An Experiment Setup for Collecting Data for Adaptive Output Planning in a Multimodal Dialogue System

Ivana Kruijff-Korbayová, Nate Blaylock, Ciprian Gerstenberger, Verena Rieser
Saarland University, Saarbrücken, Germany
korbay@coli.uni-sb.de

Tilman Becker, Michael Kaißer, Peter Poller, Jan Schehl
DFKI, Saarbrücken, Germany
tilman.becker@dfki.de

Abstract

We describe a Wizard-of-Oz experiment setup for the collection of multimodal interaction data for a Music Player application. This setup was developed and used to collect experimental data as part of a project aimed at building a flexible multimodal dialogue system which provides an interface to an MP3 player, combining speech and screen input and output. Besides the usual goal of WOZ data collection to get realistic examples of the behavior and expectations of the users, an equally important goal for us was to observe natural behavior of multiple wizards in order to guide our system development. The wizards' responses were therefore not constrained by a script. One of the challenges we had to address was to allow the wizards to produce varied screen output a in real time. Our setup includes a preliminary screen output planning module, which prepares several versions of possible screen output. The wizards were free to speak, and/or to select a screen output.

1 Introduction

In the larger context of the TALK project[1] we are developing a multimodal dialogue system for a Music Player application for in-car and in-home use, which should support natural, flexible interaction and collaborative behavior. The system functionalities include playback control, manipulation of playlists, and searching a large MP3 database. We believe that in order to achieve this goal, the system needs to provide advanced adaptive multimodal output.

We are conducting Wizard-of-Oz experiments [Bernsen *et al.*, 1998] in order to guide the development of our system. On the one hand, the experiments should give us data on how the potential users interact with such an application. But we also need data on the multimodal interaction strategies that the system should employ to achieve the desired naturalness, flexibility and collaboration. We therefore need a setup where the wizard has freedom of choice w.r.t. their response and its realization through single or multiple modalities. This makes it different from previous multimodal experiments, e.g., in the SmartKom project [Türk, 2001], where the wizard(s) followed a strict script. But what we need is also different in several aspects from taking recordings of straight human-human interactions: the wizard does not hear the user's input directly, but only gets a transcription, parts of which are sometimes randomly deleted (in order to approximate imperfect speech recognition); the user does not hear the wizard's spoken output directly either, as the latter is transcribed and re-synthesized (to produce system-like sounding output). The interactions should thus more realistically approximate an interaction with a system, and thereby contain similar phenomena (cf. [Duran *et al.*, 2001]).

The wizard should be able to present different screen outputs in different context, depending on the search results and other aspects. However, the wizard cannot design screens on the fly, because that would take too long. Therefore, we developed a setup which includes modules that support the wizard by providing automatically calculated screen output options the wizard can select from if s/he want to present some screen output.

Outline In this paper we describe our experiment setup and the first experiences with it. In Section 2 we overview the research goals that our setup was designed to address. The actual setup is presented in detail in Section 3. In Section 4 we describe the collected data, and we summarize the lessons we learnt on the basis of interviewing the experiment participants. We briefly discuss possible improvements of the setup and our future plans with the data in Section 5.

2 Goals of the Multimodal Experiment

Our aim was to gather interactions where the wizard can combine spoken and visual feedback, namely, displaying (complete or partial) results of a database search, and the user can speak or select on the screen.

Multimodal Presentation Strategies The main aim was to identify strategies for the screen output, and for the multimodal output presentation. In particular, we want to learn

[1]TALK (Talk and Look: Tools for Ambient Linguistic Knowledge; `www.talk-project.org`) is funded by the EU as project No. IST-507802 within the 6th Framework program.

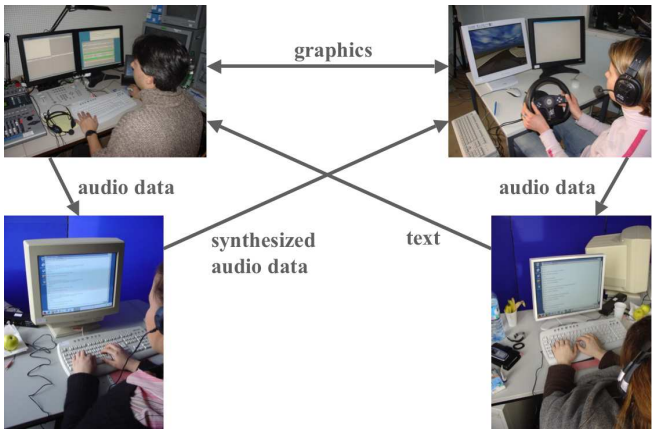

Figure 1: Multimodal Wizard-of-Oz data collection setup for an in-car music player application, using the Lane Change driving simulator. Top right: User, Top left: Wizard, Bottom: transcribers.

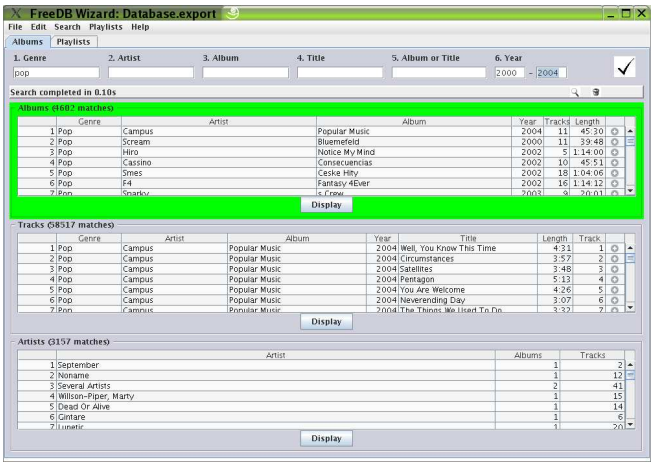

Figure 2: Screenshot from the FreeDB-based database application, as seen by the wizard. First-level of choice what to display.

when and what content is presented (i) verbally, (ii) graphically or (iii) by some combination of both modes. We expect that when both modalities are used, they do not convey the same content or use the same level of granularity. These are important questions for multimodal fission and for turn planning in each modality.

We also plan to investigate how the presentation strategies influence the responses of the user, in particular w.r.t. what further criteria the user specifies, and how she conveys them.

Multimodal Clarification Strategies The experiments should also serve to identify potential strategies for multimodal clarification behavior and investigate individual strategy performance. The wizards' behavior will give us an initial model how to react when faced with several sources of interpretation uncertainty. In particular we are interested in what medium the wizard chooses for the clarification request, what kind of grounding level he addresses, and what "severity" he indicates. [2] In order to invoke clarification behavior we introduced uncertainties on several levels, for example, multiple matches in the database, lexical ambiguities (e.g., titles that can be interpreted denoting a song or an album), and errors on the acoustic level. To simulate non-understanding on the acoustic level we corrupted some of the user utterances by randomly deleting parts of them.

3 Experiment Setup

We describe here some of the details of the experiment. The experimental setup is shown schematically in Figure 1. There are five people involved in each session of the experiment: an experiment leader, two transcribers, a user and a wizard.

The wizards play the role of an MP3 player application and are given access to a database of information (but not actual music) of more than 150,000 music albums (almost 1 million songs), extracted from the FreeDB database.[3] Figure 2 shows an example screen shot of the music database as it is presented to the wizard. Subjects are given a set of predefined tasks and are told to accomplish them by using an MP3 player with a multimodal interface. Tasks include playing songs/albums and building playlists, where the subject is given varying amounts of information to help them find/decide on which song to play or add to the playlist. In a part of the session the users also get a primary driving task, using a *Lane Change* driving simulator [Mattes, 2003]. This enabled us to test the viability of combining primary and secondary task in our experiment setup. We also aimed to gain initial insight regarding the difference in interaction flow under such conditions, particularly with regard to multimodality.

The wizards can speak freely and display the search result or the playlist on the screen. The users can also speak as well as make selections on the screen.

The user's utterances are immediately transcribed by a typist and also recorded. The transcription is then presented to the wizard.[4] We did this for two reasons: (1) To deprive the wizards of information encoded in the intonation of utterances, because our system will not have access to it either. (2) To be able to corrupt the user input in a controlled way, simulating understanding problems at the acoustic level. Unlike [Stuttle *et al.*, 2004], who simulate automatic speech recognition errors using phone-confusion models, we used a tool that "deletes" parts of the transcribed utterances, replacing them by three dots. Word deletion was triggered by the experiment leader. The word deletion rate varied: 20% of the utterances got weakly and 20% strongly corrupted. In 60% of the cases the wizard saw the transcribed speech uncorrupted.

The wizard's utterances are also transcribed (and recorded)

[2] Severity describes the number of hypotheses indicated by the wizard: having no interpretation, an uncertain interpretation, or several ambiguous interpretations.

[3] Freely available at http://www.freedb.org

[4] We were not able to use a real speech recognition system, because we do not have one trained for this domain. This is one of the purposes the collected data will be used for.

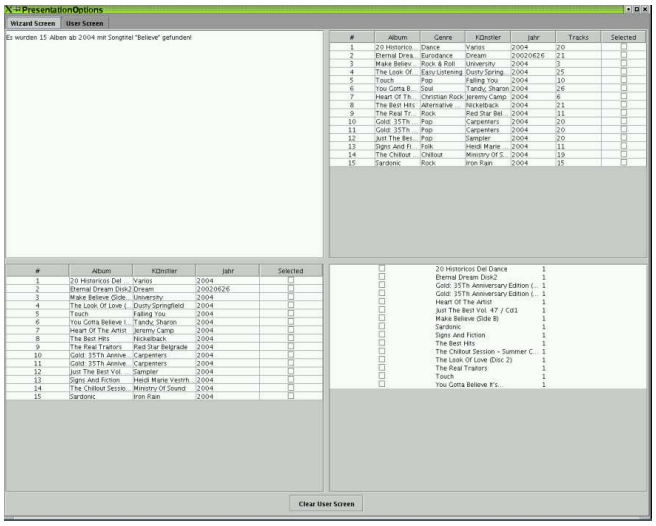

Figure 3: Screenshot from the display presentation tool offering options for screen output to the wizard for second-level of choice what to display an how.

and presented to the user via a speech synthesizer. There are two reasons for doing this: One is to maintain the illusion for the subjects that they are actually interacting with a system, since it is known that there are differences between human-human and human-computer dialogue [Duran *et al.*, 2001], and we want to elicit behavior in the latter condition; the other has to do with the fact that synthesized speech is imperfect and sometimes difficult to understand, and we wanted to reproduce this condition.

The transcription is also supported by a typing and spelling correction module to minimize speech synthesis errors and thus help maintain the illusion of a working system.

Since it would be impossible for the wizard to construct layouts for screen output on the fly, he gets support for his task from the WOZ system: When the wizard performs a database query, a graphical interface presents him a first level of output alternatives, as shown in Figure 2. The choices are found (i) albums, (ii) songs, or (iii) artists. For a second level of choice, the system automatically computes four possible screens, as shown in Figure 3. The wizard can chose one of the offered options to display to the user, or decide to clear the user's screen. Otherwise, the user's screen remains unchanged. It is therefore up to the wizard to decide whether to use speech only, display only, or to combine speech and display.

The types of screen output are (i) a simple text-message conveying how many results were found, (ii) output of a list of just the names (of albums, songs or artists) with the corresponding number of matches (for songs) or length (for albums), (iii) a table of the complete search results, and (iv) a table of the complete search results, but only displaying a subset of columns. For each screen output type, the system uses heuristics based on the search to decide, e.g., which columns should be displayed. These four screens are presented to the wizard in different quadrants on a monitor (cf. Figure 3), allowing for selection with a simple mouse click. The heuristics for the decision what to display implement preliminary strategies we designed for our system. We are aware that due to the use of these heuristics, the wizard's output realization may not be always ideal. We have collected feedback from both the wizards and the users in order to evaluate whether the output options were satisfactory (cf. Section 4 for more details).

Technical Setup To keep our experimental system modular and flexible we implemented it on the basis of the Open Agent Architecture (OAA) [Martin *et al.*, 1999], which is a framework for integrating a community of software agents in a distributed environment. Each system module is encapsulated by an OAA wrapper to form an OAA agent, which is able to communicate with the OAA community. The experimental system consists of 12 agents, all of them written in Java. We made use of an OAA monitor agent which comes with the current OAA distribution to trace all communication events within the system for logging purposes.

The setup ran distributed over six PCs running different versions of Windows and Linux.[5]

4 Collected Data and Experience

The SAMMIE-2[6] corpus collected in this experiment contains data from 24 different subjects, who each participated in one session with one of our six wizards. Each subject worked on four tasks, first two without driving and then two with driving. The duration was restricted to twice 15 minutes. Tasks were of two types: searching for a title either in the database or in an existing playlist, building a playlist satisfying a number of constraints. Each of the two sets for each subject contained one task of each type. The tasks again differed in how specific information was provided. We aimed to keep the difficulty level constant across users. The interactions were carried out in German.[7]

The data for each session consists of a video and audio recording and a logfile. Besides the transcriptions of the spoken utterances, a number of other features have been annotated automatically in the log files of the experiment, e.g., the wizard's database query and the number of found results, the type and form of the presentation screen chosen by the wizard, etc. The gathered logging information for a single experiment session consists of the communication events in chronological order, each marked by a timestamp. Based on this information, we can recapitulate the number of turns and the specific times that were necessary to accomplish a user task. We expect to use this data to analyze correlations be-

[5]We would like to thank our colleagues from CLT Sprachtechnologie `http://www.clt-st.de/` for helping us to set up the laboratory.

[6]SAMMIE stands for Saarbrücken Multimodal MP3 Player Interaction Experiment. We have so far conducted two series of data-collection experiments: SAMMIE-1 involved only spoken interaction (cf. [Kruijff-Korbayová *et al.*, 2005] for more details), SAMMIE-2 is the multimodal experiment described in this paper.

[7]However, most of the titles and artist names in the music database are in English.

tween queries, numbers of results, and spoken and graphical presentation strategies.

Whenever the wizard made a clarification request, the experiment leader invoked a questionnaire window on the screen, where the wizard had to classify his clarification request according to the primary source of the understanding problem. At the end of each task, users were asked to what extent they believed they accomplished their tasks and how satisfied they were with the results. Similar to methods used by [Skantze, 2003] and [Williams and Young, 2004], we plan to include subjective measures of task completion and correctness of results in our evaluation matrix, as task descriptions can be interpreted differently by different users.

Each subject was interviewed immediately after the session. The wizards were interviewed once the whole experiment was over. The interviews were carried out verbally, following a prepared list of questions. We present below some of the points gathered through these interviews.

Wizard Interviews All 6 wizards rated the overall understanding as good, i.e., that communication completed successfully. However, they reported difficulties due to delays in utterance transmission in both directions, which caused unnecessary repetitions due to unintended turn overlap.

There were differences in how different wizards rated and used the different screen output options: The table containing most of the information about the queried song(s) or album(s) was rated best and shown most often by some wizards, while others thought it contained too much information and would not be clear at first glance for the users and hence they used it less or never. The screen option containing the least information in tabular form, namely only a list of songs/albums with their length, received complementary judgments: some of the wizards found it useless because it contained too little information, and they thus did not use it, and others found it very useful because it would not confuse the user by presenting too much information, and they thus used it frequently. Finally, the screen containing a text message conveying only the number of matches, if any, has been hardly used by the wizards. The differences in the wizards' opinions about what the users would find useful or not clearly indicate the need for evaluation of the usefulness of the different screen output options in particular contexts from the users' view point.

When showing screen output, the most common pattern used by the wizards was to tell the user what was shown (e.g., *I'll show you the songs by Prince*), and to display the screen. Some wizards adapted to the user's requests: if asked to show something (e.g., *Show me the songs by Prince*), they would show it without verbal comments; but if asked a question (e.g., *What songs by Prince are there?* or *What did you find?*), they would show the screen output and answer in speech.

Concerning the adaptation of multimodal presentation strategies w.r.t. whether the user was driving or not, four of the six wizards reported that they consciously used speech instead of screen output if possible when the user was driving. The remaining two wizards did not adapt their strategy.

On the whole, interviewing the wizards brought valuable information on presentation strategies and the use of modalities, but we expect to gain even more insight after the annotation and evaluation of the collected data. Besides observations about the interaction with the users, the wizards also gave us various suggestions concerning the software used in the experiment, e.g., the database interface (e.g., the possibility to decide between strict search and search for partial matches, and fuzzy search looking for items with similar spelling when no hits are found), the screen options presenter (e.g., ordering of columns w.r.t. their order in the database interface, the possibility to highlight some of the listed items), and the speech synthesis system.

Subject Interviews In order to use the wizards' behavior as a model for interaction design, we need to evaluate the wizards' strategies. We used user satisfaction, task experience, and multi-modal feedback behavior as evaluation metrics.

The 24 experimental subjects were all native speakers of German with good English skills. They were all students (equally spread across subject areas), half of them male and half female, and most of them were between 20 to 30 years old.

In order to calculate user satisfaction, users were interviewed to evaluate the system's performance with a user satisfaction survey. The survey probed different aspects of the users' perception of their interaction with the system. We asked the users to evaluate a set of five core metrics on a 5-point Likert scale. We followed [Walker *et al.*, 2002] definition of the overall user satisfaction as the sum of text-to-speech synthesis performance, task ease, user expertise, overall difficulty and future use. The mean for user satisfaction across all dialogues was 15.0 (with a standard derivation of 2.9). [8] A one-way ANOVA for user satisfaction between wizards (df=5, F=1.52 p=0.05) shows no significant difference across wizards, meaning that the system performance was judged to be about equally good for all wizards.

To measure task experience we elicited data on perceived task success and satisfaction on a 5-point Likert scale after each task was completed. For all the subjects the final perceived task success was 4.4 and task satisfaction 3.9 across the 4 tasks each subject had to complete. For task success as well as for task satisfaction no significant variance across wizards was detected.

Furthermore the subjects were asked about the employed multi-modal presentation and clarification strategies.

The clarification strategies employed by the wizards seemed to be successful: From the subjects' point of view, mutual understanding was very good and the few misunderstandings could be easily resolved. Nevertheless, in the case of disambiguation requests and when grounding an utterance, subjects ask for more display feedback. It is interesting to note that subjects judged understanding difficulties on higher levels of interpretation (especially reference resolution problems and problems with interpreting the intention) to be more costly than problems on lower levels of understanding (like the acoustic understanding). For the clarification strategy this

[8] [Walker *et al.*, 2002] reported an average user satisfaction of 16.2 for 9 Communicator systems.

implies that the system should engage in clarification at the lowest level a error was detected.[9]

Multi-modal presentation strategies were perceived to be helpful in general, having a mean of 3.1 on a 5-point Likert scale. However, the subjects reported that too much information was being displayed especially for the tasks with driving. 85.7% of the subjects reported that the screen output was sometimes distracting them. 76.2% of the subjects would prefer to more verbal feedback, especially while driving. On a 3-point Likert scale subjects evaluated the amount of the information presented verbally to be about right (mean of 1.8), whereas they found the information presented on the screen to be too much (mean of 2.3). Studies by [Bernsen and Dybkjaer, 2001] on the appropriateness of using verbal vs. graphical feedback for in-car dialogues indicate that the need for text output is very limited. Some subjects in that study, as well subjects in our study report that they would prefer to not have to use the display at all while driving. On the other hand subjects in our study perceived the screen output to be very helpful in less stressful driving situations and when not driving (e.g. for memory assistance, clarifications etc.). Especially when they want to verify whether a complex task was finally completed (e.g. building a playlist), they ask for a displayed proof. For modality selection in in-car dialogues the driver's mental workload on primary and secondary task has to be carefully evaluated with respect to a situation model.

With respect to multi-modality subjects also asked for more personalized data presentation. We therefore need to develop intelligent ways to reduce the amount of data being displayed. This could build on prior work on the generation of "tailored" responses in spoken dialogue according to a user model [Moore *et al.*, 2004].

The results for multi-modal feedback behavior showed no significant variations across wizards except for the general helpfulness of multi-modal strategies. An ANOVA Planned Comparison of the wizard with the lowest mean against the other wizards showed that his behavior was significantly worse. It is interesting to note, that this wizard was using the display less than the others. We might consider not to include the 4 sessions with this wizard in our output generation model.

We also tried to analyze in more detail how the wizards' presentation strategies influenced the results. The option which was chosen most of the time was to present a table with the search results (78.6%); to present a list was only chosen in 17.5% of the cases and text only 0.04%. The wizards' choices varied significantly only for presenting the table option. The wizard who was rated lowest for multimodality was using the table option less, indicating that this option should be used more often. This is also supported by the fact that the show table option is the only presentation strategy which is positively correlated to how the user evaluated multimodality (Spearman's r = 0.436*). We also could find a 2-tailed correlation between user satisfaction and multimodality judgment (Spearman's r = 0.658**). This indicates the importance of good multimodal presentation strategies for user satisfaction.

Finally, the subjects were asked for own comments. They liked to be able to provide vague information, e.g., ask for "an oldie", and were expecting collaborative suggestions. They also appreciated collaborative proposals based on inferences made from previous conversations.

In sum, as the measures for user satisfaction, task experience, and multi-modal feedback strategies, the subjects' judgments show a positive trend. The dialogue strategies employed by most of the wizards seem to be a good starting point for building a baseline system. Furthermore, the results indicate that intelligent multi-modal generation needs to be adaptive to user and situation models.

5 Conclusions and Future Steps

We have presented an experiment setup that enables us to gather multimodal interaction data aimed at studying not only the behavior of the users of the simulated system, but also that of the wizards. In order to simulate a dialogue system interaction, the wizards were only shown transcriptions of the user utterances, sometimes corrupted, to simulate automatic speech recognition problems. The wizard's utterances were also transcribed and presented to the user through a speech synthesizer. In order to make it possible for the wizards to produce contextually varied screen output in real time, we have included a screen output planning module which automatically calculated several screen output versions every time the wizard ran a database query. The wizards were free to speak and/or display screen output. The users were free to speak or select on the screen. In a part of each session, the user was occupied by a primary driving task.

The main challenge for an experiment setup as described here is the considerable delay between user input and wizard response. This is due partly to the transcription and spelling correction step and partly due to the time it takes the wizard to decide on and enter a query to the database, then select a presentation and in parallel speak to the user. We have yet to analyze the exact distribution of time needed for these tasks. Several ways can be chosen to speed up the process. Transcription can be eliminated either by using speech recognition and dealing with its errors, or instead applying signal processing software, e.g., to filter out prosodic information from the user utterance and/or to transform the wizard's utterance into synthetically sounding speech (e.g., using a vocoder). Database search can be sped up in a number of ways too, ranging from allowing selection directly from the transcribed text to automatically preparing default searches by analyzing the user's utterance. Note, however, that the latter will most likely prejudice the wizard to stick to the proposed search.

We plan to annotate the corpus, most importantly w.r.t. wizard presentation strategies and context features relevant for the choice between them. We also plan to compare the presentation strategies to the strategies in speech-only mode, for which we collected data in an earlier experiment (cf. [Kruijff-Korbayová *et al.*, 2005]).

For clarification strategies previous studies already showed

[9] Note that engaging at the lowest level just helps to save dialogue "costs". Other studies have shown that user satisfaction is higher for strategies that would "hide" the understanding error by asking questions on higher levels [Skantze, 2003], [Raux *et al.*, 2005]

that the decision process needs to be highly dynamic by taking into account various features such as interpretation uncertainties and local utility [Paek and Horvitz, 2000]. We plan to use the wizard data to learn an initial multi-modal clarification policy and later on apply reinforcement learning methods to the problem in order to account for long-term dialogue goals, such as task success and user satisfaction.

The screen output options used in the experiment will also be employed in the baseline system we are currently implementing. The challenges involved there are to decide (i) when to produce screen output, (ii) what (and how) to display and (iii) what the corresponding speech output should be. We will analyze the corpus in order to determine what the suitable strategies are.

References

[Bernsen and Dybkjaer, 2001] Niels Ole Bernsen and Laila Dybkjaer. Exploring natural interaction in the car. In *CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation*, 2001.

[Bernsen *et al.*, 1998] N. O. Bernsen, H. Dybkjær, and L. Dybkjær. *Designing Interactive Speech Systems — From First Ideas to User Testing*. Springer, 1998.

[Duran *et al.*, 2001] Christine Duran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Aalborg, 1-2 September 2001*, pages 48–57, 2001.

[Kruijff-Korbayová *et al.*, 2005] Ivana Kruijff-Korbayová, Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Michael Kaißer, Peter Poler, Jan Schehl, and Verena Rieser. Presentation strategies for flexible multimodal interaction with a music player. In *Proceedings of DIALOR'05 (The 9th workshop on the semantics and pragmatics of dialogue (SEMDIAL)*, 2005.

[Martin *et al.*, 1999] D. L. Martin, A. J. Cheyer, and D. B. Moran. The open agent architecture: A framework for building distributed software systems. *Applied Artificial Intelligence: An International Journal*, 13(1–2):91–128, Jan–Mar 1999.

[Mattes, 2003] Stefan Mattes. The lane-change-task as a tool for driver distraction evaluation. In *Proceedings of IGfA*, 2003.

[Moore *et al.*, 2004] Johanna D. Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Sociey Conference, AAAI Press*, 2004.

[Paek and Horvitz, 2000] Tim Paek and Eric Horvitz. Conversation as action under uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000.

[Raux *et al.*, 2005] Antoine Raux, Brian Langner, Dan Bohus, Allan W. Black, and Maxine Eskenazi. Let's go public! taking a spoken dialog system to the real world. 2005.

[Skantze, 2003] Gabriel Skantze. Exploring human error handling strategies: Implications for spoken dialogue systems. In *Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[Stuttle *et al.*, 2004] Matthew Stuttle, Jason Williams, and Steve Young. A framework for dialogue data collection with a simulated asr channel. In *Proceedings of the ICSLP*, 2004.

[Türk, 2001] Ulrich Türk. The technical processing in smartkom data collection: a case study. In *Proceedings of Eurospeech2001, Aalborg, Denmark*, 2001.

[Walker *et al.*, 2002] Marylin Walker, R. Passonneau, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sandersa, S. Seneff, D. Stallard, and S. Whittaker. Cross-site evaluation in darpa communicator: The june 2000 data collection. 2002.

[Williams and Young, 2004] Jason D. Williams and Steve Young. Characterizing task-oriented dialog using a simulated asr channel. In *Proceedings of the ICSLP*, 2004.