographical Sketch

I am currently a Ph.D. student working under the supervision of Dr. Alex Rudnicky and Dr. Roni Rosenfeld in the Computer Science Department at Carnegie Mellon University. I was born in Romania and I have obtained my undergraduate Computer Science degree from "Politechnica" University of Timisoara. Outside academia, I am interested in non-fiction books, coffee, science (especially modern physics, social sciences and evolutionary theory), geeky-cool things, yoga, existentialism, more books, skiing, as well as ice-cream, tiramisu and dark comedies.

Ellen Campana

University of Rochester
Department of Brain and Cognitive Sciences
Meliora Hall, Box 270268
University of Rochester
Rochester, NY 14627

ecampana@bcs.rochester.edu
http://www.ellencampana.com

1 Research Interests

I have multiple research interests that lie at the intersection points between the fields of **human factors**, **psycholinguistics**, and **dialog systems**. Specifically, I am interested in how **dialog system usability** could be improved, particularly for users who are engaged in **multitasking,** or who have to do complicated **problem-solving**, and/or **reasoning** about the domain. I am convinced that the **methods** and findings from **cognitive psychology** can be applied to usability in dialog systems -- for improving systems directly, and for evaluating the consequences of specific **design decisions**.

I have a special interest in **referring expressions** because they are fundamental to **natural human language use in context**. As people engage in **collaborative problem-solving**, much of what they say refers to specific entities and actions, real or hypothetical. Thus, much of the "work" of language understanding in **task-focused dialog** is identifying the specific entities and actions that a speaker intended to refer to. Therefore, dialog system design decisions related to **reference resolution** and **reference production** can be expected to have large consequences on system usability.

2 Past, Current and Future Work

My research can be broadly categorized into the following 3 categories: human-human communication, improving dialog systems, and evaluating dialog systems. Due to space constraints I can only give the broadest of summaries for each, but please feel free to contact me if you would like more information about any of my projects.

2.1 Investigating Human-human Communication

In most of the projects I've worked on so far the goal has been to investigate how humans naturally communicate; how they produce and understand natural language. With one group of colleagues I investigated human comprehension of definite referring expressions in unscripted problem-solving tasks (Campana et al. 2002). With another group of colleagues I investigated human generation of instructions directed toward robot assistants developed for robot-assisted missions on Mars (Dowding, 2004). I have also worked on some projects that are only distantly related to dialog systems: fMRI during comprehension of noun phrases (definite and indefinite), comprehension of instrument verbs, and comprehension and production of hand gestures (by college students, typical children and children with high-functioning autism).

2.2 Improving Dialog Systems Directly

In two projects that I've worked on so far the goal has been to improve dialog systems directly by drawing on findings from cognitive psychology. In the first project, my colleagues and I integrated eye-movement data into a dialog system developed for directing robots on the International Space Station. This work stands out from other work of its kind because it focuses on making use of the eye-movements people naturally make while talking about objects that are present in their environment (Campana et al. 2001). The second project, which I am only beginning to be involved in, has the goal of making a system that is capable of incremental language understanding and generation. These capabilities will hopefully allow the system to interact more naturally with users. Developing this system will involve designing a new architecture, and creating a domain in which to explore and test its unique capabilities (Aist, 2004).

2.3 Evaluating Dialog Systems

In three projects that I've worked on so far the goal has been to improve dialog system evaluation using methods from cognitive psychology. In the first project, my colleagues and I used experimental design principles from experimental psychology to compare users' performance on the WITAS system augmented with targeted help capabilities to the users' performance on the WITAS system without this capability (Hockey et al, 2003). In the second project, my colleagues and I investigated the possibility of using human eye-movements during spoken language comprehension as a fine-grained evaluation metric for speech synthesis (Swift et al. 2002). In the third project, my colleagues and I extended the dual-task paradigm, a classic tool of cognitive psychology, to dialog system evaluation (Campana et al. 2004). My Ph.D. dissertation will build on this line of research: I plan to use the dual-task paradigm to

evaluate users' cognitive load as they interact with systems that differ with respect to how "natural" and/or "consistent" they are. My specific goal is to examine the limits of human learnability to find a set of empirically determined design principles for spoken dialog systems. The dissertation itself will focus on definite referring expressions, but my hope is that the method will generalize to other important domains.

3 Challenges in Spoken Dialog Systems Research

There seems to be a growing tension in spoken dialog systems research: some researchers argue that in order to be usable, systems need to more closely approximate human-human communication while other researchers argue that standardization is the key to improved usability. Some of the most important current challenges in spoken dialogs systems research relate to this tension. I think in the next few years there will be increasing pressure for researchers to take a stand on some of the following questions:

1) Should human-human communication be the gold standard for dialog system development? If so, what aspects are truly fundamental (thus necessary to implement), and which might be superfluous? If not, what else should be the gold standard, and is there a principled way to decide how to make design decisions?

2) Are people actually capable of learning how to use the systems that are being developed? How much training is necessary? How robust is the training – can people still perform well when they're under stress or trying to do several things at once? How do individual differences factor in?

3) Are different system designs optimal for different applications? For instance, is one design well suited to a multitasking or problem-solving environment while another is well-suited for repeat users rapidly accessing information from a well-defined database? How can we go about researching such potential differences?

References

Aist, G. 2004. Speech, gaze, and mouse data from choosing, placing, painting, rotating, and filling (virtual) vending carts. *International Committee for Coordination and Standardisation of Speech Databases (COCOSDA)* 2004 Workshop, Jeju Island, Korea, October 4, 2004.

Campana, E., Tanenhaus, M. K., Allen, J. F. and Remington, R.W. (2004). Evaluating Cognitive Load in Spoken Language Interfaces using a Dual-Task Paradigm. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP).* Jeju Island, Korea. October, 2004.

Campana, E., Baldridge, J., Dowding, J., Hockey, B. A., Remington, R. W., and Stone, L. S. (2001). "Using eye movements to determine referents in a spoken dialog system" in *Proceedings of the Workshop on Perceptive User Interfaces*. ACM Digital Library. November 2001.

Campana, E., Brown-Schmidt, S. and Tanenhaus, M.K. (2002). Reference resolution by human partners in a natural interactive problem-solving task. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP).* Denver, CO, September 2002.

Dowding, J. and **Campana, E.** (2004). Spoken Dialog Systems in MDRS Rotation 29. Seventh Annual Mars Society Convention, Chicago, IL, August 2004.

Hockey, B.A., Lemon, O., **Campana, E.**, Hiatt, L., Aist, G., Hieronymus, J., Gruenstein, A., and Dowding, J. (2003) Targeted help for spoken dialog systems: Intelligent feedback improves naive users' performance. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, April 2003.

Swift, M.D., **Campana, E.**, Allen, J.F., and Tanenhaus, M.K. (2002). Monitoring Eye Movements as an Evaluation of Synthesized Speech. In *Proceedings of the IEEE Workshop on Speech Synthesis*. Santa Monica, CA, September 2002.

Biographical Sketch

Ellen Campana is currently pursuing a Joint Ph.D. in Brain & Cognitive Sciences and Computer Science at the University of Rochester, USA. Her Ph.D. work is financed through a NASA GSRP fellowship. She holds a B.S. in Computer Science and a B.S in Psychology from the University of Wisconsin-Madison, and an M. A. in Brain and Cognitive Sciences from the University of Rochester. In her copious spare time Ellen enjoys world travel, dancing (lindy), playing with her four cats, knitting, and driving motorcycles. She also spends a lot of time and energy striving to improve the University of Rochester's Graduate Organizing Group, a college-wide graduate student organization which she has been the president of for 2 years.

Stephen Choularton

Centre for Language Technology
Macquarie University
Sydney

stephenc@ics.mq.edu.au
www.ics.mq.edu.au/~stephenc

1 Research Interests

My research interests lie generally in the area of **spoken dialogue systems**, with a special interest in commercial deployed **task-oriented** systems. Currently my main focus is on the problems that **speech recognition errors** cause to the successful completion of such dialogues. I am working towards a theory of error recognition and repair in spoken dialogue systems as my Ph.D. work.

2 Past, Current and Future Work

My previous honours work involved the design of a semi-automatic tool for writing task-oriented dialogues in VoiceXML, and to manage database connectivity. It exploited the idea of dialogue design patterns to provide general templates for such dialogues. Currently, I am involved in my Ph.D. work on handling speech recognition errors in spoken dialogue systems.

2.1 Corpus Work

At the heart of the work is the study of two large corpora. The first, the *Pizza Corpus* is a large corpus arising from a trial of a commercial grammar based pizza ordering system in Australia. The second is the Colorado University Communicator Corpus (*Communicator*). This is an academic n-gram based system, that provides flight information, and if requested, hotel and car hire information in the destination city. Brief comparative statistics are given in Table 1.

Feature	Pizza	Communicator
Language Model	Grammar	*N*-gram
Acoustic Model	Australia	American
Dialogues	2486	2334
Utterances	32728	40522
Words	54740	106562
Vocabulary Size	1048	2367
Error Rate	19.6 %	36.65%

Table 1: Corpora Comparison

My general position is that speech recognition errors occur because the utterance 'heard' by the computer is dissimilar to the acoustic model and/or the language model being employed by the speech engine. To be able to identify when they are occurring, I am undertaking work in both the acoustic and language domains.

2.2 Acoustic Errors

The Communicator Corpus has a sound file for each utterance. I take a more general position to the causes of errors in the acoustic domain than some other researchers (such as, for example Schriberg and Hirschberg) who appear to see hyperarticulation as the principal cause). In order to capture difference across as many dimensions as possible, I chose a wide range of features.including length, pitch, formants, jitter and shimmer, intensity and periods of silence, and automatically extracted them from the sound files associated with the corpus using Praat software.[1]

I have subjected this data to both logistic regression and support vector machine analysis (*SVM*). Predictions for unknown utterances using logistic regression are shown to the left of the table below and SVM, to the right:

	false	true	false	true
false	26	12	18	20
true	26	108	13	121

Logistic regression gave a prediction accuracy of 77.9% with a kappa of 43.3%. An SVM using a radial kernel produced gave an overall agreement of 80.8% and a kappa of 40.3%. These high levels of accuracy in prediction encourage me in thinking I will be able to identify errors successfully.

2.3 Language Errors

Pursuing the ideas involved in a double recognition approach involving a first grammar based recognition, and a second wide vocabulary n-gram based recognition (for example see Gorrell et al. [2002]), I have studied the Pizza Corpus for out of language (*OOL*) utterances. The data is very interesting and a tentatively analyze of it data as shown in table 2:
The first objective of this work is to discover if one can identify OOL utterances, which will necessarily be misrecognized by any recognizer. We also desire a metric to measure the degree of distance of such OOL utterances from the language model. We have designed a 'similarity' metric based on the Levenstein distance between the hypothesis and the utterance.

[1] http://www.fon.hum.uva.nl/praat/

	number	%
Errors on the language axis	861	7.9%
Errors on the acoustic axis	833	7.7%
Name spelling	194	1.8%
Void turns - correctly recognized	799	7.4%
Correct recognitions	8150	75.2%
Totals	10837	100%

Table 2: Taxonomy of misrecognitions in the Pizza Corpus

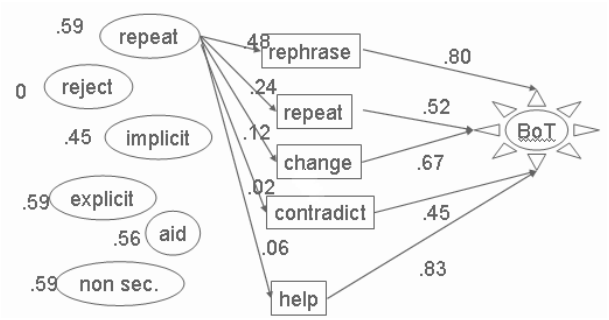


Figure 1: The probabilities of various strategies succeeding.

It is trivial to find any OOL utterances from an annotation. Of course no recognizer would be as perfect as the human annotation, so I have endeavored to introduce errors into the annotation. One well know open source recognizer, Sphinx 4[2] claims a word error rate (*WER*) of 2.7, 7.2 and 18.9% with 1,000, 5,000 and 64,000 word vocabulary n-gram language models, and this level of errors still allows one to operate at the 50% to two-thirds correct prediction of OOL utterances. I also speculate that similarity may becomes a further feature to introduce into the ultimate classifier to ensure that high levels of error recognition can be achieved.

2.4 Error Repair

2.4.1 Determining System prompt Output

I have undertaken studies on the Pizza Corpus that involve tagging for errors, the clues the system employs to inform users of an error, the reactions of users, and the length of time it takes for the dialogue to get back on track [Choularton and Dale, 2004]. The counts that are derived from such a study allow one to calculate the best repair strategy. If one defines the best strategy as the one that gets a dialogue back on track most often, one can produce the Bayesian network shown in figure 1. I have shown all the intermediate probabilities for the system clue *repeat* but omitted them for the other clues. The figures on the left show the probabilities of getting back on track for each type of clue. A pilot study was carried out on under two hundred dialogues, so the data sample is really too small to produce sound conclusions. However, as the method provided a rational approach to determining the 'best' strategies to use to repair errors, I intend to carry out the study on a larger part of the corpus in order to produce useable finding.

2.4.2 Managing Repair Subdialogues

The Pizza Corpus policy of rejecting any hypothesis below a certain tuned threshold was largely successful in ensuring that misrecongized facts or instructions did not enter the dialogue. Unfortunately, while this policy was very successful in achieving what might have been its prime objective, it produced considerable user frustration, requests for the operator and hang ups.

It is well known that the actual words employed between people when problems are being resolved can materially affect the feelings of the parties. While we can establish the correct strategy for the system to use from the work referred to in section 2.4.1. It we may require further study of 'diplomatic' language to be able to phrase the prompts in a manner that satisfies all the requirement of an effective repair subdialogue.

[2] http://cmusphinx.sourceforge.net/sphinx4/

3 Challenges in Spoken Dialog Systems Research

There are many challenges in spoken dialog systems research. I believe that one of the most important, at this time.is the problems that recognition errors cause. However, there are many other challenges particularly with dialogues that are more open than simple task-oriented dialogues. These involve achieving a more general conversational competence,

References

S. Choularton and R. Dale. User responses to speech recognitin errors: Consistency of behaviour across domains. Sydney, December 2004. Australian Language Technology Workshop, ASSTA.

G. Gorrell, I. Lewin, and M. Rayner. Adding intelligent help to mixed-initiative spoken dialogue systems. ICSLP-2002, 2002.

Biographical Sketch

2001 Completed B.Sc., Computer Science at Macquarie University.

2002 Won the Motorola Language Technology Pogram Honours Scholarship 2002 and completed First Class Honours Degree

2003 Won Australian Post-graduate Award Scholarship plus University Top-up Scholarship to undertake Ph.D. research.

Matthias Denecke

NTT Communication Science Laboratories
2-4 Hikaridai Seika Cho
Kyoto 619-0237
Japan

matthias@opendialog.org
www.opendialog.org

1 Research Interests

My research interests can be categorized into three areas. I am interested in **rapid prototyping** of spoken dialogue systems. This includes the description of dialogue and task models. Furthermore, I am interested in **statistical dialogue management** and learning from experience to optimize dialogue strategies, in order to build **robust systems**. Finally, I am interested in **Interactive restricted domain question answering systems**.

2 Past, Current and Future Work

2.1 Rapid Prototyping of Spoken Dialogue Systems

My work on rapid prototyping focusses on whether is it possible to separate application specific and generic aspects of a dialogue manager in such a way that it is possible to build a new system in a new domain and / or for a new language by reusing without changes the generic component(s) and supplying (specifications of) the application specific parts. Part of the proposed solution is the implementation of specification-driven dialogue algorithms ((Denecke and Waibel, 1997) for ontology-driven clarification question generation). This approach resulted in the implementation of the domain independent dialogue manager ARIADNE. It could be shown experimentally that ARIADNE is a platform allowing rapid development of new applications (Denecke, 2002).

2.2 Open Source Dialogue Manager

The dialogue manager ARIADNE developed to evaluate the rapid prototyping approach is available as open source software at `www.opendialog.org` under an Apache-style license. The available components provide a platform that, together with freely available Microsoft speech recognizers and synthesizers, can be used to rapidly develop new dialogue applications. It supports the development of applications in multiple languages in parallel. Languages using non-latin character set, such as Chinese and Japanese, are equally supported. Using this platform, systems are known to have been implemented in English, German, Japanese an Chinese. It has been downloaded by researchers in academia and industry and is in use in several projects in Europe, the USA and Asia.

2.3 Machine Learning of Dialogue Policies

Learning dialogue strategies using reinforcement learning based on user feedback is an attractive approach as the optimized dialogue strategies minimize on average dialogue length, or increase user satisfaction. However, the amount of data needed for optimization is a common problem. One potential solution is to provide approximative solutions of the Markov process. My colleagues and I have been looking at two ways to determine approximate solutions. One solutions consists of dynamically clustering dialogue states in groups and treat a state cluster as one state, effectively deciding that the optimal actions in all states are to be the same (Denecke et al., 2004). The second approach consists of determining optimal actions in infrequently visited states based on optimal actions in neighboring frequently visited states (Denecke et al., 2005).

2.4 Interactive Restricted Domain Question Answering

Generally, spoken dialogue system assume that the back-end system (e.g., the database or the API that is to be accessed by voice) provides some sort of structure that can be exploited by the dialogue manager. For example, fields in database tables correspond to slots in frame-based representations used in the dialogue manager, and impose therefore restrictions on the dialogue manager, for example, which values to prompt for. The situation changes dramatically once the back-end becomes an unstructured text corpus, such as a newspaper corpus or a set of technical manuals. It is no longer possible to exploit assumptions on the back-end to decide which kinds of questions to ask., or whether to terminate the dialogue. At the same time, dialogue is an important component in restricted-domain question answering systems. This is because in the desired applications, for example interactive trouble shooting applications, it is highly unlikely that the user specifies all necessary information in one turn (or even knows what information is necessary). Therefore, there is a need for a new approach to dialogue management different from dialogue management for task oriented systems.

Recently, my colleagues and I at NTT have been looking at how such an approach to dialogue management

might look like. The idea of our proposed solution is to simplify dialogue management as much as possible and shift the burden of generation of appropriate output to an example-based generation component. Simplyfing the dialogue managers' task has the advantage that the problem of data sparseness is reduced; and we can train classifiers to determine appropriate solutions for dialogue management. (see (Denecke and Tsukada, 2005) (Denecke and Yasuda, 2005)).

3 Challenges in Spoken Dialog Systems Research

It is inherently difficult to develop robust, well-working spoken dialogue systems, especially if the complexity exceeds those of database lookup systems. What makes this so difficult? I believe the question of *dialogue system specification* to be orthogonal to the question of *dialogue system optimization*. The latter is concerned with decisions related to the uncertainty of the input channels while the former is concerned with specifying the "character", for a lack of better term, of the application. So let's look at both aspects separately.

As for the question dialogue system specification, it may be instructive to look at the development of graphical user interfaces in the 80's. Together with the GUIs, object-oriented programming language became popular. In object-oriented programming languages, it became easy for developers to say things such as "I want a button just as this one, except it has to be a bit different.", since object-oriented programming languages provide mechanisms for subclassing and overriding. The important thing to note is that developers do not need to understand every aspect of GUIs in order to implement them. Going even further, (at least some) GUI guidelines can be implemented in a framework in such a way that it is difficult for developers to violate them. As far as the development for spoken dialogue systems is concerned, we do not know how to say "I want a system like this one, but it should behave differently sometimes, and here is where and how". But this idea expresses the central concern of software engineering, namely *encapsulation of concerns.*

As for the question of optimization, macine learning approaches have been proven to be useful in optimizing dialogue strategies w.r.t. to clarification questions or rejection; that is, these approaches address the issue of erroneous input.

For these reasons, I believe the following two statements to be true. First, knowledge-based approaches to dialogue management will co-exist with machine learning approaches, because both approaches address different needs. It will require additional research to decide which aspect of developing a dialogue system is better to be specified symbolically vs. learned automatically. It will also require further investigation how best to integrate the specifications with the results of the machine learning algorithms. Finally, it will be important to encapsulate results of both aspects into reusable components, because encapsulation (allowing for reusability without understanding) is the base for rapid development of new systems.

Second, similar to the case of graphical user interfaces, a solution allowing for easy specification of spoken dialogue systems, along the lines of object oriented programming languages used for GUI programming, might not be a solution coming from within the spoken dialogue system community. This solution (if it ever comes to existence) may very well be a contribution from software engineering.

References

M. Denecke and H. Tsukada. 2005. Instance-Based Generation for Interactive Restricted-Domain Question-Answering Systems. In *Proceedings of the 2nd International Joint Conference on Natual Language Processing*. Springer Verlag.

M. Denecke and A.H. Waibel. 1997. Dialogue Strategies Guiding Users to their Communicative Goals. In *Proceedings of Eurospeech, Rhodos, Greece*. Available at http://www.opendialog.org/info_papers.html.

M. Denecke and N. Yasuda. 2005. Does this answer your Question? Towards Dialogue Management for Interactive Restricted Domain Question Answering Systems. In *Proceedings of the 6th* SIGDIAL *Workshop, Lisbon.*

M. Denecke, K. Dohsaka, and M. Nakano. 2004. Learning Dialogue Policies using State Aggregation in Reinforcement Learning. In *Proceedings of the International Conference in Speech and Language Processing, South Korea*.

M. Denecke, K. Dohsaka, and M. Nakano. 2005. Fast Reinforcement Learning of Dialogue Policies using Stable Function Approximation. In *Lecture Notes in Artificial Intelligence*. Springer Verlag.

M. Denecke. 2002. Rapid Pprototyping for Spoken Dialogue Systems. In *Proceedings of the International Conference on Computational Linguistics, Taipei, Taiwan*.

Biographical Sketch

Matthias obtained a Masters in Computer Science from Karlsruhe University, a Masters in Computational Linguistics from Stuttgart University and a doctorate on rapid prototyping in spoken dialogue systems from Karlsruhe university. He interrupted his studies to work in two startup companies to see aspects of the real word side of speech processing systems. He is currently a research associate at NTT Communication Science Laboratories and, besides enjoying life in Japan, works on interactive restricted domain question answering systems.

Peter Fröhlich

ftw. Telecommunications Research Center
Vienna
Tech Gate Vienna, Donau-City-Str.1, A-1220
Vienna
Austria
froehlich@ftw.at
www.ftw.at

1 Research Interests

In the last years, promising speech user interfaces have emerged, such as mobile voice services (TellMe in the US, or the Austrian A1 voice service), multimodal research demonstrators (e.g. ftw. MONA, see Wegscheider et al 2004), or powerful supportive products for the blind (e.g. JAWS). However, most speech services tend to focus on conveying lexical, text-based information. The resulting user experience may be comparable to that of a purely textually-based visual website layout. My current PhD work is concerned with a general approach to overcome these limitations in expressiveness of today's speech-enabled services by the use of paralinguistic information, non-speech sound and multimodal representations. The overall approach is to produce prototypical design solutions, to conduct experimental user studies and to define guidelines for speech user interface design (see also Fröhlich 2005b).

2 Past, Current and Future Work

In this section, the research issues identified for my PhD project are described. The common denominator of these studies is to investigate and validate ways to enhance the expressiveness of speech-based user interfaces. If not indicated otherwise, the studies have been conducted with a research copy of the leading Austrian voice portal, the "A1 voice service". Sections 2.1 to 2.4 describe past to current work. Sections 2.5 and 2.6 relate to plans for currently planned research, which will be at a more mature advanced stage at the time of the workshop.

2.1 Expressive speech synthesis

Based on an Email reader application, expressive elements to enhance the pre-processing of non-lexical information in Email texts (i.e. certain non-speech sounds for separation and quotation marks or emoticons) were developed and validated. The results generally indicate that carefully designed non-speech sounds expressing non-lexical elements in spoken email and news texts increased user satisfaction and performance. (see Fröhlich and Hammer 2004).

2.2 Expression of time-dependent data

The basic research interest was to compare speech and non-speech sound regarding the ability to express time-dependent data in telephony speech applications, especially system-response time. In a user study, waiting cues as well as silence durations were compared with regard to error rate and subjective satisfaction (Fröhlich 2004a).
One major result was that music pieces were rated as much more appropriate and pleasant than synthetic or natural sounds. Based on this result, an electronic questionnaire study is currently being conducted, hopefully yielding more knowledge about the relevant characteristics of music pieces (e.g. familiarity, activation, simplicity, musical style). Data collection is ongoing.

2.3 Non-speech sound interaction feedback

The basic research interest was to investigate whether simple, frequently occurring speech feedback items (e.g. "The mail has been deleted" or "OK") can efficiently be replaced by non-speech sounds ("auditory icons"). One could argue that non-speech sound might decrease user workload and annoyance, however with a potential rise of user errors. The results (not yet published) do not support the replacement of simple speech feedback by sound, at least for the relatively short usage duration investigated in the study (15 minutes).

2.4 Expressiveness vs. annoyance

The main interest was to investigate the trade-off between increased richness and expressiveness (achieved by auditory icons, different voices or music) and potential subjective user annoyance. The following issues were addressed systematically (results not yet published):

- What is the optimal amount of sound to be used in voice services?
- Should sound and speech be combined sequentially or in parallel in a voice service?
- Are natural or musical sounds preferred in voice services?
- Are there certain characteristics of background music to spoken text that are relevant for subjective satisfaction?

2.5 Spatial audio

Regarding the goal to increase the expressiveness of speech systems, spatial audio in mobile devices promises several benefits, e.g. further dimensions to display auditory information, and a more realistic experience in augmented reality applications.
Currently, our research group is building two mobile prototypes in which spatial audio will also be implemented: a location- and position-aware 3D visualisation service and a presence awareness application. It is planned to conduct comparative experiments on the benefits of spatial audio for these types of application areas.

2.6 Multimodal user interfaces

The research activities so far have been conducted with auditory interfaces, especially telephony based voice services. Of course, many problems imposed by the auditory domain (such as seriality and lack of persistence) could be compensated by adding visual navigation or memory aids. Taking advantage of an emerging telecommunications standard that enables the combination of circuit- with packet-switched networking (IMS), we are planning to develop and validate a prototype of a "visually enhanced mobile voice service".

3 Challenges in Spoken Dialog Systems Research

In the context of my research work, I see the following challenges for spoken dialog systems research:

3.1 Integration of different auditory interaction styles

One important general challenge is directly attached to the research issues described above and thus I would like to mention it first. There should be a common framework for integrating linguistic, paralinguistic and non-linguistic auditory information. We should have more evidence when to enable which kind of input/output style in the user interface. One potential step to a clearer picture in this regard is to gather more empirical research evidence and to showcase promising enabling technologies and applications. This attempt is currently made by organizing a workshop "Combining speech and sound in the user interface" at ICAD05 (Fröhlich and Pucher, 2005).

3.2 Interaction robustness:

Speech-based human-computer interaction should not be impaired by interruptions and decelerations caused by inaccurate speech recognition. Therefore, basic research into models and algorithms to improve speech input performance is crucial (mainly out-of-scope for HCI researchers). But also speech output needs to be as easily understandable and natural as possible in order to support an efficient and satisfactory interaction flow (see e.g. Fröhlich 2004). One way of optimizing the output quality of mobile speech systems is to systematically compare different implementation alternatives, such as mobile device class, synthesis method, data rate or lexicon usage (see Pucher and Fröhlich, 2005).

3.3 Mobile multimodal speech applications

Due to form factor limitations of mobile devices, visual output and gestural input are resource-demanding and cumbersome. Speech interaction is widely regarded to be a promising alternative, because it is not dependent on device size (apart from technical constraints, of course). One challenge in this regard is to identify useful context characteristics for speech (concerning privacy, background noise, attention constraints etc) and to find promising services.

Further important research challenges for spoken dialogue systems relevant to me are to exploit the emotional cues inherent in the speech signal, e.g. for emotionally responsive in–car dialogue systems. Furthermore, flexible dialogue strategies for coping with speech recognition problems are important issues to be addressed.

References

Fröhlich, P. (2004). Increasing Interaction Robustness of Speech-enabled Mobile Applications by Enhancing Speech Output with Non-speech Sound. *Proc. ROBUST 2004, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, Norwich, England, August 2004.*

Fröhlich, P. (2005a). Dealing with System Response Times in Interactive Speech Applications. *Extended Abstracts of CHI 2005, Portland, Oregon, USA, April 2005.*

Fröhlich, P. (2005b). Non-Speech Sound and Paralinguistic Parameters in Interactive Speech Applications. *Extended Abstracts of CHI 2005, Portland, Oregon, USA, April 2005.*

Fröhlich P. and Hammer F. (2004). Expressive Text-to-Speech: A User-centred Approach to Sound Design in Voice-enabled Mobile Applications. *Proc. Second Symposium on Sound Design, Paris, France, October 2004.*

Fröhlich P. and Pucher M. (2005). Combining Speech and Sound in the User Interface. ICAD 05 Workshop. Position papers of the workshop to appear in: *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland, July 6-9, 2005*. Workshop homepage: http://userver.ftw.at/~frohlich/workshop.htm

JAWS, http://www.freedomscientific.com/

Pucher, M. and Fröhlich, P. (2005). A User Study on the Influence of Mobile Device Class, Synthesis Method, Data Rate and Lexicon on Speech Synthesis Quality. To appear in: *Proc. Interspeech 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisboa, Portugal.*

TellMe, http://www.tellme.com

Wegscheider F., Simon R., Tolar K. Position Paper for the *W3C Workshop on Metadata for Content Adaptation. October 12/13, 2004, Dublin, Ireland*

Biographical Sketch

Peter Fröhlich is a user interface researcher at the speech group of ftw. Telecommunications Research Centre Vienna, Austria. Prior to joining ftw., Peter has worked as a usability engineer at CURE, Center for Usability Research and Engineering in Vienna.

He holds a diploma in Psychology (University of Salzburg) and a diploma in Music Education (University Mozarteum Salzburg). He is pursuing his PhD studies in Applied Psychology at the University of Vienna.

Peter has several years experience in conducting and managing user centred design activities for both industrial and IST-funded development projects. He has been teaching User-Centred Design on several Austrian technical colleges.

Sudeep Gandhe

Institute for Creative Technologies,
University of Southern California
13274 Fiji Way, Suite #600
Marina Del Rey, CA 90292

gandhe@ict.usc.edu

1 Research Interests

My research interests lie in developing technologies for Natural Language Dialog systems, that would enable a coherent interaction between the user and the system. The Natural language technology has reached a maturity where natural language interfaces have became feasible to many applications. They are replacing simple command or menu driven inputs. The particular applications I have been looking at is interactive media retrieval. These types of systems have had great success for training and entertainment. They are designed so that users can ask any question and receive a prerecorded response. I have been specifically looking into how to make such interactions more coherent. I am also interested in applying statistical methods to dialog modeling.

2 Past, Current and Future Work

I have had the opportunity to work at Institute for Creative Technologies and have worked myself on some projects and closely observed a few more. Following is the summary of my work and what I have learned so far, organized according to the projects.

2.1 Speech-to-speech translation

We developed a speech to speech translation system for medical domain. (Narayanan et. al., 2004) Using this system an English speaking doctor can communicate with a Farsi speaking patient and carry out the medical diagnosis. The system is composed of many modules, viz. automatic speech recognizer, machine translation, dialog manager, GUI and speech synthesis. My work focused on GUI and the dialog manager. We built a GUI which facilitates the communication between patient and the doctor. Only one participant, the doctor, can control the interaction. After speaking the utterance, the doctor is presented with multiple interpretations of that utterance and the doctor can choose one from those. The GUI also shows the history of the current dialog along with possible next utterances the doctor may choose to speak. The dialog manager component in this system is different from most of the dialog systems, in the sense that it has no active participation in carrying out the dialog. It can only assist the communication process. With this goal in mind, we split the dialog in phases. viz, introduction, registration, Q&A, physical examination, diagnosis, conclusion. We also analyzed different medical cases, cardio, neuro, ent, ortho, ... The approach used here was, that a total of 1400 commonly used medical phrases were hand-tagged for phase and case. A classifier trained on these predicts the current phase and case. Based dialog history and current phase and case estimations, the next possible doctor utterances are predicted and presented to the doctor for selection. The challenge remains getting the dialogs tagged for concepts and what to present as a suggestion.

2.2 Coherent interactions

The basic idea of a question answering system, where answers are pre-recorded video segments has proved very useful in various applications for training and entertainment as well. Users are allowed to input a free-text question which in turn elicits a pre-recorded video response. Although the video response tends to have very good value in terms of immersive experience, the very design of the system allows for a lack of coherence, especially when there are no video responses directly answering the question or are not phrased in desired manner. We tried to address this issue by introducing short linking dialog between question and answer and thus bridging the gap. We carried out experiments to assess whether such linking dialogs can increase the coherence of interaction and proved that interactions with human-generated linking dialogs are statistically significant when compared to interactions without linking dialogs. (Gandhe et. al., 2004) Further analysis of human-generated linking dialogs reveals that these carry more information than present in the answer or the question. This leads us to realize the need for a knowledge base behind such a system. We have started build-

ing such a knowledge base and are experimenting with simple computer generated linking dialogs.

3 Challenges in Spoken Dialog Systems Research

Gathering natural language dialog data by role-playing and through wizard experiments is a well established practice for dialog research. This data is mainly utilized for building lexicons, language models for speech recognition and understanding the domain of interaction. Although this data can also be useful for dialog modeling, there are very few systems which use this directly to implement statistical methods for coming up with next utterance for the system. Annotating such corpus for dialog moves and then operating at that level is time-consuming. I believe, utilizing this gathered data to its full extent is an important problem to solve.

Natural language understanding has always been the important component in a spoken dialog system. It sits right before the dialog manager in the traditional pipeline architecture and is a very important input. It is not hard to believe, that however good this component may be, there will be some instances of utterances, possibly outside the domain of interaction, that will not be well understood by the NLU. This presents another important challenge for the spoken dialog systems – how to deal with failures in NLU?

References

S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Gandhe, S. Ganjavi, P. G. Georgiou, C. M. Hein, S. Kadambe, K. Knight, D. Marcu, H. E. Neely, N. Srinivasamurthy, D. Traum, and D. Wang. 2004. *The Transonics Spoken Dialogue Translator: An aid for English-Persian Doctor-Patient interviews.* in working notes of the AAAI Fall symposium on in working notes of the AAAI Fall symposium on Dialogue Systems for Health Communication, pp 97--103.

Sudeep Gandhe, Andrew Gordon, Anton Leuski, David R Traum, and Douglas W. Oard. 2004. *First steps toward linking Dialogues: mediating between free-text questions and pre-recorded video answer*. presented at the Army Science Conference.

Biographical Sketch

He is currently pursuing a PhD degree in Computer Science at University of Southern California, Los Angeles. He works at Institute for Creative Technologies under the advisement of Dr. David Traum. Before that he has completed Masters in Computer Sceince from USC and has received B. Eng. (Computer) from Mumbai University, India. His interest lie in developing techologies for coherent communication especially in question answering systems.

César González-Ferreras

Universidad de Valladolid
Departamento de Informatica
ETIT, Campus Miguel Delibes
Valladolid 47011, SPAIN

cesargf@infor.uva.es
www.infor.uva.es/~cesargf

1 Research Interests

My research interests are **spoken dialog systems**, with special focus on **accessing on-line information using speech**, and **multimodal dialog systems**, with special focus on **languages to describe the interaction**.

2 Past, Current and Future Work

I am currently working on providing access to web contents using speech. Two different approaches have been proposed. In the first one, a solution for any web site was built. In the second one, we restricted the domain to newspaper web sites, in order to develop ad-hoc strategies tailored to the contents of that domain. Both systems were developed using VoiceXML.

Browsing Internet contents using speech is becoming more and more important, mainly because the spread of mobile devices which allow web access anytime and anywhere. However, the task is not easy, because a restructuration of the information is required to adapt it to the new modality, very different from the visual one: speech interaction is sequential and not persistent. We can not present all the information at once, like in a traditional web browser, we have to dialog with the user in order to give her only the information she wants. The lack of metainformation describing web contents makes more difficult the conversion, because content authors emphasize visual appearance instead of structuring the contents properly.

In our first approach, (González-Ferreras and Cadeñoso-Payo, 2004), we have developed a system that dynamically converts HTML pages into VoiceXML ones allowing the information to be accessed using a VoiceXML browser over the telephone line. A voice application describes how the conversion has to be done for each HTML page. We created a development tool which helps in building that voice applications. Once the application is built, we use a transcoding server to access the information using speech. Five typical HTML patterns have been identified and we provide for each of them a way of accessing the information using speech. The system is useful to access HTML pages in which information changes very often but the internal structure of the page (HTML code) remains unchanged.

In our second approach, (González-Ferreras and Cadeñoso-Payo, 2005), we developed a spoken dialog system which provides access to a newspaper web site. The system is based on an information model, which structures the information, and on an interaction model, which describes how the interaction is carried out. First, the information model is built automatically from web contents. Next, the system uses that model to interact with the user using browse and search strategies. If the user has a specific information need, search can be used to access it directly. Browse can also be used to see which information is available. We carried out an evaluation of the system to measure its usability. System performance is measured, obtaining a task success rate of 92% and a word error rate of 18.09%. User satisfaction is measured, obtaining positive results. Users said the system is useful and easy to use, although the interaction is repetitive and boring.

As future work we plan to focus on search strategy, trying to overcome the limitations of automatic speech recognition (ASR). Problems with ASR affect usability, because users feel confused when the system does not understand them. The main problem is the size of the vocabulary, because news domain has an open vocabulary. We need to find ways to restrict the vocabulary for a given state of the interaction and to find a mechanism to deal with out of vocabulary words.

I have also some experience with multimodal dialog systems. We have built a first prototype, (González-Ferreras et al., 2004), that integrates spoken dialogs in a virtual reality application, in order to achieve a more natural interaction and increase interaction possibilities. We have defined a framework to develop applications using VoiceXML as language to describe dialogs and VRML as language to describe the virtual world. We have used XML to integrate both modalities. The framework uses components stored in a repository to create applications. This helps in reusing from one application to another. The control flow in our system is described by VoiceXML pages and this makes speech control the interaction. As future work we plan to create a language to describe the interaction for both modalities, in order to allow any modality to have the initiative.

3 Challenges in Spoken Dialog Systems Research

Deploying dialog systems in the real world is one of the key challenges. In the last decades there has been a lot of effort in advancing speech technology and finally we have seen some research spoken dialog systems (SDS) that interact with people in a natural and flexible way. Speech technology is prepared to be used in real systems and the focus must be on developing and deploying systems for real users.

However the development process of SDS is not mature enough. Building SDS is very different from building traditional computer systems. Speech interaction is sequential and not persistent, and this makes it very different from traditional systems, which have a graphical user interface (GUI).

We need to create a branch of software engineering to deal with SDS. We need to find a model to describe SDS at analysis and design phases of development, like relational model is used to describe data bases or Unified Modeling Language (UML) to describe object oriented systems. We need to adapt the development process to focus on the interaction, involving final users in the process, as has been done in some agile methods. We also need CASE tools tailored to SDS and reference of good practices and design patterns. All this effort will help to make easy the development process, helping to deal with the complexity of SDS. Better systems will be developed: more maintainable, flexible and extensible. The research community is aware of this, and there are some proposals of methodologies, guidelines and tools for developing SDS. I expect more work on this in the following years.

Another key challenge of SDS is usability, and few attention has been given to it until now. When a SDS is deployed in the real world, factors that affect its usability must be studied. Traditionally, systems with more natural interaction have been the goal of SDS. However, I think natural systems and usable systems are different concepts. There must be a shift from natural interaction to usable interaction. In order to study usability, evaluation on the systems has to be done using psychometrics theory. This has been done in evaluation of GUI systems and it is beginning to be used also for SDS, (Larsen, 2003).

An important factor of usability is the conceptual model of the system that the users have. Users must understand how the system works and which are its main limitations (for instance, the language the system is able to understand). A way of achieving that is to establish conventions and train the users how to use SDS. This has been done in other inventions, in which the usage is not natural but people has become used to, as described in (Heisterkamp, 2003). We must develop systems with high usability, although their interaction is not natural, like Universal Speech Interface (Rosenfeld et al., 2001).

User adaptation can also increase usability of SDS. Adapting some system models can improve performance (for instance, acoustic models). An user model can be used to tailor the interaction to each user and user preferences can be used to decide which information should be presented first. Speaker identification can determine the user in an unobtrusive way, without using any security information such as a password.

References

César González-Ferreras and Valentín Cadeñoso-Payo. 2004. Building Voice Applications from Web Content. In *International Conference on Text, Speech and Dialogue (TSD)*.

César González-Ferreras and Valentín Cadeñoso-Payo. 2005. Development and Evaluation of a Spoken Dialog System to Access a Newspaper Web Site. In *European Conference on Speech Communication and Technology (Eurospeech)*.

César González-Ferreras, Arturo González-Escribano, David Escudero-Mancebo, and Valentín Cadeñoso-Payo. 2004. Incorporación de interacción vocal en mundos virtuales usando VoiceXML. In *Congreso Español de Informática Gráfica (CEIG)*.

Paul Heisterkamp. 2003. "Do not attempt to light with match!": Some thoughts on progress and research goals in Spoken Dialog Systems. In *European Conference on Speech Communication and Technology (Eurospeech)*.

Lars Bo Larsen. 2003. *On the Usability of Spoken Dialogue Systems*. Ph.D. thesis, The Faculty of Engineering and Science, Alborg University.

Ronald Rosenfeld, Dan Olsen, and Alex Rudnicky. 2001. Universal speech interfaces. *Interactions*, 8(6):34–44.

Biographical Sketch

César González-Ferreras was born in Barcelona, Spain. He received the B.Sc. and M.Sc. degrees in computer science in 1998 and 2000 respectively, both from University of Valladolid (Spain). He is currently working toward the Ph.D. at Computer Science Department of University of Valladolid. In 2001 he joined the Department of Computer Science at University of Valladolid, where he is currently an assistant professor. His research interests include spoken dialog systems and on-line information access using speech.

Genevieve Gorrell

Department of Computer and Information Science
Linköping University
581 83 LINKÖPING
Sweden

gengo@ida.liu.se
www.ida.liu.se/~gengo

1 Research Interests

My research focus is **language modelling** for limited-domain speaker-independent spoken dialogue systems. From a background of spoken dialogue systems design, including work with dialogue management and semantics, my interests have moved more recently toward language model optimisation. I have investigated the relative merits of **grammar-based** and **stochastic** approaches, and am currently developing matrix-based methods for use in n-gram language modelling. The relevant techniques are **Latent Semantic Analysis**, **singular value decomposition**, **eigen decomposition** and **Random Indexing**.

2 Past, Current and Future Work

My research in spoken dialogue systems initially focused on the relative merits of grammar-based and stochastic language models in spoken dialogue systems, before moving on to consider how the different approaches might be combined to best effect–this work is described in the next section. More recently, I have been investigating machine-learning methods for eigen decomposition, a technique with many potential applications to language modelling. This work is described in section 2.2.

2.1 Comparing and Combining Approaches to Language Modelling

In (Rayner et al, 2001), we demonstrated the relevance of agreement constraints to the performance of a grammar-based language model. At the same time, our work on a comparison of grammar-based and stochastic language models in the same domain (Knight et al, 2001) explored the circumstances under which one can expect superior performance from a grammar-based and a robust speech understanding system respectively. We also presented a "plug-and-play" technique for grammar compilation (Rayner et al, 2001; Rayner et al, 2003), appropriate to situations where the coverage of the language model needs to be manipulated in real time.

Over the next few years I went on to investigate techniques for exploiting language modelling approaches to provide assistance to users in the case that they are misrecognised. "Targeted Help" (Gorrell et al, 2002) uses a stochastic language model to classify an utterance misrecognised by the primary grammar-based recogniser and give one of a limited number of help messages. An investigation was also done into the potential for using a back-off SLM to discover out-of-coverage vocabulary (Gorrell, 2003). Targeted Help was then extended into a system that uses layered recognisers and Latent Semantic Analysis for classification (Gorrell, 2003; Gorrell, 2004).

2.2 Eigen Decomposition for Language Modelling

Continuing my work with Latent Semantic Analysis (Deerwester et al, 1990), I am currently investigating the applicability of eigen decomposition to language modelling. Eigen decomposition describes the transformation of a matrix into a set of orthogonal vectors, each with a value describing the weight of the matrix in that orientation. Any square matrix can be decomposed in this way, and the vector set will be no greater in dimensionality than the original matrix. It may in fact be considerably smaller, since most natural matrices contain some redundancy. Since the eigenvalues give the importance of each vector within the matrix, it is often useful to discard the lower values, allowing the data to be compressed, and in some contexts, meaningfully generalised.

Singular value decomposition (SVD) extends eigen decomposition to arbitrary rectangular matrices. Instead of one set of eigenvectors being produced, it decomposes the matrix into a pair of vector sets. SVD forms the basis of Latent Semantic Analysis (LSA). In LSA, a matrix is formed from a corpus of text passages, with one line per unique word and one column per text passage. The text passages might be article abstracts or document titles, for example. The matrix is populated with frequency data and (potentially after some preprocessing) decomposed using singular value decomposition. The lower value vector pairs are discarded and the matrix reconstructed. The previously sparse matrix is now meaningfully filled out, such that if a pair of words appear frequently together, in the cases that they do not appear together, the previously zero count now contains a non-zero value. For example, a document that refers to cats exclusively as felines

might now contain a non-zero value in the “cat” slot. In this way, the technique can be said to have discovered the synonymity between “cat” and “feline”.

An important feature of human language acquisition is that human beings receive input in a continuous and unending stream of individual items. They continue to improve their model over time. I am interested in developing an approach that shares these properties. For this reason, I am currently investigating incremental approaches to eigen decomposition. The Generalized Hebbian Algorithm (Sanger, 1989) is a neurologically plausible incremental implementation of eigen decomposition. My current work involves applying the Generalized Hebbian Algorithm to LSA. Another aspect to my current work is the extension of GHA to rectangular matrices. I am also using Random Indexing (Kanerva et al, 2000) to optimise the technique.

My next steps are to apply the machine-learning techniques I have been working with to language modelling. Previous work has demonstrated the value of LSA-style techniques in language modelling, for example (Bellegarda, 2000). I hope to build on the existing work with the addition of a theoretically interesting and practically useful technique.

3 Challenges in Spoken Dialog Systems Research

The great success of stochastic language modelling, despite its simplicity and failure to embody linguistic knowledge, has meant that grammar-based language modelling has been neglected in many ways. There are undeniable advantages to an approach that is more knowledge-rich, for example, in the semantic interpretation of utterances. As stochastic language modelling reaches its peak, a transition to more knowledge-rich approaches will need to be made if performance is to continue to improve. Making this transition poses a fascinating challenge.

References

Genevieve Gorrell. 2004. *Language Modelling and Error Handling in Spoken Dialogue Systems*. Licentiate thesis, Linköping University, Sweden.

Rayner, M, Boye, J, Lewin, I and Gorrell, G. 2003. *Plug and Play Spoken Dialogue Processing*. In ”Current and New Directions in Discourse and Dialogue”. Eds. Jan van Kuppevelt and Ronnie W. Smith. Kluwer Academic Publishers.

Genevieve Gorrell. 2003. *Recognition Error Handling in Spoken Dialogue Systems*. Proceedings of 2nd International Conference on Mobile and Ubiquitous Multimedia 2003.

Genevieve Gorrell. 1999. *Using Statistical Language Modelling to Identify New Vocabulary in a Grammar-Based Speech Recognition System*. Proceedings of Eurospeech 2003.

Gorrell, G, Lewin, I and Rayner, M. 1999. *Adding Intelligent Help to Mixed Initiative Spoken Dialogue Systems*. Proceedings of ICSLP 2002.

Knight, S, Gorrell, G, Rayner, M, Milward, D, Koeling, R and Lewin, I. 2001. *Comparing Grammar-Based and Robust Approaches to Speech Understanding: A Case Study*. Proceedings of Eurospeech 2001.

Rayner, M, Lewin, I, Gorrell, G and Boye, J. 2001. *Plug and Play Speech Understanding*. Proceedings of SIGDial 2001.

Rayner, M, Gorrell, G, Hockey, B. A, Dowding, J and Boye, J. 2001. *Do CFG-Based Language Models Need Agreement Constraints?*. Proceedings of NAACL 2001.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas and Richard A. Harshman. 1990. *Indexing by Latent Semantic Analysis* vol 41:6. Journal of the American Society of Information Science.

Terence D. Sanger 1990. *Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network* vol 2. Neural Networks.

J. Bellegarda 2000. *Exploiting latent semantic information in statistical language modeling* vol 88:8. Proceedings of the IEEE.

Kanerva, P, Kristoferson, J and Holst, A. 2000. *Random Indexing of Text Samples for Latent Semantic Analysis*. In Gleitman, L.R. and Josh, A.K. (Eds.): Proceedings of the 22nd Annual Conference of the Cognitive Science Society, p. 1036. Mahwah, New Jersey: Erlbaum, 2000.

Biographical Sketch

Genevieve holds a BSc. (Hons) in psychology from the University of Central Lancashire, UK, an MPhil. in Computer Speech and Language Processing from Cambridge University, and a Lic. Fil. from Linköping University, Sweden. She has worked at Netdecisions, UK, as a senior researcher, and as a visiting scientist at NASA Ames Research Center and at the Knowledge Engineering and Discovery Research Institute, New Zealand. She is currently completing her doctoral studies at Linköping University.

Alexander Gruenstein

Spoken Language Systems Group - CSAIL
Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA 02139

alexgru@csail.mit.edu
http://www.mit.edu/~alexgru/

1 Research Interests

I am primarily interested in spoken dialogue systems at the level of words, utterances and meaning. I am particularly interested in how **computational models of context** can percolate constraints to other dialogue-system processes: for instance, how both conversational context and contextual world knowledge can be used to improve **natural language interpretation**, **natural language generation**, and **language modeling**.

2 Past, Current and Future Work

Generic Framework for Dialogue System Design At Stanford, I co-developed a framework geared toward dialogue systems that provide interaction with intelligent agents. My primary contribution was the development of generic components for interfacing the dialogue manager to an intelligent device. I developed a *Recipe Scripting Language* which can be compiled into an *Activity Model* (AM) that hierarchically models the tasks an agent can undertake in a manner particularly well suited for dialogue management (Gruenstein, 2002a). I also co-developed the *Dialogue Move Tree* (DMT) - a framework for representing conversational context which defines a search space of *information state update procedures* for updating the dialogue system's information state based on conversational contributions (Lemon and Gruenstein, 2004; Lemon et al., 2002). The DMT and AM together provide a robust framework for modeling the *context* of the conversation: while the DMT models the conversational context, the AM tracks the current, past, and future "cognitive" states of the intelligent agent. The framework is generic, and serves as the basis of several dialogue systems.

Context Sensitive Language Modeling I have explored several ways in which incorporating contextual information can be used to improve language modeling for speech recognition. For instance, contextual structure provided by the DMT can be utilized to constrain grammar-based language models by using contextual information to choose an appropriate *subset* of the grammar tuned to recognizing likely upcoming user utterances (Lemon and Gruenstein, 2004). Context-dependent language models are swapped into the recognizer, increasing recognition speed and accuracy. A back-off language model is used in a second (or parallel) pass if recognition results using the small language model do not exceed a confidence threshold.

Additionally, I have recently co-developed new techniques for incorporating context-sensitivity into n-gram language models (Gruenstein et al., 2005). By training with context-sensitive *dynamic classes* and then swapping in appropriate expansion sets at run time, n-gram language models can quickly and effectively be biased based on conversational context. For instance, utilizing context, the word *two* in the following training corpus snippet can be tagged as belonging to the class TIME rather than the more generic class DIGIT:

```
System:  When would you like to depart?
User:  Around two, please
```

The context-sensitive class TIME can then be appropriately populated at runtime.

Natural Language Generation I have also worked on mechanisms to integrate conversational context into natural language generation. In particular, the AM and DMT can be leveraged to ensure that an agent's conversational contributions are *relevant*, *still true when uttered*, *concise*, *non-repetitive*, and make natural use of *referring expressions* (Lemon et al., 2003). Additionally, effective contextual models can be used to provide natural language summaries of the progress of both the conversation and the tasks which are being undertaken (Gruenstein and Cavedon, 2004) – this allows autonomous agents in dynamic environments to accurately convey their "cognitive state."

Fragment Interpretation Contextual models of discourse can be used to effectively interpret utterances – especially non-sentential utterances – in context. For example, I have explored using the AM and DMT in the interpretation of corrective fragments such as *No, not there, the fire station* which are only felicitous in particular conversational contexts (Lemon and Gruenstein, 2004). I have also briefly explored this problem (Gruenstein, 2002b) using the *sign coercion* approach developed in (Ginzburg and Cooper, 2004).

Probabilistic Dialogue State Modeling I am also interested in modeling dialogue state probabilistically. I have done initial work in the area, allowing persistent uncertainty through multiple weighted attachment to the DMT (Gruenstein et al., 2004). Future work in this area might include the integration of Monte-Carlo methods with dialogue state update functions.

3 Challenges in Spoken Dialog Systems Research

Dialogue systems typically tend to rely on a "piped" architecture. For instance, in a typical architecture, the speech recognizer passes an n-best list to a parser, which hands off a set of parses to the dialogue manager, which then interprets the parse and produces a response, which is turned into natural language by a generation component and then realized via TTS. This approach yields disconnects which can be off-putting – such as user utterances which are "recognized" but not "understood." An important challenge is to find practicable paths for better integrating the disparate components which comprise a dialogue system, so that knowledge modeled in one part of the system informs processing in other components.

A second challenge lies in developing robust systems based on little or no training data. System design currently tends to rely on experts making their best guesses as to what a natural interaction with the system will be like, often hand-crafting language models and dialogue strategies – sometimes based on small wizard-of-oz studies – which are then iteratively refined as newer versions of a system are deployed. Natural language interfaces, unlike other computer interfaces, tend to engender specific user expectations – as users are already fluent in the medium of communication, and come with preconceived notions of how to accomplish the task at hand. Thus, it is a major challenge to *rapidly* develop robust, natural interfaces.

A third challenge is robust integration of multiple modalities in a portable environment. As portable computing power becomes cheaper and high-speed network connectivity pervasive, an opportunity emerges for integrating visual, tactile, and speech modalities in a persistent, personalized framework. Researchers should be looking in this domain to create truly *natural* systems, which should aim to leverage existing personal and persistent electronic devices such as cell phones and digital music players.

References

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates. *Linguistics and Philosophy*, 27(3):297–366.

Alexander Gruenstein and Lawrence Cavedon. 2004. Using an activity model to address issues in task-oriented dialogue interaction over extended periods. In *Proceedings of AAAI Spring Symposium on Interaction Between Humans and Autonomous Systems over Extended Periods*.

Alexander Gruenstein, Lawrence Cavedon, John Niekrasz, Dominic Widdows, and Stanley Peters. 2004. Managing uncertainty in dialogue information state for real time understanding of multi-human meeting dialogues. In *Proceedings of the 8th Workshop on Formal Semantics and Pragmatics of Dialogue*.

Alexander Gruenstein, Chao Wang, and Stephanie Seneff. 2005. Context-sensitive statistical language modeling. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

Alexander Gruenstein. 2002a. Conversational interfaces: A domain-independent architecture for task-oriented dialogues. Master's thesis, Stanford University.

Alexander Gruenstein. 2002b. English corrective fragments: Syntactic and semantic considerations. Unpublished manuscript. http://www.mit.edu/ alexgru/fragments.ps.

Oliver Lemon and Alexander Gruenstein. 2004. Multithreaded context for robust conversational interfaces: context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction*, 11(3):241–267.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. 2002. Collaborative activities and multi-tasking in dialogue systems. *Traitment automatique des langues*, 43(2):131–154. Special issue on dialogue.

Oliver Lemon, Alexander Gruenstein, Randolph Gullett, Alexis Battle, Laura Hiatt, and Stanley Peters. 2003. Generation of collaborative spoken dialogue contributions in dynamic task environment. In Reva Freedman and Charles Callaway, editors, *Working Papers of the 2003 AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 85–90. AAAI Press.

Biographical Sketch

Alexander Gruenstein is a second year PhD student in the Spoken Language Systems group at M.I.T. He received B.S. and M.S. degrees in Symbolic Systems from Stanford University, concentrating in Artificial Intelligence and Natural Language Technologies. He enjoys tennis, hiking, photography, and kite-flying.

Jeongwoo Ko

Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15206, USA

jko@cs.cmu.edu
http://www.cs.cmu.edu/~jko

1 Research Interests

My research interests lie generally in the dialog system for the automotive environment, with a special focus on flexible dialog switching and system-initiative dialogs in such systems.

2 Past, Current and Future Work

I have been working on the dialog system for mobile environment in the CAMMIA project. CAMMIA (Conversational Agent for Multilingual Mobile Information Access) is the spoken dialog system to provide route guidance and information service in English and Japanese. It has been extended to the multi-modal dialog system in order to support speech, touch screen and navigation map. The prototype systems have been developed especially for the automotive environment.

2.1 Flexible dialog switching

In the automotive environment, the direction dialog for route guidance is one of the most important tasks. But the direction dialog is very different from other kinds of information seeking dialogs in that it tends to be long (sometimes more than several hours) and does not need to be active all the time. It can be active only when telling the next direction to the user. Otherwise, it can be running in the background and the user may have other dialogs with the system. For example, the user can ask the system to retrieve auxiliary information (e.g., parking locations at the destination) while the primary task is already undersay (e.g., navigation to the destination).

To support multiple dialogs at the same time, the Dialog Manager (DM) maintains a dialog stack and supports the user interruption. Figure 1 shows an example. As you can see in S5, the system interrupts the user turn to notify the next direction and then continues the dialog by resuming the interrupted weather dialog.

2.2 System-initiative dialogs

I have been working on the system-initiative dialogs to provide useful information without any request from the user.

S1: What can I do for you today?
U1: I want to go to Carnegie Mellon University.
S2: Do you want to go to Carnegie Mellon University?
U2: Yes.
S3: The distance to the destination is 100 miles.
It takes about 2 hours.
U3: I would like to know weather.
S4: Please tell me the area and the date.
U4: Pittsburgh
(Navigation System sends the next direction to the Dialog Manager)
S5: To go to Carnegie Mellon University, please make a left turn here.
Please tell me the date for Pittsburgh.

Figure 1. Sample dialog for dialog switching

The driving direction is one type of the system-initiative dialogs. Even though the user initiates the dialog by setting the destination, the system should interrupt the user to tell the next direction (Figure 1). In our system, the next direction has the highest priority and the interruption is done immediately to tell it to the user without any negotiation steps.

I have been extending this framework to provide other types of information in the tour guide dialogs. For example, the system initiates a dialog when it finds famous sightseeing sites or user's favorite food near the car location. The dialog starts with a negotiation and the user may withdraw or delay it.

This work involves important research issues to find out when and how to provide information to the user. Especially, the user is driving a car and the interruption should be designed not to affect driving safety. I am working on simple reasoning component called IDA (Intelligent Dialog Agent). The IDA gets information from the user profile and other sub systems, and then predicts the proper timing and contents. Here are some features the IDA uses to make a decision.

- User preference
- Current location
- Status of wireless signal
- Time of day/Day of week

- Talking status (from DM history)
- Driving speed (from location change)
- Driving time (elapsed, route length)

When integrating the system with a car, we are going to add more features such as car speed, car speed change, gas status, temperature and microphone input (human conversation) My current research is to apply machine learning techniques with the features.

2.3 Anaphora/Ellipsis/Ambiguity resolution

I am also interested in Anaphora/Ellipsis resolution to support natural dialogs between the user and the system. This is also important to support flexible dialog switching. For example, the user is searching the sightseeing sites near the destination. During the search, the user can request the weather information ("Tell me the weather for the destination today"). This time, the system switches to the weather dialog by passing the city name where the destination is located.

Ambiguity resolution is another interesting area. One user utterance can be used in different ways based on the dialog context. For example, in Japanese, "どのくらいかかるの?" has several meanings: How much, How far, how wide, how tall, etc. The system needs to judge the meaning based on the dialog context. This is one of future work.

3 Challenges in Spoken Dialog Systems Research

Here are some challenges in Spoken Dialog Systems Research.

3.1 Grammar design for flexible dialog switching

The flexible dialog switching needs many active grammars to catch the user utterance for topic switching. As more dialogs are supported, the bigger grammars are needed, which makes recognition rate get lower.

As one solution, when we built the first prototype system, we designed the entry grammar which only includes the dialog names (e.g., weather, restaurant, sightseeing, direction). This does not increase the grammar size much even though the number of the dialogs is increased. But one problem is that this does not support natural dialogs between the user and the system. The user always has to say command-like words such as "weather", "restaurant" or 'sightseeing" to start a dialog. How to design the efficient grammars is an interesting challenge for this task.

3.2 Reasoning about user interruption

The system-initiative dialogs needs to interrupt the user. In the automotive environment, the user is driving the car and the interruption should be designed not to affect driving safety. Reasoning to predict proper contents and proper timing is one of the interesting challenges.

References

Hataoka, N., Y. Obuchi, I. Akahori, M. Tateishi, S. Judy, J. Ko, T. Mitamura, E. Nyberg. 2003. *Development on Speech Dialog Management System, CAMMIA*. Proceedings of the Acoustical Society of Japan, Autumn Conference, Nagoya, Japan

Eric Nyberg, Teruko Mitamura, Paul Placeway, Michael Duggan, Nobuo Hataoka. 2002. *Dynamic Dialog Management with VoiceXML*, Proceedings of HLT

Project web site
http://www.lti.cs.cmu.edu/Research/CAMMIA

Biographical Sketch

Jeongwoo Ko is a Ph.D student in Language Technologies Institute, School of Computer Science at Carnegie Mellon University. She received her Bachelor Degree in Computer Science from Ewha Womans University, Korea and her Master Degree in Information Systems Management from Carnegie Mellon University.

Ian R. Lane

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ATR SLT Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0288, Japan

ian@ar.media.kyoto-u.ac.jp
www.ar.media.kyoto-u.ac.jp/~ian

1 Research Interests

My research interests span the areas of speech recognition, natural language processing and information retrieval. I am particularly interested in incorporating "*high-level*" knowledge into the ASR (automatic speech recognition) framework. I believe such approaches are necessary to realize robust speech recognition and understanding. Towards this goal, I am currently investigating approaches that incorporate "*topic*" information for various tasks within spoken language systems.

In the future spoken language interfaces will become a necessary component for many information retrieval applications. By combining voice with other modalities of interaction, effective and efficient human-computer interfaces can be realized. Spoken language interfaces are also relevant for tasks other than information retrieval, for example facilitating communications between users, via speech-to-speech translation (Takezawa, 2005) or by acting as a third participant in dialogue.

2 Past, Current and Future Work

In my current research, I am focusing on spontaneous dialogue between native Japanese and English speakers via a speech-to-speech translation system (Takezawa, 2005). For this task, dialogue is simple and semantically explicit compared to natural human-human dialogue, between speakers of the same language. However, it is more spontaneous than typical human-computer interaction, and thus provides a good platform for spoken dialogue research.

For this application, I have investigated; topic-dependent recognition (Lane, 2005a), based on topic detection and language model switching, out-of-domain utterance detection (Lane, 2004), and confidence measure generation for ASR output (Lane, 2005b). Overviews of these three research topics are given in the following subsections.

2.1 Topic-dependent speech recognition for multi-domain spoken dialogue

When performing spoken dialogue over multiple domains, topic- or sub-task-dependent language modeling increases both the accuracy and efficiency of the system. However, current dialogue systems that use multiple topic-dependent language models typically adopt a system initiative approach (Wessel, 1999) where the appropriate LM is applied based on the system's prompt, determined by the dialogue flow of the system.

To realize a flexible ASR framework, I developed a novel architecture combining topic detection and topic-dependent language modeling (Lane, 2005a). In this framework, the inferred topic is automatically detected from the user's utterance, and speech recognition is then performed, applying an appropriate topic-dependent language model. This allowed users to seamlessly switch between domains while maintaining high recognition accuracy. To improve system robustness, a hierarchical back-off mechanism was incorporated where detailed topic models were applied when topic detection was confident and wider models that cover multiple topics were applied in cases of uncertainty.

2.2 Out-of-domain utterance detection

The second area of research I have focused on is the detection of OOD (out-of-domain) utterances that cannot be handled by the backend system. To operate effectively and realize robust speech recognition, spoken dialogue systems are specifically designed to operate over a limited and definite domain, as defined by the back-end application. For users, however, the exact definition of the application domain is not necessarily clear, and users, especially novice users, often attempt OOD utterances. For an effective user interface, systems must provide feedback to the user, informing them when an OOD utterance is encountered. This will enable users to determine whether to continue the current task after being confirmed as in-domain, or to halt, after being informed that it is OOD and cannot be handled by the back-end system.

In (Lane, 2004) I investigated OOD utterance detection based on topic classification and in-domain verification. In the proposed approach, the application domain of the system was assumed to consist of multiple sub-domain topic-classes. OOD detection was performed by first calculating classification confidence scores for all of

these classes, and then applying an in-domain verification model to the resulting confidence vector.

2.3 ASR confidence scoring

In recent work (Lane, 2005b), I have investigated utterance verification incorporating "*high-level*" knowledge. Previous approached for assessing the confidence of ASR output are typically based on feature-based methods, explicit model-based schemes, or posterior-probability-based approaches. These approaches, however, estimate recognition confidence based on the "*low-level*" information that is available during decoding. There are apparently knowledge sources outside the ASR framework, which have not been well exploited for estimating recognition confidence.

Two confidence measures were investigated. The first, *in-domain confidence*, (as derived in out-of-domain utterance detection) is a measure of match between the input utterance and the application domain of the back-end system. The second, *discourse coherence*, is a measure of the consistency between consecutive utterances in a dialogue session. A joint verification confidence was generated by combining these two measures with a conventional measure based on the GPP (generalized posterior probability (Lo, 2005)) of the ASR output. Incorporating the two proposed measures significantly improved utterance verification accuracy compared to using GPP alone, and when negligible ASR errors (that do not affect translation) were ignored, further improvement was achieved.

3 Challenges in Spoken Dialog Systems Research

To realize robust spoken dialogue systems challenges that must be considered include; the detection and handling of errors during dialogue, and methods to incorporate "context" knowledge into the speech recognition, understanding and dialogue management components.

As errors will occur in human-computer spoken dialogue, to realize an effective user interface, the system should detect the type of error and provide informative feedback to the user. This will enable the user to determine the reason for system failure and recover from it. For example, if the input utterance is OOD, the user should be informed of the application domain of the system, if an utterance contains OOV (out-of-vocabulary) words, the user should be informed that this word is unknown. Similar feedback should be provided for acoustic (noise and channel mismatch), and other linguistic errors. Effectively detecting a wide range of errors and realizing cooperative dialogue is one significant challenge for spoken dialogue research.

Current ASR systems rely on only acoustic and linguistic information during decoding. To realize robust speech recognition, external information relating to the "context" of the current utterance is required. Such information can relate to a wide variety of knowledge, including "context" from the application domain, external-knowledge, such as time-of-day, and even user-dependent knowledge via user-modeling. An effective framework to combine a large number of varied knowledge sources is required to realize such an architecture.

References

T. Takezawa et. al 1998. *A Japanese-to-English speech translation system: ATR-MATRIX*, In Proc. ICSLP, pp. 957–960.

I. Lane, T. Kawahara, T. Matsui and S. Nakamura, 2005. *Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching*, IEICE Trans., Vol.E88-D, No.3, pp.446–454.

I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, 2004. *Out-of-Domain Detection based on Confidence Measures from Multiple Topic Classification*, In Proc. ICASSP, Vol.1, pp.757–760.

I. Lane, and T. Kawahara, 2005. *Utterance Verification Incorporating In-domain Confidence and Discourse Coherence Measures*, Accepted EUROSPEECH, 2005.

F. Wessel and A. Baader, 1999. *Robust Dialogue-State Dependent Language Modeling using Leaving-one-out* In Proc. ICASSP, Vol.2, pp. 741–744.

W. K. Lo, and F. K. Soong, 2005. *Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences*, In Proc. ICASSP, pp. 85–89.

Biographical Sketch

Ian R. Lane received the B.Tech degree in information engineering from Massey University, New Zealand in 2000. He was awarded a Monbukagakusho scholarship to study in the School of Informatics, Kyoto University, where he received the M.E. degree in 2003 and is currently undertaking research towards a Ph.D. degree. He is also an intern researcher at ATR Spoken Language Translation Research Laboratories. Mr. Lane is a member of the Acoustical Society of Japan (ASJ), and IEEE.

Piroska Lendvai

Tilburg University
Dept of Language and Information Science
PO Box 90153, 5000 LE Tilburg
Netherlands

p.lendvai@uvt.nl
ilk.uvt.nl/~piroska

1 Research Interests

I am a postdoc researcher at the University of Tilburg in the Netherlands, working in the Induction of Linguistic Knowledge research group. I apply **supervised machine learning techniques** for **pragmatic-semantic processing** of natural language data. The data typically describe shallow properties of **spoken user input** to a dialogue system, such as **prosodic measurements** of the acoustic signal, a **bag-of-words** representation of ASR output, **dialogue history**. The goal of my research is **robust partial understanding** of the user input. The understanding is based on the identification of a simple set of **dialogue acts** and **semantic entities** in the user turn, but also on spotting whether the user signals **awareness of communication problems** during the interaction, and on predicting possible **future interaction problems** based on the shallow properties.

Apart from facilitating full understanding of the user input, such partial interpretation can be fed back to the speech recognition and the dialogue manager of the dialogue system for utilisation. The interpretation process is developed to cope with **noise** in spoken input (such as **disfluencies**) and in the shallow representation of such input (such as a flat bag-of-words vs a more structured **representation of ASR output**, and to account for **multi-layeredness** both in the input content and in the classification task designed for learning. Although such a method is generally language and domain independent, it is to be integrated in a specific application. The phenomena that are encoded in the learning process as features or as classes therefore inevitably contain **domain knowledge**.

This entails that for example dialogue acts or semantic entities might be differently formulated depending on whether the application concerns air travel, spare time activity, health, etc. Another focus of my research is therefore to investigate by machine learning methods the extent to which a certain identifiable category (e.g. 'user is aware of problem', 'user has filled a specific slot that has been asked for by the system') is correctly formulated, or whether a task can be **more optimally formulated** for machine learning applied in the dialogue system. For example, it can be that one **machine learning algorithm bias** prefers simultaneous learning of more dialogue phenomena, whereas another technique performs better when those tasks are learnt in isolation. It can also turn out that some **task is over- or underspecified** and is suboptimally learnable: for example, too many or too few categories of 'communication problem' were defined for the learner. This entails the investigation whether it is sufficient to define machine learning tasks for spoken dialogue system design based on human background knowledge, or whether it would also be beneficial to automatically infer some (simple) **domain ontology** for this purpose.

2 Past, Current and Future Work

My recently accomplished PhD project focused on applying machine learning techniques to extract pragmatic-semantic information from spoken user input in the Dutch OVIS dialogue system (train travel domain). The learning task involved simultaneous task-related act and information unit type classification, as well as bidirectional problem detection. I investigated the following research issues: (i) to what extent a memory-based and a rule induction machine learner can be used for interpretation of user turns in spoken dialogue systems, (ii) whether the complex learning task of four-level interpretation can be optimised by decomposing it to subtasks, and (iii) whether filtering noise from spoken input on the basis of higher-level linguistic information leads to improved learning performance on the interpretation task.

Systematic search in the space of possible subtask combinations revealed that it is possible to find significantly better co-learnable compositions of pragmatic-semantic interpretation subtasks than identifying each component in isolation or in full combination, and that the optimal component combinations can be meaningfully different per learning algorithm.

It was also found that when ASR output was filtered from disfluent words (training on the Corpus of Spoken Dutch that contains transcribed human-human dialogues), from syntactically less dominant words, or from words that do not frequently occur in the recognition hypotheses, disfluency filtering and chunk non-head filtering had a positive but statistically insignificant impact

on the partial interpretation subtasks, whereas frequency-based filtering deteriorated classification performance. We furthermore investigated five different flattened representations encoding the ASR lattice. The results show that the most robust type of these representations is the flat, binary bag-of-words encoding of the recognition hypothesis. The following papers have been published on this research: (Lendvai et al., 2002a; Lendvai et al., 2002b; Lendvai, 2003; Lendvai et al., 2003a; Lendvai et al., 2003b; Lendvai and Maruster, 2003; Lendvai et al., 2004; Lendvai, 2004; Lendvai and van den Bosch, 2005).

Currently I participate in the Dutch national IMIX project (Interactive Multimodal Information eXtraction) that aims at developing a spoken dialogue system for question answering in the medical domain. The current version of the system is being used to collect human-machine dialogues, to which the approach described above is applied for detecting pragma-semantic information. I am also working on automatic processing of background medical documents, so that – based on a simple extracted domain ontology – candidate answers to user questions can be disambiguated. The adaptation of dialogue act and semantic entity types to the specificities of medical information-seeeking dialogues is planned according to the obtained results, so that the system can perform an improved analysis of user input on top of the current domain concept and semantic relation type spotting.

3 Challenges in Spoken Dialog Systems Research

An important issue for state-of-the-art spoken dialogue systems is how pragmatic and semantic information need to be encoded for optimal processing of the user input. Knowledge-based definition of fine-grained categories prove unnecessary when robust approaches drawing on shallow, automatically obtainable dialogue properties can produce a similar result. It is however so far uncharted how much of the manually predefined information types is necessary to provide to learning algorithms (either in the form of features or of classes), and how fine-grained those should be? It needs to be empirically investigated whether there is a difference between dialogue domains or processing techniques in the way they utilise different representations of dialogue phenomena. Clearly, there are interdependencies between feature design and task design when training a machine learner, as both encode domain knowledge, but it can be that some phenomena are better encoded as (one or more) features, and some as (simple or more fine-grained) classes. It is also an issue whether domain ontologies can be of help here, in case it turns out that there are differences in this respect between various deployed dialogue systems. Machine learning provides a framework for evaluating differently designed setups of pragma-semantic processing, and after training the processing results can be fed back into the analysed system.

References

P. Lendvai and L. Maruster. 2003. Process discovery for evaluating dialogue strategies. In *Proc. of ISCA Workshop on Error Handling in Spoken Dialogue Systems*, pages 119–122.

P. Lendvai and A. van den Bosch. 2005. Robust ASR lattice representation types in pragma-semantic processing of spoken input. In *Proc. of AAAI-05 workshop on Spoken Language Understanding*.

P. Lendvai, A. Van den Bosch, E. Krahmer, and M. Swerts. 2002a. Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting. In *Proc. ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*.

P. Lendvai, A. Van den Bosch, E. Krahmer, and M. Swerts. 2002b. Multi-feature error detection in spoken dialogue systems. In *Proc. Computational Linguistics in the Netherlands (CLIN '01)*. Rodopi Amsterdam.

P. Lendvai, A. van den Bosch, and E. Krahmer. 2003a. Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In *Proc. of EACL Workshop on Dialogue Systems: Interaction, adaptation and styles of management*, pages 69–78.

P. Lendvai, A. van den Bosch, and E. Krahmer. 2003b. Memory-based disfluency chunking. In *Proc. of Disfluency in Spontaneous Speech Workshop (DISS'03)*, pages 63–66.

P. Lendvai, A. van den Bosch, E. Krahmer, and S. Canisius. 2004. Memory-based robust interpretation of recognised speech. In *Proc. of SPECOM '04, International Conference on Speech and Computer, St. Petersburg, Russia*, pages 415–422.

P. Lendvai. 2003. Learning to identify fragmented words in spoken discourse. In *Proc. of EACL Student Research Workshop.*, pages 25–32.

P. Lendvai. 2004. *Extracting Information from Spoken User Input*. Ph.D. thesis, Tilburg University, Netherlands.

Biographical Sketch

Piroska Lendvai received the Ph.D. degree in computational linguistics from Tilburg University, Netherlands in 2004, and the M.A. degrees in English and in Russian philology from Janus Pannonius University, Pécs, Hungary in 1998. She is currently a postdoc researcher at Tilburg University working on robust language understanding for question-answering dialogues in the Interactive Multimodal Information eXtraction (IMIX) project.

Udhyakumar.N

Sanyo LSI Technology India Pvt. Ltd,
Unit 03, Level 08, Discoverer Block,
ITPL, Whitefield, Bangalore,
INDIA – 560066.

Email : udhyakn@sanyo.co.in
Webpage : http://udhyakumar.tripod.com

1 Research Interests

My primary research objective is to design and deploy real-world **spoken language dialog systems** for illiterate masses that are almost untouched by IT revolution. I am particularly focusing on building speech interfaces (**recognition** and **synthesis**) for the dialog systems in a **multilingual** environment. I firmly believe that these spoken language dialog systems address poor literacy issues in taking technology to the developing regions of the world.

2 Past, Current and Future Work

I have worked on developing robust, multilingual speech recognizers for telephone based information retrieval systems in Indian languages. Recently I worked with the speech group of TIER team, University of California, Berkeley on the design and evaluation of Tamil Market, a dialog system which provides information on weather and crop prices. I am also working on Interactive Voice Mail, a speech interface for sending and receiving voice messages. A small overview of the above projects is given in the following subsections.

2.1 Multilingual speech recognition

I built a multilingual speech recognition system for Hindi and Tamil, the most widely spoken Indian languages as my final year project. This work is conducted mainly as an investigative study to analyze various challenges in developing ASRs for dialog systems in Indian environment. I have listed some of the issues we have addressed:

- Language switching is common in India, where there is a general familiarity of more than one language. This demanded a multilingual ASR to decode words from several languages (N.Udhya kumar et. al 2004).
- A spoken language dialog system in a multilingual environment should be able to respond to the user in his language. Hence integrated language identification is performed to inform the later stages (speech synthesis) the language being used (C.S.Kumar. et. al 2005).
- To address the sparseness of training data, cross-lingual bootstrapping of English acoustic models is employed.
- The system is planned to be embedded in information retrieval applications. So we included an automatic grapheme to phoneme converter to handle dynamic vocabulary.

2.2 Tamil Market

Tamil Market is a spoken language dialog system which enables illiterate farmers to get information about weather and price of various crops. It is one of the projects undertaken by the TIER (Technology and Infrastructure for Emerging Regions) group of UC, Berkeley. The main aim is to develop a user-interactive dialog system with minimum resources, effort and time. The system comprises of a Tamil speech recognizer, dialog manager and voice response generation unit (Chuck Wooters 2004). My responsibility in the team is to assist them in the design of the dialogues and to improve the accuracy of the ASR used.

We face a lot of design challenges in Tamil Market. Tamil has a lot of dialects which vary with the geographical location. This heavily affects the accuracy of the speech recognizer. The dialog system is completely state-based and only a subset of words is active in each state. Thus we are able to handle a medium-sized vocabulary since the recognizer has to search only within a small set of words at each stage. The design also takes into account the errors made by users, as they haven't been exposed to computers before. Machine-directed voice response is used to guide the user prompting him to produce short answers for a series of questions. Fall back routines are also provided to correct the user when he makes a mistake.

We identified different villages in Tamilnadu for speech collection to have a representative data of the entire group. We also evaluated our design among illiterates in those villages and collected their feedbacks. We used a "Wizard of Oz" study in the current evaluation since the speech recognizer is not perfected yet due to lack of training data. The project is still in the design phase and the users' feedbacks and speech data are used to improve the system.

2.3 Interactive Voice Mail

Interactive voice mail aims to provide a speech-based interface for sending and receiving voice messages. With the advance of wireless networking, internet is fast becoming an accessible technology in developing regions. People still find the email difficult as they are not conversant using computers. Hence this solution is particularly suited for illiterate users who want to experience the use of internet messaging.

The users can login to the system, browse and listen to the incoming messages, compose new messages and update their address book through an interactive speech based interface. Appropriate voice response in users' native language is provided to guide them since they are unaware of the limitations of computer interaction. Similar to the Tamil Market, the recognizer's accuracy is taken into account in the dialog design. The system is still to be evaluated for usability study.

3 Challenges in Spoken Dialog Systems Research

I have presented a few topics below which I believe to be some of the important and current challenges in spoken dialog systems research. Recently there is a growing interest in developing innovative, real-world applications for masses for which computing resources remain largely out of reach. There has been relatively little work in the development of effective user interfaces for these applications, especially in an environment with low literacy and with a wide range of languages and dialects. Spoken language dialog systems can enable these applications by providing support for small-vocabulary, hands-free speech recognition and generation, which is critical for developing regions. There are a lot of challenges in deploying such systems which call for novel research methods.

Multilingual systems which handle a number of languages in a single framework have become popular now. They address several issues like lack of training data, quick adaptation to a new target language and compact integration. Multilingual spoken dialog system is a challenging research area which has a lot of real-world applications, since the users can interact with the system in their native language. The system components should be made as language-independent as possible. There are also other issues like data collection, evaluation and flexibility of adding new languages which need to be addressed (James Glass et. al 1995).

Usability evaluation is an important process in the development of Spoken Dialog systems. The purpose of evaluation is to analyze design errors and estimate how well the system fits its purpose and meets actual user needs and expectations. In spite of its key importance far less resources have been invested in the usability evaluation measures of SDSs over the years than in its component technologies. Currently there is no standard as to which evaluation criteria to use. Metrics should be formulated which can include component testing but should assess the overall performance of the system in users' perspective (Alicia Abella et. al 1997).

References

Alicia Abella, et. al 1997, "Evaluating interactive dialogue systems: Extending component evaluation to integrated system evaluation" In Proc, ACL/EACL Workshop on SDS, Madrid, Spain.

Chuck Wooters 2004, "Automatic Speech Recognition for Developing Regions", Berkeley ICT4B Workshop, USA

James Glass et. al 1995, "Multilingual Spoken language understanding in the MIT Voyager System", Speech Communication, pp 1-18.

C.S.Kumar et. al 2005, "Language Identification for Multilingual speech recognition systems", accepted for presentation, HCII, Lasvegas, USA.

N.Udhyakumar et. al 2004, "Multilingual speech recognition for Information retrieval in Indian context", In Proc. HLT/NAACL 04 Student research workshop, Boston, MA, USA

N.Udhyakumar, C.S.Kumar and R.Srinivasan 2004, "Decision tree learning for automatic grapheme to phoneme conversion for Tamil", In Proc. SPECOM, St.Petersburg, Russia.

Biographical Sketch

Udhyakumar was born in Tamilnadu, India in 1982. He received his B.E in Electronics and Communication from Amrita Institute of Technology in 2004. He is currently working as a design engineer in the Speech and Audio group of Sanyo LSI. His interests include machine learning, digital signal processing and human-computer interaction. He is currently interested in designing spoken language dialog systems focusing the needs of emerging markets like India.

Udhay has good experience in developing multilingual speech interfaces for Indian languages. His project on "Multilingual speech recognition for Hindi and Tamil" has been selected as the best undergraduate project of the year. He has published three international and two national papers. Udhyakumar is a member of IEEE, IETE and ISCA.

Tim Paek

Microsoft Research
One Microsoft Way
Redmond, WA 98052
timpaek@microsoft.com

1 Research Interests

Ever since computers were invented, people have dreamed of having conversations with them. When speech recognition came around and dazzled the world with dictation, the dream seemed ever closer. That was several decades ago. Since then, researchers have realized that dialog is a much more complicated process than previously thought. Dictation is a good example. According to usability studies, the primary reason why so many users try dictation but subsequently stop using it is because of frustration involving dialog about what words were misrecognized, what sociolinguistics have called "repair." On the other hand, repair comes very natural to human beings, and it is a known fact that people with less than perfect hearing still manage to engage in conversation and accomplish their tasks. So, how is it that people with various kinds of deficiencies, such as poor hearing, unfamiliarity with the native language, or even aphasia, still manage to maintain a conversation with other people, while a computer with similar deficiencies cannot?

This question began my interest in dialog research, and in particular **dialog management**. How can we teach dialog systems to manage misunderstandings as effectively as human beings, and hopefully, in such a way that their performance will degrade gracefully with word error rate? In attempt to answer this and related questions, I have explored **decision-theoretic**, and more recently, **reinforcement learning** (technically, a subset of the former) approaches to optimizing dialog repair strategies. I am also interested in simplifying the **authoring** of dialog systems as well as automatically pinpointing areas where various kinds of **tuning** may be required, and finally, as a systems builder, in deploying real-time dialog **applications** that are well-designed and serve useful purposes.

2 Past, Current and Future Work

In examining how dialog systems might manage misunderstanding in a more "human" way, in previous research, I have tried to represent the process by which people achieve mutual understanding of their utterances and actions, a process known as **grounding**, using probabilistic techniques, and more generally, decision theory (Paek & Horvitz, 1999; Paek & Horvitz, 2000). I have tried to argue that grounding and decision theory together provide a good theoretical basis for selecting appropriate repairs (Paek & Horvitz, 2003) and can even be used to deal with other speech-related issues such as the speech target problem (Paek et al., 2000).

With respect to current research, I have recently been exploring methods for training dialog systems using reinforcement learning in an online fashion; that is, as a user interacts with the system. When users are engaged in speech interaction, the most natural thing to do when confronted with an error is to correct the system with utterances like "no" or "wrong." Unfortunately, most dialog frameworks are not designed to take user feedback like corrections and adapt in real-time their strategy or policy. I have implemented a voice-enabled browser that learns in an online fashion from user feedback. The browser is controlled by a graphical model for a finite-horizon Markov Decision Process for optimizing clarification dialogs. The parameters of the graphical model are updated in real-time using sampling. Since dialog interaction is potentially endless, and may or may not be stationary with regards to its transition probabilities and rewards, the system must deal with the "explore vs. exploit" dilemma. I have been investigating how different exploration strategies fare in this domain. All of this is forthcoming in a series of papers soon to be submitted for publication.

With respect to future research, I have been interested in simplifying the dialog authoring process so that designers who are not familiar with statistical methods can maximally exploit statistical learning without having to understand any of the mathematical details. Since the company I work for also builds a speech application SDK, I have been working with the product team to create simple tuning tools for diagnosing problematic situations in dialog interaction.

3 Challenges in Spoken Dialog Systems Research

IMHO, the reason why dialog research is so challenging and in many ways, slower in progress than other fields, deals primarily with the engineering difficulties of building a system on which to forge and evaluate re-

search ideas. Dialog research is similar to robotics in that much time is spent just putting together all of the different parts of the system. Furthermore, the system that is built is oftentimes tied to the application domain, which makes it difficult for researchers to generalize their results. Finally, evaluation is difficult to do "correctly" and also depend in many ways on the application domain (see Paek, 2001).

Another major challenge for dialog research is the lack of collaboration between academic research and industry practice. Many research ideas are looked upon by product designers as impractical, despite evaluation measures showing its increased performance. For example, despite the fact that reinforcement learning has been demonstrated to dialog management as good as if not better than hand-crafted rules, most developers are loathe to work within such a statistical framework. Even if they familiarized themselves with reinforcement learning, updating the performance of the system would require extensive support, especially if a change was needed in the state space. In short, to promote dialog research, I believe it will be very beneficial for the field to have more cooperative exchanges with industry. Ultimately, it will be industry and the drive for profits that will bring dialog systems to the general public. To that end, I hope to represent one of the companies interested in advancing both the state of art in dialog research as well as in speech application products.

References

Paek, T. 2001. Empirical methods for evaluating dialog systems. In *Workshop on Evaluation Methodologies for Language and Dialogue Systems*. ACL/EACL.

Paek, T. & Horvitz, E. 1999. Uncertainty, utility, and misunderstanding. In *AAAI Fall Symposium on Psychological Models of Communication*, North Falmouth, MA, November 5-7, 85-92.

Paek, T. & Horvitz, E. 2000. Conversation as action under uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 445-464. Morgan Kaufmann.

Paek, T., & Horvtiz, E. 2003. On the utility of decision-theoretic hidden sub-dialog. In *Error Handling in Spoken Dialogue Systems*, ISCA Tutorial and Research Workshop, 95-100.

Paek, T., Horvitz, E., & Ringger, E. 2000. Continuous listening for unconstrained spoken dialog. In *6th International Conference on Spoken Language Processing*. Beijing, China.

Biographical Sketch

Tim Paek is a researcher in the Machine Learning and Applied Statistics group at Microsoft Research. He received his M.S. in Statistics and Ph.D. in Cognitive Psychology from Stanford University. His primary research focus is on natural language understanding and dialog, user modeling, and user interface with a particular focus on applying statistical techniques.

Antoine Raux

Carnegie Mellon University
Newell-Simon Hall
5000 Forbes Avenue
Pittsburgh, PA 15232
USA

antoine+@cs.cmu.edu
www.cs.cmu.edu/~antoine

1 Research Interests

My research interests cover a broad range of topics from **speaker adaptation** for ASR, to **prosody modeling** for TTS, to **dialog management** and **turn-taking**. The main purpose of my work is to improve human-machine spoken interaction by enhancing the low level abilities (ASR, TTS, channel establishment) of spoken dialog systems.

2 Past, Current and Future Work

2.1 Non-Native Speakers

In the past few years, my main topic of research has been non-native speakers, and how to improve their experience with spoken dialog systems. From data collected with the CMU Let's Go bus information system (Raux et al, 2003), I studied linguistic (Raux and Eskenazi, 2004), phonetic (Raux, 2004), and acoustic (Raux and Singh, 2004) differences between native and non-native users. In the context of this system, and perhaps not surprisingly, while linguistic and phonetic differences did appear, acoustic differences seem to be the most important factor influencing the quality of the recognition of non-native speech. Hence while acoustic adaptation brought a significant gain in recognition accuracy, phonetic and language model adaptation brought only marginal improvement. In addition to improving the system's understanding of non-native speech, I proposed a method based on lexical entrainment to guide non-native users' language towards the lexical and syntactic structures expected by the system. While the initial experiment was not conclusive due to the poor quality of the overall interaction (mainly speech recognition), I am planning to reproduce this experiment in the near future, with improved language and acoustic models, as well as more controlled conditions.

2.2 Prosody Generation for TTS

In relation with my work on non-native speakers, I investigated ways to create more appropriate and natural prosodic contours for dialog utterances. In (Raux and Black, 2003), we describe a new method to generate F0 contours by concatenating portions of natural contours from recorded utterances, in the same way that concatenative speech synthesis generates utterances by concatenating portions of waveforms from recorded utterances. Human listeners preferred the utterances whose contour were generated with our method to those using a standard rule-based F0 contour generation method. The effect was even larger when modeling emphasized speech using a database specially designed and recorded for that purpose. Overall, this approach allowed us to build natural yet flexible (since prosodic and spectral content are modeled separately) contours in a very limited amount of time (assuming a database of the target phenomenon is available) and without requiring an expert to write prosodic rules.

2.3 Turn-Taking

While most research in spoken dialog systems has focused on task and on dealing with uncertainty in speech recognition and understanding, very little attention has been given to the way the system and the user manage turn-taking. In the vast majority of cases, spoken dialog systems assume a rigid turn-taking behavior where user input goes through a serial "understanding-dialog management-generation" pipeline and turn boundaries are detected using only pause information. This results in interaction issues like turn overtaking and unnecessary pauses. In addition, while it is known that prosody plays a central role in conversational speech, current systems' speech synthesizers typically use prosodic models trained on read speech or more natural but less flexible task-specific recordings. This results in the need for the system to almost always produce full sentences, as opposed to the prosodically rich fragments that are typically used in human-human conversation, particularly for confirmation and other grounding functions.

My current research aims at addressing these issues by designing a general framework that performs interaction management in parallel to traditional dialog management. As preliminary work, I am examining turn transitions in task-oriented human-human and human-machine conversation with the goal of identifying the salient features (semantic, syntactic, prosodic...) of turn transitions (TTs). This will lead to the definition of a "turn tran-

sition space". By analyzing the distribution of TTs in this space, I hope to identify types of human-human TTs and to compare the distributions of human-human and human-computer TTs. Ultimately, I hope to correlate metrics in the TT space with human perception of the quality and rhythm of a conversation. This should lay the ground work for me to investigate, design, and evaluate different interaction management algorithms.

3 Challenges in Spoken Dialog Systems Research

3.1 Building Avanced Yet Robust Systems

From a practical point of view, I believe that one challenge our field is currently facing is to build practical systems that go beyond today's commercially available menu-like voice-driven information access systems while still being usable by the vast majority of people. This involves two competing constraints: on the one hand we need to find tasks or means of interaction that significantly extends the reach of dialog systems, on the other hand we need to keep those systems robust enough so as not to harm their usability. One example of such a combined effort is natural language based systems that allow the user to utter complex utterances while detecting and correcting misunderstandings due to the open nature of their input. In what other directions can such a dual effort be conducted? For example, how can we give our systems more personality and social ability without harming their effectiveness on task? How can we create systems that handle complex tasks without frustrating customers with inefficient and/or inappropriate responses?

3.2 Towards Dynamic Conversation Management

As reflected in my current work, I believe that one major hurdle towards human-like interaction with machines (assuming this should be one of our goals) is the rigidity of the interaction. Certain conversational acts like confirmation and acknowledgment are far more frequent in human-human converations than they are in human-computer conversations of the same nature. Yet, the *cost* of such information-poor turns in term of disruption from the main task is much higher in human-computer interaction. My opinion is that this and other interactional problems we are facing today are the result of two related weaknesses: our systems' inability to dynamically deal with converation management signals such as backchannels, and the lack of flexibility of our synthetic voices, particularly in terms of prosody, which plays an important role in these signals. First, there is a need to acquire more knowledge about the form and function of conversation management signals, both from other fields such as linguistics and Conversation Analysis and by conducting our own empirical studies. Second, we should build new, more dyanmic system architectures that are able to incorporate such knowledge into human-computer interaction.

References

Antoine Raux and Brian Langner and Maxine Eskenazi and Alan Black 2003. *The Let's Go Spoken Dialogue System*, Eurospeech 2003, Geneva, Switzerland.

Antoine Raux and Maxine Eskenazi 2004. *Non-Native Users in the Let's Go Spoken Dialogue System: Dealing with Linguistic Mismatch*, HLT 2004, Boston, Massachusetts.

Antoine Raux 2004. *Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition* ICSLP 2004, Jeju Island, Korea.

Antoine Raux and Rita Singh 2004. *Maximum Likelihood Adaptation of Semi-Continuous HMMs by Latent Variable Decomposition of State Distributions* ICSLP 2004, Jeju Island, Korea.

Antoine Raux and Alan W Black 2003. *A Unit Selection Approach to F0 Modeling and its Application to Emphasis* ASRU 2003, Saint Thomas, US Virgin Islands.

Biographical Sketch

I am a 3rd year PhD student at Carnegie Mellon University, in the Language Technologies Institute. Prior to coming to CMU, I got a Master degree from Kyoto University, Japan. My Master's thesis was about Intelligibility Assessment and Adaptive Drill Generation for a Computer-Assisted Pronunciation Learning System. I pursued my undergraduate studies at Ecole Polytechnique (Paris, France), majoring in Computer Science/Engineering. As far as extra-curricular activities are concerned, I am married and the proud father of a little Yuma, which uses up most of my time outside school :)

Verena Rieser

Department of Computational Linguistics
Saarland University
Saarbrücken, D-66041
vrieser@coli.uni-sb.de
http://homepage.mac.com/verenarieser/

1 Research Interests

My research interests lie generally in multi-modal dialogue systems, with a particular interest on **multi-modal feedback generation** and **robust interaction strategies**. More specifically, I am applying **reinforcement learning** to **multi-modal clarification strategies**. Clarification in dialogue requests additional information to resolve understanding problems. My work explores how to combine various sources of interpretation uncertainty to decide whether to engage in clarification sub-dialogue and what kind of clarification strategy is most rewarding with respect to user satisfaction and task success.

2 Past, Current and Future Work

I started working on dialogue systems using simple handcrafted methods developing towards more "intelligent" and data-driven models. My interests in the field of usability engineering narrowed down to clarification strategies to assure robust interaction.

2.1 Usability Design in XML-based dialogue frameworks

I began working on spoken language interfaces searching for design strategies to enhance usability of information seeking dialogues. In my work I explored guidelines for speech system design (such as (Bernsen et al., 1998)) and used iterative prototyping. For dialogue management we used SDML, a XML-based modelling language allowing finite-state like dialogue models, (Brey et al., 2000). I built a talking washing machine named "Hermine" for exhibition at the CeBit 2003[1] (Rieser, 2003). The system used strategies such as context-sensitive help, iterative prompting, priming user utterances, grammar switching and short-cuts for expert users to assure robust interaction. Furthermore, graphical feedback was used to display current slot values, to reinforce the turn-taking signal and to signal non-understanding. The most noticeable result was that the personalisation of the system made it a main public attraction. By personalising the system we violated central design principles for task-oriented dialogues (such as "be informative", or "be relevant"). This raises the question whether abstract design guidelines can guarantee "good" dialogue design. The major drawback of the applied dialogue framework is the restrictive dialogue model and the use of "canned" output.

[1] www.cebit.de

2.2 Generation of Clarification Requests in the ISU-based Approach

In the Information State Update (ISU) approach to dialogue modelling, the information state represents contextual information and allows a more flexible control strategy as specified by update-rules. My work was aiming for a more robust and natural clarification strategy within this framework. A contrastive analysis of clarification requests in different corpora helped to identify features that influence human clarification strategies (Rieser and Moore, 2005). Furthermore, we identified form-function correlations which can inform the generation of clarification requests. These results fed into a prototype system named FRAGLE[2] that is able to request clarification on all levels of grounding (Rieser, 2004).

2.3 Learning a Multi-Modal Clarification Strategy

My current work aims to learn a multi-modal clarification strategy based on features in the information state that maximises task success and user satisfaction. Like previous rule-based systems that dealt with clarification (Schlangen, 2004; Purver, 2004), we assume that speakers and listeners ground their utterances on several levels of understanding. But instead of defining the clarification strategy deterministically, we aim to learn the action with the highest reward by applying reinforcement learning. In contrast to decision theoretic models, reinforcement learning incorporates the possibility of "delayed rewards", i.e. to sacrifice short-term gains for greater long-term gains. This property is especially interesting with respect to clarification subdialogues, as they are considered to have a high immediate cost. To boot-strap an initial system, data on multi-modal interaction was gathered within the MP3 domain. For data collection we used a modified Wizard-of-Oz setting developed by the TALK project.[3] Active learning will be used to find an initial strategy based on the behaviour of six wizards. Rein-

[2] FRAGLE stands for fragmentary clarifications on several levels

[3] www.talk-project.org

forcement learning will help to optimise the strategy applying a statistical user model. The work will contribute to design robust mechanisms for handling interpretation uncertainty and generating multi-modal feedback.

3 Challenges in Spoken Dialog Systems Research

Despite the obvious need for robust mechanisms to handle poor results from speech recognition and other kinds of understanding problems, the following problems tend to re-occur while working on dialogue systems.

Does anthropomorphism help? Anthropomorphismwas considered to be problematic for spoken dialogue systems as the system pretend to have human capabilities, (Shneiderman, 1998). In my experience, making a system more human-like increases its acceptance. People like to interact with dialogue systems that are flattering and people tend to be polite to systems by themselves. Furthermore, human-like behaviour is a familiar interaction paradigm to users.

How to define usability? The PARADISE method (Walker et al., 1997) is a widely accepted framework to measure usability. However, the definition of user satisfaction as function of task success and dialogue costs seems to be problematic. Studies have shown that user satisfaction does not correlate with task completion times (Williams and Young, 2004) and the perceived task success depends on different error handling strategies (Skantze, 2003).

How to determine a strategy for multi-modal output planning? The claim by (Cohen and Oviatt, 1995) that speech is the primary input mode can be confirmed by results from Wizard-of-Oz studies undertaken by the TALK project. The SACTI-2 corpus and the MP3 corpus both show that users rarely take advantage of a click interface. But multi-modal input behaviour increases with noise as introduced by speech recognition and the wizard can encourage the user to act multi-modally (Schatzmann, 2004). Multi-modal utterances are more likely on the side of the wizard especially when handling interpretation ambiguities. In sum, multi-modal output generation seems to be especially promising in the presence of uncertainty.

Acknowledgements

The author would like thank Ivana Kruijff-Korbayová, Oliver Lemon and all the people from the TALK project for help and discussion.

References

N. O. Bernsen, H. Dybkjær, and L. Dybkjær, 1998. *Designing Interactive Speech Systems — From First Ideas to User Testing*. Springer, Berlin, Heidelberg, New York.

T. Brey, G. Hanrieder, P. Heisterkamp, L. Hitzenberger, and P. Regel-Brietzmann. 2000. *Issues in the Evaluation of Spoken Dialogue Systems - Experience from the ACCeSS Project*. In Proc. of LREC.

P. Cohen and S. Oviatt. 1995. *The Role of Voice-Input for Human-Machine Communication*. Proc. of the National Academy of Sciences, 92(22), 9921-9927, Washington DC.

M. Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London.

J. Schatzmann. 2004. *The SACTI-2 Data Collection: Observations and Analysis*.

D. Schlangen. 2004. *Causes and Strategies for Requestion Clarification in Dialogue*. Proc. of the 5th SIGdial Workshop on Discourse and Dialogue.

B. Shneiderman. 1998. *Designing the User Interface*. Addison Wesley Publishing.

G. Skantze. 2003. *Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems*. ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems.

V. Rieser and J. Moore. 2005. *Implications for Generating Clarification Requests in Task-oriented Dialogues*.

V. Rieser. 2004. *Fragmentary Clarifications on Several Levels for Robust Dialogue Systems*. MSc thesis, School of Informatics, University of Edinburgh.

V. Rieser. 2003. *Design and Implementation of a Speech Interface for a Washing Machine*. MA thesis, Department of Information Science, University of Regensburg.

M. A. Marylin A. Walker and D. J. Litman and C. A. Kamm and A. Abella. 1997. *PARADISE: A general Framework for evaluating Spoken Dialogue Agents*. Proc. of the 35th Annual General Meeting of the ACL/EACL.

J. D. Williams and S. Young. 2004. *Characterizing Task-Oriented Dialog using a simulated ASR Channel*. Proceedings of the ICSLP.

Biographical Sketch

Since November 2004 the author is a Ph.D. candidate at the International Graduate College in Saarbrücken, Germany. Her thesis is supervised by Dr. Oliver Lemon (School of Informatics, Edinburgh) and Prof. Manfred Pinkal (Computational Linguistics, Saarbrücken). She received the MSc in Informatics in "Language Engineering" from the School of Informatics, Edinburgh. Before she took a MA in Information Science and Linguistics at the University of Regensburg, Germany. During her studies she concentrated on spoken human-computer interaction and usability evaluation.

Antonio Roque

USC - Institute for Creative Technologies
13274 Fiji Way, Suite 600
Marina del Rey, CA 90292

roque@ict.usc.edu
http://www-scf.usc.edu/~aroque/

1 Research Interests

My interests are in **information-state approaches** to dialogue management, **automatic generation** of dialogue managers, and **evaluation**.

2 Past, Current and Future Work

Currently I am building dialogue managers for speech-enabled radio agents at the University of Southern California's Institute for Creative Technologies. The domain is military, in which forward observers are trained to perform artillery strike requests over the radio. Recently, VR simulators have been used for this training, but so far the trainees have been communicating over the radio with human trainers. The goal of this project is to build dialogue agents to handle the radio communication tasks, leaving the human trainers free for assessment or to manage multiple simultaneous training sessions. I am currently investigating the relative effectiveness of dialogue systems driven by human-authored Finite State Machines, automatically generated FSMs, and human-authored Information State dialogue agents.

Prior to this I worked with the Why2-Atlas project at the University of Pittsburgh. This was an natural language intelligent tutoring system which analyzed, in real-time, brief essays about qualitative college-level physics questions, and then held tutoring dialogues to correct misconceptions and lead the student to add missing information to their essays.

As an undergraduate I was interested in literary theory, but moved away from that in favor of disciplines that take a more direct approach to solving real-world problems. I enjoy building dialogue systems that solve such problems, and hope to continue doing so in the future, while still being able to work on the larger theoretical issues involved.

3 Challenges in Spoken Dialogue Systems Research

Dialogue research, like AI in general, is both an engineering discipline and a science. Because of this it offers its practitioners the ability to work on three levels over the course of a career: solving problems at hand; contributing to the dialogue community by addressing questions that it needs to have resolved; and contributing to the development of the sciences of language and cognition. If one wishes to work in all three, how does one choose the problems to attack over the course of a career?

Secondly, what practical techniques do we have for evaluating dialogue systems, and dialogue managers in particular? We have several traditions to call from. We can run experiments as done in, for example, psychology of education, with multiple concurrent human participants testing full systems, with control groups and measures such as time to task completion, quality of task completion, or learning gain. But this requires a large number of subjects, which is complicated when the domain requires specially-trained subjects such as those with basic medical or military skills; furthermore, it is not very fine-grained, requiring the testing of an entire system (unless variations of the systems are tested, which increases the number of subjects required.) We can use corpus approaches, as in statistical natural language processing, to calculate accuracy and efficiency measures, but there are questions to this, such as what it means for a dialogue system to be accurate, and how we can model dialogues to test the systems against. For researchers to be able to compare dialogue systems to one another, some further understanding of typical approaches needs to be developed.

References

Carolyn P. Rose, Dumisizwe Bhembe, Antonio Roque, Stephanie Siler, Ramesh Srivastava, Kurt VanLehn. "A Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals." Intelligent Tutoring Systems Conference, June 2-8 2002.

Carolyn P. Rose, Dumisizwe Bhembe, Antonio Roque, Kurt VanLehn. "A Hybrid Text Classification Approach for Analysis of Student Essays." Proceedings of the HLT-NAACL 03 Workshop on Educational Applications of NLP.

Kurt VanLehn, Pamela W. Jordan, Carolyn Penstein Rose, Dumisizwe Bhembe, Michael Bottner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Michael A. Ringenberg, Antonio Roque, Stephanie Siler, Ramesh Srivastava. "The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing." Intelligent Tutoring Systems Conference. Biarritz/San Sebastian, June 2-8 2002.

Biographical Sketch

Antonio began his CS PhD studies in January 2004 at the University of Southern California. He received his undergraduate degree from the University of Michigan in English with a concentration in Creative Writing, and drifted into a programming and research career during the late 90s.

Mihai Rotaru

University of Pittsburgh
5420 Sennott Hall
210 S. Bouquet St.
Pittsburgh, PA 15217, USA

mrotaru@cs.pitt.edu
www.cs.pitt.edu/~mrotaru

1 Research Interests

My research interests are focused on improving the quality of human-computer interaction via spoken or multi-modal dialogue interfaces. Specifically, I am interested in **task/dialogue design** and **dialogue management**.

2 Past, Current and Future Work

My research so far has focused on understanding the effect of user emotions on the dialogue flow, more specifically in tutoring dialogues. There are two ingredients to this problem. First, we need to reliably detect user emotions. In (Rotaru and Litman, 2005a) we show that using word level features instead of more commonly used turn level feature helps in emotion prediction. Second, we need to examine the interaction between user emotions and various dialogue phenomena. In (Rotaru and Litman, 2005b), we show that system's rejections of the current user turn are followed by more emotional user responses. Surprisingly, in our data, we find no relationship between user emotions and recognition problems within a turn nor between previous turn user emotions and current turn recognition problems.

My current research is focused on understanding what characterizes a high quality dialogue between a human and a computer and how this knowledge can be used for better task/dialogue design. I believe that the answer depends on the characteristics of the underlying domain. In my research I am interested in three types of domains: information access domains, large scale domains and intelligent tutoring.

In information access domains (e.g. air travel domain), the user knows the characteristics of the items of interest and gains access to such items via interaction with the system. Typically, the task structure is relatively simple: the system has to acquire user's constraints and retrieve the items that satisfy those constraints (sometimes a negotiation phase is also present). Thus, critical to the success of the interaction in these domains is the system's robustness to channel errors. Various recovery strategies have been proposed: addressing the error immediately (Bohus 2005), fallback to speak-and-spell (Filisko and Seneff, 2004), move on and approach the problematic information from a different task perspective (Skantze, 2003). There is little work on recovery strategies from multiple subsequent errors. In such cases, the users are forced in specific subtasks with little or no means to recover except for restarting the interaction. Strategies that are able to reuse part of the information exchanged can speed up recovery in such cases. I am interested in studying such strategies.

I think that a key step to higher quality dialogue systems in information access domains is our understanding of the effects of these recovery strategies. Also, understanding the degree of domain independence of these strategies is an important issue. Once these issues are addressed, these strategies can come become integral part of generic dialogue managers.

In large scale domains (e.g. the real estate domain), besides the robustness to channel errors, there is another key ingredient: the system's ability to assist user in finding the item(s) of interest. This includes helping the user navigate the domain, providing relevant information, assisting the user in his decision process, etc.

During exploration of large scale domains, users frequently find themselves lost in the data space with no clear direction of what to do next. Also, given the complexity of such domains, the exploration history might fade in user's mind. To handle these issues users employ a variety of techniques like using a reference point in their exploration or exploring using a predefined exploration strategy (which might be devised after several failed attempts to browse the data space). The process of exploration is an important user activity and allows the user to become familiar with the data space. Familiarity is associated with user building a "cognitive map" of the domain. Once the user is familiar with the domain, he can use his expertise for a variety of decision tasks. For example, in the real-estate domain, once a user has become familiar with the area, he can proceed to choosing a suitable house.

My proposal is to build a navigation model that attempts to discover how the user explores the data space. This approach is in contrast with current approaches where the system has a plan and the user has to follow it. In other words, instead of the user following the system's plan, the system follows user's (exploration) plan. Having an accurate representation of the user's plan has application for many tasks in these domains. For exam-

ple, a graphical rendition of the navigation model can be presented to the user in multimodal interfaces in this way possibly scaffolding the construction of user's cognitive map and increasing the chances of success in exploration tasks. Knowing the user's plan and his exploration patterns can lead to better recommendations for query relaxation and better summarization of subspaces in the data space. It can also inform the system about relevant information that the user is expecting (e.g. if the user's exploration plan is to ask about the schools and golf courses in a city and then move to finding houses in that city, whenever the user explores a new city, school and golf course information can be provided without waiting of user's request).

Dialogue-based tutoring can be viewed as a guided navigation of the problem's solution. In contrast to guided navigation of a task like in information access domains, there is flexibility in the order in which the task is performed. The flexibility is responsible for good/intuitive tutoring (e.g. combine backward chaining with forward chaining) or for an awkward or hard to understand tutoring (e.g. using only forward chaining). Moreover, current research in my group (Forbes-Riley et al., 2005) shows that in human-human tutoring, student attempts to drive the dialogue by introducing new concepts correlates with learning. I am currently exploring whether the navigation model proposal for large scale domains can be adapted to the tutoring domain and if it will result in more flexible and student-controlled tutoring dialogues.

Finally, I am also interested in how techniques from the Recommendation Systems field can be applied to improve the quality of current dialogue systems. The main idea is to use information from previous interactions with other (similar) users to drive the conversation with the current user. Specifically, the navigation model can benefit a lot from this information source.

3 Challenges in Spoken Dialog Systems Research

I believe that an important issue in the (spoken) dialogue systems community at this point is the lack of standards and tools sharing/availability. To draw a parallel, a key to the success to graphical interfaces has been the development of standards and middleware components. For example, in any graphical application users expect a help menu item or that they can access (sometimes context-sensitive) help by pressing F1. In the same way, one can imagine generic dialogue commands that every dialogue system should implement (like help, cancel, restart, etc). This will also have a positive effect on users since it will educate them about what things they can do when they are in trouble. As a side effect of the lack of standards, the amount of reuse and availability of dialogue tools and components is very limited.

In terms of robustness to channel errors, most systems employ very few recovery strategies. I think that an important challenge will be to understand how radically different recovery strategies (confirmations, speak-to-spell, restarts) can be combined in the same system to increase the task robustness.

Finally, I think that methodologies and tools for semi-supervised offline analysis of corpora collected from previous system runs can help improve existing dialogue systems in an incremental fashion.

References

D. Bohus and A. Rudnicky. 2005. *Sorry, I Didn't Catch That! - An Investigation of Non-understanding Errors and Recovery Strategies,* To appear in SIGdial.

E. Filisko and S. Seneff. 2004. *Error Detection and Recovery in Spoken Dialogue Systems*. HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing

K. Forbes-Riley, D. Litman, A. Huettner and A. Ward. 2005. *Dialogue-Learning Correlations in Spoken Dialogue Tutoring,* To appear in International Conference on Artificial Intelligence Education (AIED).

M. Rotaru and D. J. Litman. 2005a. *Using Word-level Pitch Features to Better Predict Student Emotions during Spoken Tutoring Dialogues*, To appear in European Conference on Speech Communication and Technology (Interspeech-2005/Eurospeech).

M. Rotaru and D. J. Litman. 2005b. *Interactions between Speech Recognition Problems and User Emotions,* To appear in European Conference on Speech Communication and Technology (Interspeech-2005/Eurospeech).

G. Skantze. 2003. *Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems*, ISCA Workshop on Error Handling in Spoken Dialogue Systems.

Biographical Sketch

Mihai Rotaru is currently a 4th year Ph. D. student in the Department of Computer Science, University of Pittsburgh, USA. He works under the supervision of Dr. Diane J. Litman. He was born in Romania and received his B.S and M.Sc. from West University, Timisoara, Romania. His extra curricular interests include good movies, traveling, cool gadgets, friends and (last but not the least) good beer. These interests often overlap with his academic activities at conferences (especially during evenings).

Jost Schatzmann

Cambridge University
Fallside Lab, Engineering Department
Trumpington Street, Cambridge CB21PZ
United Kingdom

[js532@cam.ac.uk]
[http://mi.eng.cam.ac.uk/~js532]

1 Research Interests

My research interests lie in the areas of user modeling and statistical approaches to spoken dialogue systems. I am particularly interested in systems that learn from experience what constitutes a good dialogue strategy and I am interested in developing user simulation tools for training such systems.

2 Past, Current and Future Work

2.1 Overview of Current Work

My work is part of the growing field of research on applying reinforcement-learning techniques to dialogue management design. The main motivation driving research in this field is the hope to learn dialogue strategies from data rather than having to rely on handcrafted rules. However, it is very rarely the case that enough data is available to sufficiently explore the vast space of possible dialogue states and strategies and it is not guaranteed that the truly optimal strategy is indeed present in the training corpus.

I am interested in approaching these problems with the help of user simulation tools. The simulated user (typically modeled on the abstract intention-level rather than the word- or acoustic-level) allows us to generate any number of training episodes. It also enables us to explore strategies which are not present in the training corpus, so that new and potentially better strategies can be found.

The feasibility of simulation-based learning has been shown by a number of research groups, and various user simulation techniques have been presented in the literature (Levin et al, 2000, Scheffler and Young, 2002; Pietquin, 2004). Yet, the approach struggles to find broad acceptance in the field because the quality of the simulated dialogues is often poor and the robustness of the learned strategies is uncertain.

2.2 Proposed Approach

I consider the current lack of solid user models and rigorous evaluation standards as major roadblocks to further progress on strategy learning. The goal of my PhD research is to introduce better evaluation methods and to develop user- and error-modeling techniques that enable us to learn strategies which outperform competing handcrafted strategies when tested on human users.

Over the past six month, I have been mainly working on evaluation techniques to improve the reliability of simulation-based strategy learning.

- Introduction of *simulation quality metrics* and evaluation of the state of the art in domain-independent stochastic user simulation (Schatzmann et al, 2005a)
- Investigation of the *effect of the user model on the learned strategy*, re-assessment of standard evaluation practices for testing learned strategies and investigation of *user model-independent evaluation* techniques (Schatzmann et al, 2005b)

In the coming two years, I hope to work on the following projects:

- *Introduction of strategy confidence scores*, indicating how likely a learned strategy is to work well on human users. My motivation for work in this area is to enable the system designer to optimize a strategy not only with respect to some reward function but also with respect to its expected reliability.
- *Markov Decision Process (MDP)-based user models*. MDPs have been used for building dialogue managers and I am interested in applying the framework to user modeling. I intend to use *Agendas* (Xu and Rudnicky, 2000) to represent states, as these naturally combine the user goal and the user state, encode dialogue history and priority of user actions.
- *Error Generation*. I am interested in learning strategies under system specific error-conditions and I plan to compare simulated errors with errors generated by feeding synthesized acoustic-level user output into the recognition and understanding components of our system.
- *Empirical Evaluation.* To conclude my PhD project, I want to test simulation quality and strategy performance using human users.

2.3 Results and Previous Work

Results of my work on simulation-based strategy learning can be found in (Schatzmann et al, 2005a) and (Schatzmann et al, 2005b). Prior to research in this area, I worked on multimodal systems and investigated how gestures can be used to constrain the speech recognizer in order to improve performance. (Schatzmann, 2004)

3 Challenges in Spoken Dialog Systems Research

3.1 Uncertainty in Dialogue Management

A key challenge for SDS research exists in finding dialogue models that support the notion of uncertainty. In current systems, a dialogue strategy is essentially a mapping from dialogue states to system actions. Realistically however, the true state of the dialogue is never known. Since spoken human-computer dialogue always involves a noisy communication channel, the dialogue manager can never be certain that the recognized input is actually correct. I am interested in current work by (Williams et al., 2005) on Partially-Observable Markov Decision Processes (POMDPs) which maintain a distribution over dialogue states rather than a single current state. POMDPs are much harder to compute than current MDP models and their use in large-scale systems is an interesting challenge for SDS research.

3.2 Modeling and Inference of User Goals

Research on spoken dialogue systems is often justified by claims that speech is the "most natural" or "most efficient" form of interface. However, traditional input devices have become so familiar that many users find it quicker, more intuitive and less error-prone to use the mouse or keyboard. The majority of users is reluctant to change to a voice interface if it delivers no real efficiency gain or additional functionality.

A key challenge for future research is thus to identify how spoken dialogue can indeed outperform competing forms of interfaces. A starting point could be to reconsider the advantages speech has over keyboard or mouse-based input. One of these is that speech can transmit compact, high-level goals in place of several specific low-level commands. This is frequently used in human-human communication, for example by saying "Cancel today's group meeting and inform all participants by email!". Current computer systems require us to break up such a sentence into atomic instructions: "Open calendar! Go to current date! Select group meeting...". At this level, each spoken command is equivalent to a mouse-click, meaning that there is no performance gain in using speech. But whereas it is impossible to replace a large number of sequences of mouse-clicks or keystrokes with a single click or a single stroke, this should one day be possible for speech. Research on modeling, inferring and decomposing user goals is a great challenge, but it may offer speech interfaces a "unique selling proposition" for competing with traditional interfaces.

References

Levin, E., et al. (2000). *A Stochastic model of human-machine interaction for learning dialogue strategies.* IEEE Transactions on Speech and Audio Processing, 8(1):11-23.

Pietquin, O. (2004). *A Framework for Unsupervised learning of Dialogue Strategies*. PhD Thesis, Faculte Polytechnique de Mons.

Schatzmann, J., K. Georgila, and S. Young. (2005a). *Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems*. Proc. of SIGdial 2005, Lisbon, Portugal

Schatzmann, J., M. Stuttle, K. Weilhammer, S. Young. (2005b). *Effects of the User Model on Simulation-based Learning of Dialogue Strategies*. Proc. of ASRU 2005, Cancun, Mexico (submitted)

Schatzmann, J. (2004). *Speech Recognition in a Multimodal Context*. MPhil Thesis, Cambridge University

Scheffler, K. and S. Young. (2002). *Automatic Learning of Dialogue Strategy using Dialogue Simulation and Reinforcement Learning*. Proceedings of HLT, San Diego, California

Williams, J. D., P. Poupart, and S. Young. (2005) *Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management.* Proceedings of SIGdial 2005, Lisbon, Portugal

X. Wu and A. Rudnicky. (2000) *Task-based Dialogue Management using an Agenda*. ANLP/NAACL Workshop on Conversational Systems, Seattle, WA

Biographical Sketch

Jost Schatzmann is a PhD student working for Prof. Steve Young at the University of Cambridge. Prior to his PhD studies, Jost completed his undergraduate degree at Imperial College, London, and a Masters degree at Cambridge University. He is loosely affiliated with the EU TALK Project and his studies are funded by the Bill Gates Cambridge Trust. Jost is also a member of Darwin College and spends most of his free time rowing for the Darwin College Boat Club.

Gabriel Skantze

Dept. for Speech Music and Hearing, KTH
Lindstedtsvägen 24
100 44 Stockholm
Sweden

[gabriel@speech.kth.se]
[www.speech.kth.se/~gabriel/]

1 Research Interests

My research interests lie generally in the area of **error handling in spoken dialogue systems**. I am trying to integrate techniques for **dialogue modelling**, **robust interpretation**, **error detection**, **error recovery** and **grounding** into a complete system and test it with users.

2 Past, Current and Future Work

One of the greatest challenges when building spoken dialogue systems (SDS) is to deal with error and miscommunication. Errors may arise from several sources, but the most common source is often the speech recogniser. In conversational dialogue systems, where the user may speak relatively freely, one should always expect a certain amount of errors from the speech recogniser. However, natural human conversation has built-in mechanisms for handling miscommunication. Conversational language may also contain redundancies that allow some loss of data. If these mechanisms are understood correctly, other components in a SDS may be developed to utilize techniques for detecting, ignoring and repairing errors, and fruitful dialogues may still be carried out. My work is aiming at investigating such techniques, as well as modelling the way humans handle miscommunication.

2.1 Exploring human error-handling

Initially, I conducted an experiment similar to Map Task, where humans gave directions to each other, but where a speech recogniser was used in one direction to introduce errors (Skantze, 2005). The data were analysed to find out which strategies humans employ to recover from total non-understanding. While the WER was fairly high, the subjects reported in post-interviews that they were almost always understood. The analysis showed that humans signal non-understanding only after about 30% of the actual cases of non-understanding. This is different from the behaviour in most SDS, where the default strategy after non-understanding is to signal non-understanding, by for example requesting the user to repeat or rephrase. Instead of focusing on the previous utterance, the subjects in this experiment focused on the task and asked task-related questions, which helped recovering from the problem. Such behaviour significantly helped recovering from non-understanding. A regression analysis of the user's experience of task success showed that signalling non-understanding decreased their rating, while the actual number of non-understanding had no such effect.

Another observation from the experiments was that humans were extremely good at early error detection, i.e. detection of errors in the speech recognition results. This lead to a very low number of misunderstandings. Early error detection is something that current SDS are not very good at, resulting in a lot of verification subdialogs or misunderstandings. To investigate which factors the subjects benefited from when detecting errors, a new set of subjects were given the task to go through the dialogues and mark which words in the speech recognition results they believed were incorrect, with access to varying amount of information from context, word confidence scores and 5-best lists (Skantze and Edlund, 2004a). The results showed that both word confidence scores and 5-best lists had a positive effect. However, only the immediately previous system utterance improved the performance; longer context had no effect. Different machine learning algorithms were also tested for the task (ibid.).

2.2 The HIGGINS spoken dialogue system

A spoken dialogue system, called HIGGINS, is now under development as a test bed for exploring error handling techniques (Edlund et al., 2004). The initial domain for the system is similar to the one in the experiment mentioned above: pedestrian navigation. The user is moving in a 3D virtual city. The system's task is to guide the user to a specific goal. The system has no access to the user's position but has to rely on the user's description of the environment. The domain is challenging in that robust interpretation and error awareness needs to be combined with deep semantic analysis of longer, more complex, user utterances, containing a lot of referring expressions.

The initial effort has been put on developing a robust interpreter, called PICKERING, for spoken language with deep semantic structures. It allows insertions and non-agreement inside phrases, and combines partial results to return a limited list of semantically distinct solutions. A preliminary evaluation has shown that the

interpreter performs well under error conditions, and that the built-in robustness techniques contribute to this performance (Skantze and Edlund, 2004).

2.3 Concept level grounding and error handling

Most dialogue systems have the capability to explicitly and implicitly verify what the user say. These verifications are often realised as complete turns and address whole user utterances. A challenging issue is to clarify and give feedback on (i.e. ground) the system's understanding of parts of the user utterances, based on the system's confidence in the individual words or concepts. To make the dialogue more natural and efficient, the system should be capable of producing clarification ellipsis, implicitly ground referents by carefully selecting the system's realisation of referring expressions, as well as interpreting and integrating the user's responses to such grounding behaviour. To do this, a discourse modeller, called GALATEA, has been developed and integrated in HIGGINS, which tracks the grounding status of each individual concept that has been contributed to the discourse. This grounding status is based on concept confidence scores in the deep semantic structures from the robust interpreter, as well as the history of who has said what and when. The grounding status is then used for selecting grounding and clarification strategies, as well as removing erroneous concepts from the system's model, when problems are detected.

2.4 Future work

The next step is to conduct user experiments with the complete system to evaluate the techniques mentioned above, as well as testing the error recovery strategies the subjects used in the experiment mentioned in 2.1. Also, the way users signal problems after concept level grounding and clarification will be studied and techniques for detecting them will be tested.

3 Challenges in Spoken Dialog Systems Research

Effective communication between humans often involves the use of fragmentary expressions, such as clarification and grounding ellipsis, conversational grunts, etc. I believe that such behaviour is one of the keys to efficient conversational communication, not least in the handling of errors and miscommunication. However, current spoken dialogue system cannot handle them very well. Modelling such behaviour requires better understanding and modelling of pragmatic and a prosodic phenomena, and I think it is an important topic for future research. As we build such behaviour into dialogue systems, users will probably make use of it to a greater extent.

Another aspect of natural language that makes it useful is the capabilities to express complex relations and negations, something that may be very hard to do in graphical interfaces. I believe that such expressions are very important to handle as well, if speech is to succeed as a modality for human-computer interaction, in other areas than those where our eyes and hands are busy.

References

Jens Edlund, Gabriel Skantze and Rolf Carlson. 2004. Higgins - a spoken dialogue system for investigating error handling techniques. In *Proceedings of ICSLP 2004*, 229-231. Jeju, Korea

Gabriel Skantze and Jens Edlund. 2004a. Early error detection on word level. In *Proceedings of ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction 2004*. Norwich, UK.

Gabriel Skantze and Jens Edlund. 2004b. Robust interpretation in the HIGGINS spoken dialogue system. In *Proceedings of ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction 2004*. Norwich, UK.

Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3), 207-359.

Biographical Sketch

Gabriel Skantze was born July 13, 1975 in Stockholm. After a few years, his parents moved to Karlskrona, a small town in the southern part of Sweden, where he grew up. After some initial academic studies in philosophy, he began studying cognitive science at Linköping University. There, he came in contact with their dialogue system research, working in the summers with implementation of dialogue systems. For the final term, he moved to Stockholm to do his master thesis at the Department for Speech, Music and Hearing, KTH, in the area of multimodal dialogue systems. After the receiving a Master's degree in 2000, he started to work as an application developer and user interface designer for the company PipeBeach (later bought up by HP), which developed a VoiceXML-platform. After 15 months, he returned to academia again. In 2002, he was accepted as a PhD student at KTH, under the supervision of professor Rolf Carlson, partly funded by the Graduate School for Language Technology – a national program for integrating research on speech and language processing.

Matt Stuttle

Department of Engineering
Cambridge University
Trumpington Street
Cambridge CB2 1PZ, UK

mns25@eng.cam.ac.uk
http://mi.eng.cam.ac.uk/~mns25 (delete if n/a)

1 Research Interests

I have several interests in the field of speech technology and human computer interaction.

Firstly, I'm interested in the area of **speech recognition and language understanding technology**, particularly their use in practical applications. These technologies are maturing, but many issues in their use and implementation remain open. Particularly of interest is how to utilise all the information from the parser and the recogniser to leverage performance and implement new strategies in a dialogue system.

Related to this, I'm interested in **statistical dialog systems**, specifically methods of including the n-best or full recognition and parsing results into the recogniser. By this, it is possible to represent the errors in the output of the speech understanding components, and also the possibility of the user's goals changing over the course of the dialogue.

Finally, I'm also interested in the design and application of **multimodal dialog systems**. Adding a multimodal interface greatly increases the bandwidth of information available to both the user and the system. Furthermore, it could be possible to develop strategies for interaction and goal determination, as well as providing a new interface for clarification and disambiguation of intent.

2 Past, Current and Future Work

Prior to my current position as a researcher, I was a PhD student at Cambridge University as well. My work was focussed on acoustic modelling for speech recognition. More specifically, I was looking at alternative front-end parameterisations for speech recognition systems (Stuttle and Gales, 2002).

Currently I am working at Cambridge University on the FP6 project "TALK" (Talk and Look: tools for Ambient Linguist Knowledge). I have worked on a method for collecting dialogue data in a Wizard of Oz using a simulated ASR channel (Williams and Young, 2004). The simulated channel models typical confusions present in a real ASR channel, but is much faster to set up for a given domain, and can be set to yield an arbitrary word error rate. The data collected exhibited similar behaviour to that of human/computer dialogue (Stuttle et al., 2004). Two corpora have been collected using this setup, and the data is being used to train elements of dialog systems.

Together with Edinburgh University, I have developed a baseline system based on the same tourist information domain. The dialogue manager is written in DIPPER and we are working on getting novel update rules for strategies for fragmentary clarifications based on n-best parsing.

I am also working on developing and maintaining the ATK (Application Tool Kit) open-source speech recognition system (Young, 2005). Current work has lead to the inclusion of n-best lists and the use of multiple simultaneous decoders in the recognition. Future development will see the implementation of improved word confidence metrics and word posterior lists (Evermann and Woodland, 2000). In addition, we arere working on a statistical parser (the Hidden Vector State mode) to incorporate into the ATK framework as a complete speech understanding component (He and Young, 2005). The overall aim is to incorporate these features into a multimodal dialogue system in the tourist information domain. The dialogue manager will be the Hidden Information State model, which enforces a hierachical structure onto the information/dialogue state, and effectively maintains an n-best approximation to the complete belief state.

3 Challenges in Spoken Dialog Systems Research

- At present, speech is an established technology afor certain tasks and domains. However, there are still a number of issues when running full open recognition, particularly when "boot-strapping" recognition.
- Our implementation of a multimodal Wizard of Oz experiment found users unwilling to use two modes of interaction in parallel when information seeking. What are issues in developing multimodal dialogue systems, and what tasks or strategies are they appropriate to? Where can the increased functionality help?
- How can we leverage all the information from the

speech recognition including word confidences and n-best results for a dialogue system?

- Currently reinforcement learning approaches can perform well in terms of increasing the return of a reward function for a given task. However, the exact nature of these reward functions remains an open question. What are useful values for rewards in dialogues, and can they be standardised for all users?

References

G. Evermann and P.C. Woodland. 2000. Large vocabulary decoding and confidence estimation using word posterior probabilities. *Proceedings ICASSP*.

Y. He and S.J. Young. 2005. Semantic processing using the hidden vector state model. *Computer Speech and Language*, 19(1):85–106.

M. Stuttle and M.J.F. Gales. 2002. Combining a Gaussian mixture model front end with MFCC parameters. In *Proceedings ICSLP*, pages 1565–1568.

M.N. Stuttle, J.D. Williams, and S.J. Young. 2004. A Framework for Dialog Systems Data Collection using a Simulated ASR Channel. In *Proc. ICSLP*.

J.D. Williams and S.J. Young. 2004. Characterizing Task-Oriented Human-Human Dialog using a Simulated ASR Channel. In *Proc. ICSLP*.

S. Young. 2005. *The ATK Book, Version 1.5*. Cambridge University Engineering Department.

Biographical Sketch

Matt Stuttle was grew up in Weymouth on the south coast of the UK. He studied Electronic Engineering as an undergraduate at the University of Southampton. After that, he completed a PhD with the Machine Intelligence group at Cambridge University, working on alternative front ends for speech recognition.

After a short period working in London for Sony Computer Entertainment Europe, Matt managed to break the lab record for returning to the research group from industry, and is currently an RA there. He is working on EU FP6 project TALK (Talk and Look: tools for Ambient Linguist Knowledge). His main interests are: the implementation of speech recognition, novel human computer interfaces and machine learning in dialogue.

Outside of the university, Matt enjoys running (he'll be running a half-marathon immediately after Eurospeech) and hashing in particular. He is a sporadic kitesurfer and a fan of most sports which have the suffix "-boarding".

Stefanie Tomko

Carnegie Mellon University
Language Technologies Institute
5000 Forbes Avenue, 4502 NSH
Pittsburgh PA 15217 USA

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

1 Research Interests

My research interests focus on the **usability** of and **error-handling** in spoken dialog systems. I am particularly interested in the potential of **restricted- or subset-language interfaces** and how such systems can be made more palatable to users. I am also interested in the use of **multimodal interfaces** in different conditions and what their effect is on how well users work with unimodal applications.

2 Past, Current and Future Work

To date, my research work has centered on the Speech Graffiti (*a.k.a.* Universal Speech Interface) project. This project is an attempt to create a standardized interaction style for speech communication between humans and simple machines. Our belief is that by implementing and promoting a standard interface, speech applications can become more usable and desirable tools. The interface that we have been developing in this project comprises a set of keywords and interaction guidelines that users learn in order to enable them to explore and use any application that is designed using the Speech Graffiti standard.

One of the main design principles for Speech Graffiti is that it should be more restricted than natural, conversational language, yet less restricted than typical application-specific command-and-control interfaces. The proposed benefits of such a design are

- less complex grammars and vocabularies can be used, resulting in lower speech recognition error rates;
- the overall system is less complex than for natural language interfaces, which allows the system to be used in small devices and supports the creation of application-generating toolkits;
- using a specific speaking style encourages the user to view the system as a tool, which should make the user less likely to overestimate its capabilities.

In my master's thesis work, I reported on a user study comparing Speech Graffiti to a natural language speech interface (using the same telephone-based, movie information database backend) and found that Speech Graffiti users had higher levels of user satisfaction, lower task completion times, and similar task completion rates (Tomko, 2003). Such benefits come with a lower overall system development cost, since a toolkit is available to facilitate the development of new Speech Graffiti applications (Toth *et al*, 2002). I also found that task success and user satisfaction with Speech Graffiti were significantly correlated with how often users spoke within the grammar (*grammaticality*) (Tomko & Rosenfeld, 2004). This indicates that it is important to help users speak within the grammatical bounds of spoken dialog systems, particularly subset language ones. Based on the experience of users in this study, 80% grammaticality appears to be a reasonable preliminary goal for effective interaction. Nearly all participants with Speech Graffiti grammaticality scores over 80% gave positive user satisfaction scores, and more than half of the participants in this study achieved this level. Furthermore, users with grammaticality above 80% completed an average of 6.9 tasks, while users with grammaticality below 80% completed an average of only 3.5 tasks.

However, even though they had completed a tutorial session prior to the study, some users had difficulty speaking within the Speech Graffiti language. The experiences of the six out of 23 participants who preferred the natural language system are illustrations of frustrating communication. In the Speech Graffiti interactions, they accounted for the six highest word- and concept-error rates, the six lowest task completion rates, and the four lowest grammaticality rates. One defining characteristic of these six participants was that all but one of them belonged to the group of thirteen study participants who did not have computer programming backgrounds.

My current work is designed with these users in mind. The previous study indicated that even if users appeared to "get" the language during a pre-use tutorial session, they often forgot crucial aspects of the style during the actual interaction. I am now working to enhance the project with a *shaping* component: for any non-Speech-Graffiti-grammatical input, the system will attempt to provide feedback which will encourage the user to speak within the Speech Graffiti grammar. The baseline level for such feedback will involve simply confirming users' non-Speech-Graffiti-grammatical

input with the equivalent Speech Graffiti form. Thus, if a user were to tell the system *I want to know what's playing at the Manor Theater tomorrow*, the system will respond with the Speech Graffiti-grammatical confirmation `theater Manor, day tomorrow, what movies?` before providing the query result.

This approach to increasing interaction proficiency has three main components. First, an *expanded grammar* allows the system to accept more natural language input than is allowed by the canonical Speech Graffiti language. I hypothesize that the use of the expanded grammar will reduce training time and allow the system to be more forgiving for novice users, which should increase user satisfaction. Separate language models will be built based on the target and expanded grammars so that each user input will be decoded twice. Second, *shaping confirmation* will provide the appropriate shaping response for non-Speech Graffiti input that is accepted by the expanded grammar. Finally, an error classification and response strategy will provide context-appropriate, *shaping help* for situations in which the recognized input string is accepted by neither the target nor the expanded grammars.

I believe that by receiving this type of feedback, one-time users of the system will be able to complete tasks with very little training. As people use the system more frequently, the shaping feedback can help them learn a more efficient style for interacting with the system, without having to go through a separate training process. The use of both a more natural language and a restricted language in the same system will support the investigation of the following research question: *How malleable are users? That is, given the option of speaking with more natural language, will users allow their input to be shaped to a more efficient style?*

3 Challenges in Spoken Dialog Systems Research

Spoken dialog systems are the means through which users experience speech technology. Because this technology is imperfect, and is likely to be so for the next several years, spoken dialog systems must be designed to handle gracefully the problems rooted in the imperfections of the technology. Therefore, I think two of the greatest challenges for spoken dialog systems research are error handling and evaluation.

Error handling is crucial to making interactions seamless, effective, pleasant, and natural for users. I think one interesting challenge for error handling is determining what levels of transparency should be used in presenting error situations to users. For instance, in which cases is repeating the exact string of recognized words back to the user helpful, as opposed to confusing or frustrating? Are there clear distinctions between situations when users should be made aware of errors, and those when a dialog system should try to work at correcting an error in a less explicit way?

Evaluation of spoken dialog systems is of crucial importance to the future of such systems because it is important to take potential real-world success into account when designing user interfaces. Although usability models (*e.g.* PARADISE) have been devised and "user satisfaction" obviously carries substantial weight in the evaluation of a system's success, I think that it is still not exactly clear how the components of user satisfaction contribute differently under varying conditions. I think it is important to devise user testing scenarios that more accurately reflect and capture the expereinces, needs, and perceptions of users in real-world sutations.

References

Tomko, S. "Speech Graffiti: Assessing the User Experience." *CMU LTI Tech Report CMU-LTI-04-185*, 2004. www.cs.cmu.edu/~stef/papers/mthesis.ps

Tomko, S. and Rosenfeld, R. 2004. "Speech Graffiti habitability: What do users really say?" In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA, pp. 81-84.

Toth, A., Harris, T., Sanders, J., Shriver, S., and Rosenfeld, R. 2002. "Towards Every-Citizen's Speech Interface: An Application Generator for Speech Interfaces to Databases." In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.

Biographical Sketch

Stefanie Tomko is a Ph.D. candidate working with Dr. Roni Rosenfeld at the Language Technologies Institute in Carnegie Mellon University's School of Computer Science. She received a Master's degree from the same department and holds bachelor's degrees in Linguistics and English from the University of Washington. Stefanie is active in the Women@SCS group at Carnegie Mellon, and is particularly involved with the Roadshow unit of that group, which makes presentations to girls and young women to encourage them to see computer science as an intriguing and challenging field to pursue. When she's not working, she enjoys running marathons, hiking, traveling, and entering contests of all kinds.

Keith Vertanen

University of Cambridge
Inference Group, Cavendish Laboratory
Madingley Road, Cambridge, CB3 0HE
United Kingdom

kv227@cam.ac.uk
www.inference.phy.cam.ac.uk/kv227/

1 Research Interests

My research interest is in spoken and multimodal interfaces for the creation and editing of text. In particular, I am interested in interfaces for the **correction of speech dictation** which are faster, less frustrating, and more accessible than existing solutions. Allied research interests include **multimodal interfaces**, **speech-based navigation**, recognition **confidence measures**, and **speech phenomenon during corrections** (such as hyperarticulated speech).

2 Past, Current and Future Work

Despite significant advances in speech recognition, the use of speech for text creation and editing is not widespread. Current solutions force users into a dialog in which they dictate their text, scan for errors, specify the error locations, and then provide corrections. The best strategy for correction varies from situation to situation and users often face cascading sequences of errors.

While a person can dictate at around 125-150 words per minute (WPM) (Feng and Sears, 2004), correcting recognition errors is often slow. User studies have shown corrected entry rates of 8-15 WPM (Karat et al., 1999; Sears et al., 2001). Often users are found to fall back to keyboard correction, an impossibility for people with certain disabilities. User frustration is also a problem, in one study 7 out of 8 users had quit using their speech dictation software after 6 months (Koester, 2003).

My research is about trying to improve this state of affairs. How might we better let users visualize and navigate the space of recognition hypotheses? If we invent a new interface, how do we accurately evaluate its performance? How do users really want to create and edit their text using their voice?

2.1 Correction via pointing

When a user dictates text to a speech recognizer, often the right answer lies somewhere in the recognizer's lattice of results. In order to navigate this result space, I have developed a new interface based on Dasher (Ward et al., 2000). Dasher was originally designed to provide efficient text entry without a keyboard. Users control Dasher with a pointing device such as a mouse or eye tracker. Users enter text by zooming in on letters appearing on the display. Letters and sequences of letters appear in size proportional to their probability in an underlying language model.

In Speech Dasher (Vertanen, 2005), users first dictate their desired text. The recognition results are used in combination with other lattice expansion techniques to create a dynamic language model. Using this model, in one fluid step the user can proofread their text, select the best hypothesis, and provide any words not explicitly predicted.

Experiments on the Hub1 test set show that Speech Dasher significantly reduces the information required by a user to create text. A user trial is currently underway evaluating human performance using Dasher, Speech Dasher, and a conventional dictation interface.

2.2 Evaluating text creation

In most user studies of predictive text or speech dictation, users are asked to transcribe text from a source such as a newspaper or book (Suhm et al., 1999; Ingmarsson et al., 2004). I believe this is an unrealistic task that requires users to write in a style and vocabulary different from their own. It can also necessitate shifts of the user's attention between the interface and the transcription text (MacKenzie and Soukoreff, 2002).

To address these concerns, I am currently developing a new short text composition task set. Users compose sentence-length responses based on short fictitious situations. The user is told specifically what their response should include, allowing particular vocabulary of interest to be elicited (such as difficult to recognize words, proper names, etc). While targeted words in their response might be unfamiliar, the bulk of their entry will be in the user's own style and vocabulary. The user is asked to think of their response before starting entry, hopefully concentrating the majority of the user's "think time" at the start of the task.

2.3 Future work

In the future, I would like to investigate other possible interfaces for the creation and editing of text. Rather than

impose my own preconceived notions, I would like to first explore how users would like things to work. This could be done in a Wizard of Oz style experiment in which users are allowed to interact with a simulated system controlled by a human. How do users proofread their writing? What are the different ways in which they select and correct errors? What speech phenomenon are observed during correction episodes?

Using insights generated from such a Wizard of Oz experiment, I plan to work toward a system which more closely mirrors human expectations.

3 Challenges in Spoken Dialog Systems Research

A large amount of effort has gone into making speech recognizers get it right the first time. But little work has been done in getting it right the second time. How can we improve user's success and satisfaction during error recovery? Should we be using specially trained or adapted acoustic models during error resolution? Can we build systems that know not only when a user is out-of-grammar but also know when they are frustrated or when they are trying to correct a system gone astray?

For a frequently used dialog system, the user typically appreciates that the system has limited capabilities. They tend to learn one way to do their normal activities and little else. The actual system capabilities might far exceed those within a user's normal comfort zone. How do we promote exploration of a dialog system without annoyance? Are there opportunities to use other non-audio mediums to communicate the power and flexibility of a system?

The deployment of statistically-based dialog systems in industry would seem to face several problems. One issue would be communicating the systems capabilities and behaviors with stakeholders such as product marketing and quality assurance. While in my experience it is possible to communicate the intricacies of a hand-tooled dialog system using aids such as diagrams, would the diagrams representing a learnt system be as useful? How easily could a learnt system be adjusted in the face of ad hoc requests for a given behavior or new feature?

References

Jinjuan Feng and Andrew Sears. 2004. Are we speaking slower than we type?: exploring the gap between natural speech, typing and speech-based dictation. *SIGACCESS Access. Comput.*, (79):6–9.

Magnus Ingmarsson, David Dinka, and Shumin Zhai. 2004. Tnt: a numeric keypad based text input method. In *CHI '04: Proceedings of the 2004 conference on Human factors in computing systems*, pages 639–646. ACM Press.

Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 568–575. ACM Press.

Heidi Horstmann Koester. 2003. Abandonment of speech recognition by new users. *RESNA 26th International Annual Conference*.

I. S. MacKenzie and R. W. Soukoreff. 2002. Text entry for mobile computing: Models and methods, theory and practice. *Human-Computer Interaction*, 17:147–198.

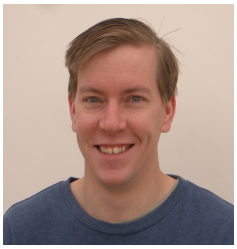
Andrew Sears, Clare-Marie Karat, Kwesi Oseitutu, Azfar S. Karimullah, and Jinjuan Feng. 2001. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society,*, (1):4–15.

Bernhard Suhm, Brad Myers, and Alex Waibel. 1999. Model-based and empirical evaluation of multimodal interactive error correction. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 584–591. ACM Press.

Keith Vertanen. 2005. Efficient computer interfaces using continuous gestures, language models, and speech. *University of Cambridge, Computer Laboratory*, 2005(UCAM-CL-TR-627).

David J. Ward, Alan F. Blackwell, and David J. C. MacKay. 2000. Dasher - a data entry interface using continuous gestures and language models. In *UIST '00: Proceedings of the 13th annual ACM symposium on User interface software and technology*, pages 129–137. ACM Press.

Biographical Sketch



Keith Vertanen is a PhD student of David MacKay at the Inference Group, University of Cambridge. He has completed a BA at the University of Minnesota, a MS at Oregon State University, and a M.Phil at the University of Cambridge. Prior to his return to academia, Keith spent four years designing and building spoken dialog systems in industry. In his free time, he enjoys traveling, photography, and all manners of outdoor pursuits.

Jason D. Williams

jdw30@cam.ac.uk
http://svr-www.eng.cam.ac.uk/~jdw30/

Speech Research Group
Machine Intelligence Laboratory
Department of Engineering
University of Cambridge

1 Research Interests

I am interested in several related aspects of human-computer dialogue.

First, I'm interested in **understanding people's behaviour in human/computer dialogue**. I view the speech recognition interface (including end-pointing, speech recognition, parsing, etc.) as a "**noisy channel**" which not only corrupts a user's meaning, but also alters a user's behaviour. By understanding both of these components, I seek to gain insights into basic properties of human behaviour and also create both models of the channel and user.

Second, I'm interested in exploring **dialogue models which explicitly represent uncertainty** which is introduced by the channel. For example, I'm interested in dialogue models which maintain a probability distribution over *many* possible states of a dialogue. I seek to create and test these models using insights and data gained from studies of the channel and user behaviour.

Finally, I am interested in approaching **dialogue management as a machine learning task** cast as "planning under uncertainty." I am interested in approaches in which a dialogue designer can specify desired outcomes of a dialogue and rely on an optimisation algorithm, provided with a dialogue model, to work out the details of the machine's plan.

2 Past, Current and Future Work

Prior to starting my Ph D, I worked in industry for several years building commercial speech recognition systems for the telephone. While in industry, I focused on usability testing and dialogue design. I explored a variety of human/computer interaction issues such as how users react to various types of prompt strategies (Williams et al., 2003a), the effects of using earcons and "naming" the system (Williams et al., 2003b), how the "domain" influences caller preference for different dialogue structures (Witt and Williams, 2003), and how these factors all relate to the "Call routing" task (Williams and Witt, 2004).

In industry, I found that the speech recognition component was typically much easier (and less expensive!) to build than the "design" of the dialogue. This realisation prompted me to explore machine learning approaches to dialogue management.

Thus far in my Ph D, I have explored two areas. First, I have developed a method for collecting dialogue data using a "Simulated ASR Channel" (Stuttle et al., 2004). The channel models the properties of a typical speech recognition interface, and two experimental subjects attempt to accomplish tasks using the channel. Because the method does not require a dialogue system to be built, I believe the "Simulated ASR Channel" is a faster and cheaper way to collect dialogue data. Using the Simulated ASR Channel, behaviour was observed to be similar to that in human/computer dialogue (Williams and Young, 2004). Thus I am hopeful that this dialogue data will be suitable for training a user model.

I have proposed a method for representing a dialogue model & manager as a Partially Observable Markov Decision Process (POMDP) trained using reinforcement learning (Williams, 2002). Reinforcement learning can require a great deal of training data, and I have spent some time looking at how a Reinforcement-Learning based dialogue manager might be "bootstrapped" with small amounts of data (Williams and Young, 2003).

I am now exploring practical issues of representing a dialogue model & manager as a POMDP. In the remainder of my Ph D, I intend to explore how the training data collected using the Simulated ASR Channel can be incorporated into a POMDP-based dialogue model & manager.

3 Challenges in Spoken Dialogue Systems Research

I see several broad challenges for Spoken Dialogue Systems:

- Dialogue systems are expensive to create, and little is reusable from one domain to another. How can we build systems more quickly? How can we leverage work in one domain in another domain?

- Funding for a research area is often driven by a target application. What are the next "killer" applications for spoken dialogue systems?

- Multi-modal interfaces provide new opportunities for engaging human/computer interaction. Whereas commercial handbooks are readily available for the design of spoken dialogue systems, multi-modal dia-

logue system design is less well-understood. What are the "design constructs" and best practices for multi-modal design?

References

Matthew Stuttle, Jason D. Williams, and Steve Young. 2004. A Framework for Wizard-of-Oz Experiments with a Simulated ASR-Channel. ICSLP, Jeju, South Korea.

Jason D. Williams. 2003. *A Probabilistic Model of Human/Computer Dialogue with Application to a Partially Observable Markov Decision Process.* Ph D first year report. Department of Engineering, University of Cambridge.

Jason D. Williams, Andrew T. Shaw, Lawrence Piano, and Michael Abt. 2003a. Preference, Perception, and Task Completion of Open, Menu-based, and Directed Prompts for Call Routing: a Case Study. Eurospeech, Geneva, Switzerland..

Jason D. Williams, Andrew Shaw, Lawrence Piano, and Michael Abt. 2003b. Evaluating real callers' reactions to Open and Directed Strategy prompts. Applied Voice Input/Output Society (AVIOS). San Jose, California.

Jason D. Williams and Silke M. Witt. 2004. A Comparison of Dialog Strategies for Call Routing. *International Journal of Speech Technology* (Vol 7, No 1). pp 9-24.

Jason D. Williams and Steve Young. 2003. Using Wizard-of-Oz simulations to bootstrap Reinforcement-Learning-based dialog management systems. In *Proceedings of the 4th SIGDIAL Workshop on Discourse and Dialogue*, Sapporo, Japan.

Jason D. Williams and Steve Young. 2004. Characterizing Task-Oriented Dialog using a Simulated ASR Channel. ICSLP, Jeju, South Korea.

Silke M. Witt and Jason D. Williams. 2003. Two Studies of Open vs. Directed Dialog Strategies in Spoken Dialog Systems. Eurospeech, Geneva, Switzerland.

Biographical Sketch

Jason Williams grew up in Hampton, New Hampshire. He studied Electrical Engineering as an undergraduate at Princeton University and Computer Speech and Language Processing as a Masters student at Cambridge University in the UK. Since then, he as worked at McKinsey & Company, Tellme, and Edify. At Tellme and Edify, Jason focused on creating spoken dialogue systems for large US and European companies. At Tellme, he also managed the recognition tuning infrastructure and started a European professional services organization based in Brussels, Belgium.

Jason is now working on his Ph D in the Machine Intelligence Lab at Cambridge University. His main interests are: understanding human behaviour in the speech recognition channel, modelling the uncertainty in human/computer dialogue, and applying machine learning to dialogue management.

Outside of work, Jason is an avid runner,[1] maniacal hiker, eager cook, and committed oenophile. He also volunteers for several organizations, including the Princeton Club of Great Britain.

[1] His marathon time of 4 hours is poor, but he did run the whole way.

List of All Submitted Topics

Each of the participants in the Young Researchers' Roundtable on Spoken Dialog System has proposed three topics for discussion at the workshop. The complete list of submitted topics is available below (in no particular order).

- For some years now, progress has been made in learning dialogue strategies from corpora or using user feedback. While the resulting systems most often show better performance than hand-crafted systems, knowledge-based approaches stubbornly refuse to go away. It would be interesting to see why that is, and how a combination of both statistical and knowledge-based approaches can be beneficial.
- What can be re-used from one dialog system to the next?
- My own research indicates that something like half of speech recognition errors are of no semantic significance, that is, the misrecognised hypothesis is semantically equivalent to the real utterance. I have seen little research on this. Being able to ignore half of misrecognitions would be very useful. Does anyone else have work or observations on this?
- Current speech systems do not take into account human emotional expression. However, emotion recognition algorithms are improving, and first commercial synthesis systems offer affective expression combined with good sound quality.
- Creating a dialogue system (for research purposes) is a challenging task that requires a number of people just to get off the ground. Can we define a set of common, open-source components/resources we could share which could make this an easier task? Examples would include a dialogue manager/shell (like TrindiKit), a common (freeware) speech recognizer, common base grammars, etc.
- Dialogue systems are usually composed by stringing together several components: speech recognizer, parser, dialogue manager, databases, ontologies, etc. In current systems, these components are usually ordered such that the output of one process feeds into the input of another process, essentially making each process a "black box" (e.g. an n-best list is produced by the recognizer which is then shipped to the parser and then the parse is shipped to the dialogue manager). However, at an intuitive level is doesn't seem like humans operate this way. Instead they seem to make use of these resources in parallel; e.g. they integrate real-world knowledge as they parse words they hear on the fly. Is this a model to aim for? What technological boundaries stand in the way? What has been accomplished thus far?
- How can we better reconcile statistical and non-statistical techniques in dialog?
- The interaction between dialog systems and speech recognition - (how) can we use dialog information to improve speech recognition? (How) can we use more of the information from the recognizer to perform better on dialog processing?
- A fair amount of current research on spoken dialogue systems is justified by claims that speech is the most natural or most efficient form of interface. But traditional input devices have grown to be so familiar that most users find it quicker, more intuitive and less error-prone to use the mouse or keyboard. What kinds of new functionality could future Spoken Dialogue Systems offer in order to overcome users' reluctance to change? What application domains are best suited for voice-driven interfaces?

- How to determine a strategy for multi-modal output planning? The claim by (Cohen and Oviatt, 1995) that speech is the primary input mode can be confirmed by results from Wizard-of-Oz studies undertaken by the TALK project. The SACTI-2 corpus and the MP3 corpus both show that users rarely take advantage of a click interface. But multi-modal input behavior increases with noise as introduced by speech recognition and the wizard can encourage the user to act multi-modally (Schatzmann, 2004). Multi-modal utterances are more likely on the side of the wizard especially when handling interpretation ambiguities. In sum, multi-modal output generation seems to be especially promising in the presence of uncertainty.
- Evaluation of a spoken dialog system is to analyze design errors and estimate how well the system fits its purpose and meets actual user needs and expectations. It is critical for the performance of the system. Despite of its key importance fewer resources have been invested in the usability evaluation measures of spoken dialogue systems over the years than in its component technologies. Currently there is no standard as such, to which evaluation criteria to use. How are spoken dialogue systems evaluated so far? Are they application dependent? What are the factors a developer has to consider in choosing a particular metric? How effectively the evaluation results can be used in the modification of the system?
- Due to the typically linear decoding chain of pragmatic-semantic information present in the user input, often additional unification methods need to be created for combining or embedding those levels, or to guarantee the legitimacy of their combinations by constraints or by dependency relations between concepts. In case some pragmatic and/or semantic input components are empirically found to be more optimally identified together than in isolation, part of these problems can be solved. Machine learning provides a framework for evaluating differently designed setups of pragma-semantic processing. What other automatic techniques can be applied for identifying optimally co-identifiable phenomena from dialogue corpora? What pragmatic-semantic aspects can be shown to be better identified in parallel than in isolation, and based on which dialogue properties? What practical knowledge is gained from such experiments concerning well-formedness of tasks, hierarchy and probability of different types of pragmatic-semantic pieces of information?
- Should human-human communication be the gold standard for dialog system development? If so, what aspects are truly fundamental (thus necessary to implement), and which might be superfluous? If not, what else should be the gold standard, and is there a principled way to decide how to make design decisions?
- Spoken Dialog Systems and the Real World: Where Can They Have an Impact? Today simple task-oriented spoken dialog systems are transitioning into day-to-day use and becoming the norm for the phone-based customer service industry. What is the next major application where SDS can have a real impact? (Personal assistants? interaction with robots/entertainment? smart homes? etc). What are the challenges raised by each of these applications?
- Apply psychometrics theory to study the factors that affect spoken dialog systems usability. A shift from natural interaction to usable interaction.
- A large amount of effort has gone into making speech recognizers get it right the first time. But little work has been done in getting it right the second time. How can we improve user's success and satisfaction during error recovery? Should we be using specially trained or adapted acoustic models during error resolution? Can we build

systems that know not only when a user is out-of-grammar but also know when they are frustrated or when they are trying to correct a system gone astray?

- Dialogue research, like AI in general, is both an engineering discipline and a science. Because of this it offers its practitioners the ability to work on three levels over the course of a career: solving problems at hand; contributing to the dialogue community by addressing questions that it needs to have resolved; and contributing to the development of the sciences of language and cognition. If one wishes to work in all three, how does one choose the problems to attack over the course of a career?
- Are there meaningful partitions of the space of different varieties of dialogue systems that have been built or one can imagine building? Can we describe a space which many of us are interested in and find a way to create and share generic components which are useful in that space? For example, along what dimensions can we characterize the differences between a dialogue system for retrieving movie show times and one that controls the appliances in a smart house? What are our motivations for making such distinctions, and how can we translate these motivations into a useful partitioning of the space?
- Researchers often work toward building dialogue systems which engage in "natural" interactions. To what extent should human/computer interaction model human-human interaction? What does "natural" mean in the context of human-computer interactions? Is this a worthwhile goal?
- Currently reinforcement learning approaches can perform well in terms of increasing the return of a reward function for a given task. However, the exact nature of these reward functions remains an open question. What are useful values for rewards in dialogues, and can they be standardized for all users?
- Speech synthesis is seldom discussed in spoken dialog systems work. What are the weaknesses of state-of-the-art TTS when it comes to its use in spoken dialog systems? What about the trade-off between naturalness and intelligibility in our field? Is there a need to synthesize more "conversational" speech? For example, is the prosody of our synthetic voices appropriate for conversational agents?
- How, when, and how much should training on system use be a prerequisite to (or an aspect of) interacting with a dialog system?
- Recently there is a growing interest in developing real-world technology for emerging markets to which computing resources remain largely out of reach. Spoken language systems can serve as effective user interfaces in these regions. There has been relatively little work in developing spoken dialogue systems, especially in an environment with low literacy and with a wide range of languages and dialects. There are a lot of design challenges that call for novel research methods. Is this a viable research area? What are the difficulties? How are we going to address them? Can existing techniques be used or they have to be modified to accomplish the task?
- Spoken dialog systems currently tend to be concentrated in information access applications. What are the bleeding-edge applications for SDS? Entertainment? Communication with robots? Interactive guidance systems? What else can we think of, and how feasible are these ideas?
- In what way can humans learn to speak more like machines, to enable us to communicate with them better on "their terms." CMU's Universal Speech Interface is a good approach, but it might also be useful to, for example, write and debug a computer program entirely through a headset. What would a dialogue or a programming environment that allows this kind of activity look like?

- Are people actually capable of learning how to use the systems that are being developed? How much training is necessary? How robust is the training? Can people still perform well when they're under stress or trying to do several things at once? How do individual differences factor in?
- How to define usability? The PARADISE method (Walker et al., 1997) is a widely accepted framework to measure usability. However, the definition of user satisfaction as function of task success and dialogue costs seems to be problematic. Studies have shown that user satisfaction does not correlate with task completion times (Williams and Young, 2004) and the perceived task success depends on different error handling strategies (Skantze, 2003).
- Although in research environments this fact can be easily overlooked, spoken dialog systems do not exist in a vacuum. What kinds of evaluation strategies can be designed in order to more accurately capture and assess users' real-world situations and needs? For instance, such evaluation strategies could take into account the urgency of the information desired and the location of access (e.g. home, car, public kiosk), and should be able to assess the factors for success and the costs of failure for spoken dialogue systems in various situations.
- In current spoken dialogue systems, the strategy used by the dialogue manager is typically explicitly coded; however, there is no guarantee that this strategy is optimal from the point-of-view of the user. A more effective approach is to train or adapt the dialogue management system based on training data or via explicit user feedback. Is this a viable approach? What technical boundaries stand in the way?
- What practical techniques do we have for evaluating dialogue systems, and dialogue managers in particular? We have several traditions to call from. We can run experiments as done in, for example, psychology of education, with multiple concurrent human participants testing full systems, with control groups and measures such as time to task completion, quality of task completion, or learning gain. But this requires a large number of subjects, which is complicated when the domain requires specially-trained subjects, such as those with basic medical or military skills; furthermore, it is not very fine-grained, requiring the testing of an entire system (unless variations of the systems are tested, which increases the number of subjects required.) We can use corpus approaches, as in statistical natural language processing, to calculate accuracy and efficiency measures, but there are questions to this, such as what it means for a dialogue system to be accurate, and how we can model dialogues to test the systems against. For researchers to be able to compare dialogue systems to one another, some kind of agreement on typical approaches is needed.
- The dialog system may initiate a dialog to push useful information without any user request. What are the issues and challenges?
- Dialogue systems are expensive to create, and little is reusable from one domain to another. How can we build systems more quickly? How can we leverage work in one domain in another domain? What is an appropriate way to divide the "pieces" of a dialogue system to maximize reuse of components from one domain to another?
- While being able to talk to machines (and being understood) is often something perceived as being desirable, it might actually be not appropriate in all situations. It would be interesting to see whether some sort of user interface guideline can be developed to decide which UI is best for a given application.

- Why is tuning a deployed dialog system so difficult, and how can we improve the process?
- Due to form factor limitations of mobile devices, visual output and gestural input are resource-demanding and cumbersome. Speech interaction is widely regarded to be a promising alternative, because it is not dependent on device size (apart from technical constraints, of course). One challenge in this regard is to identify useful context characteristics for speech (concerning privacy, background noise, attention constraints etc) and to find promising services.
- For a frequently used dialog system, the user typically appreciates that the system has limited capabilities. They tend to learn one way to do their normal activities and little else. The actual system capabilities might far exceed those within a user's normal comfort zone. How do we promote exploration of a dialog system without annoyance? Are there opportunities to use other non-audio mediums to communicate the power and flexibility of a system?
- Most of the research in NL Dialog Modeling has concentrated on dialog systems which have one participant as an agent which actively takes part in the dialog. But there is a need for systems which can monitor the dialog between two parties without much of an active participation. For e.g. tracking the dialog between two parties communicating with help of a speech to speech translation system OR systems that review and analyze the dialogs between the participants like After action review. Do we need different set of abilities in these systems and what would they be? Would these be very similar to systems that are multi-party ready?
- For those of us who work with dialog management, speech synthesis may often be overlooked as a core aspect of an SDS. Yet speech synthesis is presumably a large component of a system's personality. How can we assess and manage the affect of speech synthesis on SDS?
- What additional technology (to what already exists) would it take to create a dialogue system front-end to my operating system (e.g., Linux)? For example, we would need a way for each installed program to provide specifications about its abilities and vocabulary.
- Error Handling in Spoken Dialog Systems. One of the major challenges in today's spoken dialog systems is their lack of robustness when faces with understanding errors. What can we do to alleviate this problem? Some of the more specific questions I see are: how does a system know that it does not know? How can systems build more accurate beliefs for the information they hold? Is there a set of agreed-upon strategies for recovery and what are their merits relative to each other? How can a system learn optimal error handling behavior from experience?
- Maybe the greatest problem with spoken dialogue systems today is their poor ability to handle errors. How can we improve error handling in spoken dialogue systems? What can we learn from human error handling and how can we collect such data?
- Should we try to make dialogue systems handle natural conversational behavior (and handle the errors that arise), or should the user learn a "special" language (with potentially less errors). What are the major problems with the first approach? For the second approach, is it realistic to think that users actually will learn a new set of commands for each interface, or is it possible to agree on a standard? Is speech as a modality really useful (compared to graphical interaction), if we only can handle command-based spoken interaction?

- How can we better use the full information from a speech recognition system? What strategies of confirmation or repair can be used given a full n-best parser output? Is it worth conditioning the parser output on the dialog state?
- Does anthropomorphism help? Anthropomorphism was considered to be problematic for spoken dialogue systems as the system pretend to have human capabilities, (Shneiderman, 1998). In my experience, making a system more human-like increases its acceptance. People like to interact with dialogue systems that are flattering and people tend to be polite to systems by themselves. Furthermore, human-like behavior is a familiar interaction paradigm to users.
- Funding for research areas such as spoken dialogue systems and multi-modal systems is often driven by a target application. Telephone-based spoken dialogue systems are well-established and are becoming more pervasive. What are the next "killer" domains and applications for spoken dialogue systems? Although multi-modal dialogue systems are intuitively appealing, they have enjoyed little commercial success. Why is this so? What are the obstacles faced by multi-modal dialogue systems?
- Help users to build a conceptual model of spoken dialog systems. Train users to understand how dialog systems work and what are their main limitations. Establish conventions to make different systems interact in a similar way, allowing users to reuse their experience with any system.
- For interaction with robots, it is often assumed that natural language dialogue is an appropriate means of communication. However, due to the open-endedness of the application (ask your robot 'bring me X' where X can be anything), it is more difficult for the user to understand the limitations of the user interface than it is in task-oriented dialogue systems. What kind of techniques and user interface patterns are needed to alleviate the problem?
- What are the key issues in developing a multimodal dialog system? How can the increased interactivity be utilised without confusing a inexperienced users? What new types of dialog strategy can be used with such interfaces? When is multimodal interaction either not used or not appropriate?
- How does the system adapt its behavior according to the user skill level?
- Most commercial systems being deployed appear to be grammar based, written fairly quickly and cheaply, and without extensive research by concatenating the recognition and understanding process. Is this the way of the future? Are wide vocabulary n-gram based systems destined to stay in the research lab?
- Are different system designs optimal for different applications? For instance, is one design well suited to a multitasking or problem-solving environment while another is well-suited for repeat users rapidly accessing information from a well-defined database? How can we go about researching such potential differences?
- In spoken dialogue, errors due to acoustic, linguistic and knowledge mismatch are inevitable. To realize robust spoken language systems, such errors should be detected, and handled appropriately. For example, providing informative feedback to the user. What types of errors most affect communication? What strategies are required to overcome these errors?
- Multilingual systems have become popular which allow the users to interact with the system in their own native language. They have a lot of real world applications particularly in environments like Europe and India where there is a familiarity of more

than one language. Designing a multilingual spoken language dialog system requires each component to be as language independent as possible. There are also several other issues like collection of training data, evaluation of the system and flexibility in design to include a new language which need to be addressed. Can a single frame work accommodate all the languages? What are design issues? Are they tractable? The data collection and evaluation is very much time consuming. How it can be reduced?

- The deployment of statistically-based dialog systems in industry would seem to face several problems. One issue would be communicating the systems capabilities and behaviors with stakeholders such as product marketing and quality assurance. While in my experience it is possible to communicate the intricacies of a hand-tooled dialog system using aids such as diagrams, would the diagrams representing a learnt system be as useful? How easily could a learnt system be adjusted in the face of ad hoc requests for a given behavior or new feature?
- A 5% word error rate still means one word in 2O is wrong. At what level will recognition errors stop being a problem?
- Error recovery strategies
- Will computer games be the breakthrough for speech technology? What is done today and what are the possibilities? Which types of games could be enhanced by adding speech?
- Task design
- Computers and humans clearly have different strengths. In what ways can we exploit the abilities that artificial systems have and humans don't? In what ways will our options increase in this respect as technology progresses?
- What are the resources needed to support standardized comparative evaluations of spoken dialog system technologies? In other fields (e.g. speech recognition) there are standardized resources and methodologies for evaluation. Not so in spoken dialog systems. Is it possible to accomplish this in spoken dialog systems? What is required to perform comparative evaluations of spoken dialog systems? How can we encourage this more?
- In many other disciplines there are fairly standard test/training sets (e.g. wall street journal partitions for parsing) as well as regular competitions. Dialogue systems research is lacking in this area. What useful corpora are currently available? What are being developed? What should be developed? Would some sort of annual or bi-annual competition be useful/interesting/appropriate?
- Incremental improvement of dialogue systems by semi-supervised analysis of previous runs
- To realize robust spoken understanding systems, one approach is to closely integrate knowledge from multiple components in the dialogue system into a single decoder. Rather then searching for the word sequence with maximum likelihood, the concept/concepts with maximum likelihood are the search target. Is such an approach technically viable? What knowledge sources should be incorporated into such a framework?
- NLP research community has witnessed a change. From early analysis driven, rule-based systems to current - more empirical methods based statistical NLP. Is such a change on the horizon for NL dialog research? If no, what is it that prevents it and if yes, what can

help it? What kind of methods or corpora do we already have or need to build to make this happen.

- What do we consider to be the main obstacles to human-level performance in spoken language systems? What forms are the solutions likely to take?
- A lively line of investigation in spoken dialogue systems is domain-independent architecture. What features do we suppose the human "dialogue system architecture" possesses?
- How can we support phenomena such as turn-taking and (system) backchanneling in dialogue systems? Will we have to have proper incremental language processing to do this, or are there shortcuts?
- Most current dialog systems use a rigid turn-taking mechanism conditioned by an utterance end-pointing based solely on low-level signal processing (energy, silence) and by a strict pipeline architecture where each utterance has to be recognized, understood, go through a dialog manager, and the corresponding answer must be generated and synthesized before the system can take into account the user's next utterance. What are the issues raised by such an approach? How much of an impact does it have on the quality of the interaction (including metrics such as task success)? What are alternative turn-taking mechanisms have been/can be proposed?
- How can the user be made aware of the capabilities and limitations of a spoken dialog system? I guess this question arises mainly in task-oriented dialog systems. But the more fundamental question would be - is there a need for making user aware of the limitations? Are there ways to hide the limitations and fending off understanding difficulties using similar approaches to chatter-bots. What would such a hybrid solution look like and would it be effective?
- In automatic processing of pragmatic and semantic information of the user's input to a spoken dialogue system dialogue acts or semantic entities might be differently formulated depending on whether the application concerns air travel, spare time activity, health, etc. An important issue is how pragmatic and semantic information need to be encoded for optimal processing of the user input. Knowledge-based definition of fine-grained discourse categories prove unnecessary when robust approaches drawing on automatically obtainable --but perhaps domain or application-specific-- dialogue properties can produce a similar result. Is using fine-grained categories worth the effort of manually definition, or should we better aim at using rougher super-categories? Empirically obtained results on dialogue data might be able to trace where the balance is: e.g. if both a manually defined dialogue act taxonomy and rough pragma-semantic categories established for a given application are compared in automatic processing of (spoken) user input. What kind of automatic processing techniques are best applicable to such tasks?
- Whereas usability is a major concern when creating commercial software, it is seldom discussed in the context of research spoken dialog systems. How can we evaluate and improve the usability of our systems? How can we advance the complexity and "intelligence" of these systems without harming usability by making them more fragile? Are there design/engineering/research methods that could help evaluate/maintain the robustness of systems that explore new research directions?
- How can we make dialog authoring environments simple to use so that people without any familiarity with speech, dialog, or even programming can build and deploy dialog systems?

- Currently used models of dialogue still lack elegant ways of explicitly accounting for uncertainty. Yet, human-machine dialogue always involves a noisy communication channel and the dialogue manager can never be certain that the output of the recognition and understanding unit is error-free. Better models of dialogue do exist, but our experience in using them is still limited. A) What are the main upcoming milestones in this area of research? B) Once we have dialogue models that incorporate the notion of uncertainty, can we design new (and better) dialogue strategies? What could these look like?
- Error detection and handling of miscommunication is a central issue in designing dialogue systems. Various types of applications conducting information-seeking dialogues might differ in the way they can handle situations with inaccurate identification of concepts. It seems that systems operating in a less restricted domain can opt to provide a more general answer to the user which still seems natural without frustrating him, so that in the following turn the user can specify the input again. In a restricted domain, users that are confronted with a reply that does not satisfy their request immediately (as it is semantically more general) will sense miscommunication. What are optimal ways of designing more general, error-handling system prompts in these two types of dialogue systems? Can a domain ontology or semantic word-nets be of help in this issue? How do users express their dissatisfaction with or acceptance of prompts that are semantically more general than the semantics of the user turn? Are these phenomena different in human-machine and than in human-human communication?
- A problem experienced by many users of current SDS is that one can never know in advance what the system will understand. Users are forced to try out different instructions until they finally find the one that the system can successfully respond to. Dialogue Systems are far away from correctly understanding any type of input, and yet their capabilities often exceed simple command language. How do we bridge the gap between the users' expectations and the systems capabilities? Is it possible to standardise voice interfaces but still cover a broad range of domains? Are users willing to adapt to simpler (but possibly constraining or less natural) interfaces such as Speech-Graffiti? Do we have alternative ways of approaching this problem?
- Designers of user interfaces for the auditory modality are often forced to make a fundamental decision: when to use speech, when to use non-speech sound, and how to combine these? Related design questions are about constraints and opportunities for sound and speech for different types of systems, e.g. multimedia systems, mobile applications, speech telephony services or special applications for handicapped people.
- Create a branch of software engineering to deal with spoken dialog systems. Proposals of methodologies, guidelines and tools for developing and deploying SDS in the real world. A model to describe SDS at analysis and design phases of development. CASE tools tailored to SDS and reference of good practices and design patterns.
- To build SDS needs lots of work including corpus analysis, dialog management, backend access, grammar, language model, etc. What are the easy ways to build SDS for a new domain?