Presentation Strategies for Flexible Multimodal Interaction with a Music Player

Ivana Kruijff-Korbayová, Nate Blaylock, Ciprian Gerstenberger, Verena Rieser
Department of Computational Linguistics
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 present an ongoing project building a multimodal dialogue system for a music player supporting natural, flexible interaction and collaborative behavior. Since the system functionalities include searching a big MP3 database, multimodal output is needed.

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. The system functionalities include playback control, manipulation of playlists, and searching a large MP3 database. We aim at a system that will engage in natural, flexible interaction and collaborative behavior. We believe that in order to achieve this, the system needs to provide advanced adaptive multimodal output.

To determine the interaction strategies and range of linguistic behavior that humans naturally use in the music player scenario, we have conducted Wizard-of-Oz experiments. Our goal was not only to collect data on how potential users interact with such a system, but also (and importantly) to observe what range of interaction strategies humans naturally use and how efficient they are. We therefore used a setup where the wizard had freedom of choice w.r.t. their response and its realization in single or multiple modalities.

When developing our system, we design the multimodal output presentation strategies and the range of linguistic realization options based on experience gathered during the experiment and an analysis of the corpus.

We briefly describe our experiments and the collected data (Section 2), present initial observations on the presentation of database search results in speech and on screen (Section 3), and sketch the main system components involved output generation (Section 4).

2 SAMMIE Data Collection

We conducted two series of data-collection experiments: SAMMIE-1 involved only spoken interaction, SAMMIE-2 was multimodal, with speech and screen input and output.[2].

In both experiments, the users performed several tasks, such as finding a song or an album and playing it or adding it to a playlist. In some tasks, the users were given rather concrete specifications, such as a name (e.g, *Play Crazy by Aerosmith*), in other tasks they got more vague characteristics, such as period,

[1]TALK (Talk and Look: Tools for Ambient Linguistic Knowledge; http://www.talk-project.org), funded by the EU 6th Framework Program, project No. IST-507802.

[2]SAMMIE stands for Saarbrücken Multimodal MP3 Player Interaction Experiment.

genre or type of music (e.g., *Play a pop song from 2004*, or *Make a playlist with 4 of your favorite songs*). This resulted in interactions where the users were exploring the database contents and adding search criteria depending on what was found.

In SAMMIE-1, there were 24 subjects, who each participated in one session with one of two wizards. Each subject worked on eight tasks, for maximally 30 minutes in total. Tasks were of three types: finding a specified title, selecting a title satisfying certain constraints and building a playlist satisfying certain constraints.

In SAMMIE-2, there were 24 subjects, who each participated in one session with one of six wizards. Each subject worked on two times two tasks.[3] 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. (See (Kruijff-Korbayová et al., 2005) for details.)

The wizards, playing the role of the music player, had 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.[4] We used multiple wizards and gave them freedom to decide about their response and its realization in order to collect data with a variety of interaction strategies.

Both users and wizards could speak freely. The interactions were in German (although most of the titles and artist names in the database are English). In the multimodal setup in SAMMIE-2, the wizards could use speech only, display only, or to combine speech and display, and the users could speak and/or make selections on the screen.

[3]For the second two tasks there was a primary task using a *Lane Change* driving simulator (Mattes, 2003).

[4]FreeDB is freely available at http://www.freedb.org

Since the wizard cannot design screens on the fly, because that would take too long, we implemented modules supporting the wizard by providing automatically calculated screen output options the wizard could select from.

The types of screen output were: (i) a simple text-message conveying how many results were found, (ii) a list of just the names (of albums, songs or artists) with the Bcorresponding number of matches, (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 used heuristics based on the search to decide, e.g., which columns should be displayed. The wizard could chose one of the offered options to display to the user, or decide to clear the user's screen. Otherwise, the user's screen remained unchanged.

We are currently analyzing and annotating the data w.r.t. the interaction strategies and other aspects. The interaction strategies observed in the collected data are driving the design of turn- and sentence-planning (cf. Section 4). We also interviewed both the 'subjects and the wizards after the experiments individually. Their feedback provides us with additional insight concerning the output generation decisions made by the wizards and how successful they were according to the users.

3 Search Results Presentation

Here we present preliminary observations on the presentation of database search results. In speech-only interaction, the wizards typically say the number of results and list them, when the number is small (up to approx. 10, cf. (1)). For more results, they often say the number, and sometimes ask whether or not to list them (cf. (2)). For very large sets of results, the wizards typically say the number and ask the user to narrow down the search, (cf. (3)).

(1) I found 3 tracks. Blackbird, Michelle and Yesterday.
(2) I found 17 tracks. Should I list them?
(3) I found 500 tracks. Please constrain the search.

In multimodal interaction, a commonly used pattern is to simultaneously display screen output and describe what is shown (e.g., *I'll show you the songs by Prince*). Some wizards adapted to the user's requests: if asked to show something (e.g., *Show me the songs by Prince*), they showed it without verbal comments; but if asked a question (e.g., *What songs by Prince are there?* or *What did you find?*), they answered in speech as well as showed the screen output.

"Summaries" A common characteristic in both setups is that the wizards often verbally summarize the search results in some way: most commonly by just reporting the number of results found, as in (3). But sometimes they describe the similarities or differences between the results, as in (4).

(4) 200 are from the 70's and 300 from the 80's.

Such descriptions may help the user to make a choice, and are a desirable type of collaborative behavior for a system. Their automatic generation provides an interesting challenge: It requires the clustering of results, abstraction over specific values and the production of corresponding natural language realization. We are working on static cluster definitions (e.g., production years, genre, album names, etc.), and define suitable ways of referring to them in the turn and sentence planners (e.g., reference to decades). Clusters could also be computed dynamically, which poses two challenges: (a) deciding which clusters are most useful to the user (depending, e.g., on a user model); (b) automatically generating cluster descriptions.

Screen Output Options There were differences in how the wizards rated and used the different screen output options: The table containing most detailed information about the queried song(s) or album(s) was rated best and shown most often by some, while others thought it contained too much information and hence they used it less or never. The screen option containing 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. 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.

The subjects found the multi-modal presentation strategies helpful in general. However, they often thought that too much information was displayed. They found it distracting, especially while driving. They also asked for more personalized data presentation. We therefore need to develop intelligent ways to reduce the amount of data 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).

4 System Components

In this section, we briefly describe the components that are involved in output generation as part of the end-to-end dialogue system for the MP3 player domain we are developing.

Dialogue Management The dialogue manager is based on an agent-based model which views dialogue as collaborative problem-solving (Blaylock et al., 2003). It is implemented in the information-state update approach using DIPPER.[5] Utterances are viewed as negotiation of a shared collaborative problem-solving state, to do things such as determining joint objectives, finding and

[5]DIPPER is available at http://www.ltg.ed.ac.uk/dipper/

instantiating recipes to accomplish them, executing the recipes and monitoring for success.

Turn Planning In monomodal dialogue systems the propositional content is typically realized rather straightforwardly, producing written or spoken output w.r.t. to the issues of *what to say* and *how to say it*. In multimodal dialogue the relationship between the propositional content determined by the dialogue manager and the content realized as output is more complex as the content needs to be reasonably distributed over the available modalities in contextually appropriate ways. This also means that planning multimodal output needs to comprise the issue of *when to present what* according to the available modalities. To meet these challenges, our implementation of the turn planning component is based on a production rule system called *PATE*. Originally developed for the integration of multimodal input (Pfleger, 2004), this component provides an efficient and elegant way of realizing complex processing rules.

Sentence Planning and Realization Our sentence planner is also being implemented in PATE. One of its tasks is to plan the verbal summaries discussed in Section 3. It is also responsible for decisions pertaining to contextualized linguistic realization, such as information structure and referring expressions. Regarding sentence realization, the requirement of contextually appropriate spoken output calls for tools that allow for controlled variation in, e.g, syntactic structure and intonation. We use the OpenCCG system[6] for parsing and generation, and develop a German grammar for it (Gerstenberger and Wolska, 2005).

Speech Synthesis To produce spoken output in German we use the TTS system Mary[7], which enables us to produce contextually appropriate synthesized spoken output by controlling the intonation using a markup based on the German version of the ToBI standard.[8]

Screen Output We are using the generic table presentation tool we developed for the experiment to display tables, lists or text messages generated from the search results. The user can also graphically select items from the respective presentation. For use in the in-car system this table presenter is being adapted to the constraints of the driving situation, e.g., small display with large fonts and a limited number of rows. We are also adding a GUI for controling the MP3 player.

Later in the project, we will perform usability tests, where standard measures such as user satisfaction and task success will be used. The presentation strategies will be tested and evaluated in more specialized experiments with human judges comparing alternative outputs in specific contexts.

References

[Blaylock et al.2003] N. Blaylock, J. Allen, and G. Ferguson. 2003. Managing communicative intentions with collaborative problem solving. In *Current and New Directions in Discourse and Dialogue*, pages 63–84. Kluwer, Dordrecht.

[Gerstenberger and Wolska2005] C. Gerstenberger and M. Wolska. 2005. Introducing Topological Field Information into CCG. In Proc. of the ESSLLI 2005 Student Session, Edinburgh. To appear..

[Kruijff-Korbayová et al.2005] I. Kruijff-Korbayová, N. Blaylock, C. Gerstenberger, V. Rieser, T. Becker, M. Kaißer, P. Poller, and J. Schehl. 2005. An experiment setup for collecting data for adaptive output planning in a multimodal dialogue system. Submitted.

[Mattes2003] S. Mattes. 2003. The lane-change-task as a tool for driver distraction evaluation. In *Proc. of IGfA*.

[Moore et al.2004] J. D. Moore, M. E. Foster, O. Lemon, and M. White. 2004. Generating tailored, comparative descriptions in spoken dialogue. In *Proc. of the Seventeenth International Florida Artificial Intelligence Research Sociey Conference, AAAI Press.*

[Pfleger2004] N. Pfleger. 2004. Context based multimodal fusion. In *ICMI '04: Proc. of the 6th international conference on Multimodal interfaces*, pages 265–272, New York, NY, USA. ACM Press.

[6] OpenCCG is available at http://openccg.sourceforge.net/

[7] Mary TTS is available at http://mary.dfki.de/

[8] http://www.uni-koeln.de/phil-fak/phonetik/gtobi/