The SAMMIE System: Multimodal In-Car Dialogue

Tilman Becker, Peter Poller, Jan Schehl
DFKI
First.Last@dfki.de

Nate Blaylock, Ciprian Gerstenberger, Ivana Kruijff-Korbayová
Saarland University
talk-mit@coli.uni-sb.de

Abstract

The SAMMIE[1] system is an in-car multimodal dialogue system for an MP3 application. It is used as a testing environment for our research in natural, intuitive mixed-initiative interaction, with particular emphasis on multimodal output planning and realization aimed to produce output adapted to the context, including the driver's attention state w.r.t. the primary driving task.

1 Introduction

The SAMMIE system, developed in the TALK project in cooperation between several academic and industrial partners, employs the Information State Update paradigm, extended to model collaborative problem solving, multimodal context and the driver's attention state. We performed extensive user studies in a WOZ setup to guide the system design. A formal usability evaluation of the system's baseline version in a laboratory environment has been carried out with overall positive results. An enhanced version of the system will be integrated and evaluated in a research car.

In the following sections, we describe the functionality and architecture of the system, point out its special features in comparison to existing work, and give more details on the modules that are in the focus of our research interests. Finally, we summarize our experiments and evaluation results.

2 Functionality

The SAMMIE system provides a multi-modal interface to an in-car MP3 player (see Fig. 1) through speech and haptic input with a BMW iDrive input device, a button which can be turned, pushed down and sideways in four directions (see Fig. 2 left). System output is provided by speech and a graphical display integrated into the car's dashboard. An example of the system display is shown in Fig. 2.

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

Figure 1: User environment in laboratory setup.

The MP3 player application offers a wide range of functions: The user can control the currently playing song, search and browse an MP3 database by looking for any of the fields (song, artist, album, year, etc.), search and select playlists and even construct and edit playlists.

The user of SAMMIE has complete freedom in interacting with the system. Input can be through any modality and is not restricted to answers to system queries. On the contrary, the user can give new tasks as well as any information relevant to the current task at any time. This is achieved by modeling the interaction as a collaborative problem solving process, and multi-modal interpretation that fits user input into the context of the current task. The user is also free in their use of multimodality: SAMMIE handles deictic references (e.g., *Play this title* while pushing the iDrive button) and also cross-modal references, e.g., *Play the third song (on the list)*. Table 1 shows a typical interaction with the SAMMIE system; the displayed song list is in Fig. 2. SAMMIE supports interaction in German and English.

3 Architecture

Our system architecture follows the classical approach (Bunt et al., 2005) of a pipelined architecture with multimodal interpretation (fusion) and

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 57–60,
Sydney, July 2006.

U:	Show me the Beatles albums.
S:	I have these four Beatles albums. [shows a list of album names]
U:	Which songs are on this one? [selects the Red Album]
S:	The Red Album contains these songs [shows a list of the songs]
U:	Play the third one.
S:	[music plays]

Table 1: A typical interaction with SAMMIE.

fission modules encapsulating the dialogue manager. Fig. 2 shows the modules and their interaction: Modality-specific recognizers and analyzers provide semantically interpreted input to the multimodal fusion module that interprets them in the context of the other modalities and the current dialogue context. The dialogue manager decides on the next system move, based on its model of the tasks as collaborative problem solving, the current context and also the results from calls to the MP3 database. The turn planning module then determines an appropriate message to the user by planning the content, distributing it over the available output modalities and finally co-ordinating and synchronizing the output. Modality-specific output modules generate spoken output and graphical display update. All modules interact with the extended information state which stores all context information.

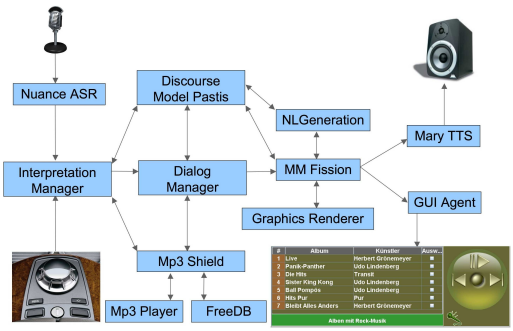

Figure 2: SAMMIE system architecture.

Many tasks in the SAMMIE system are modeled by a plan-based approach. Discourse modeling, interpretation management, dialogue management and linguistic planning, and turn planning are all based on the production rule system PATE[2] (Pfleger, 2004). It is based on some concepts of the ACT-R 4.0 system, in particular the goal-oriented application of production rules, the activation of working memory elements, and the weighting of production rules. In processing typed feature structures, PATE provides two operations that both integrate data and also are suitable for condition matching in production rule systems, namely a slightly extended version of the general *unification*, but also the discourse-oriented operation *overlay* (Alexandersson and Becker, 2001).

4 Related Work and Novel Aspects

Many dialogue systems deployed today follow a state-based approach that explicitly models the full (finite) set of dialogue states and all possible transitions between them. The VoiceXML[3] standard is a prominent example of this approach. This has two drawbacks: on the one hand, this approach is not very flexible and typically allows only so-called system controlled dialogues where the user is restricted to choosing their input from provided menu-like lists and answering specific questions. The user never is in control of the dialogue. For restricted tasks with a clear structure, such an approach is often sufficient and has been applied successfully. On the other hand, building such applications requires a fully specified model of all possible states and transitions, making larger applications expensive to build and difficult to test.

In SAMMIE we adopt an approach that models the interaction on an abstract level as collaborative problem solving and adds application specific knowledge on the possible *tasks*, available *resources* and known *recipes* for achieving the goals.

In addition, all relevant context information is administered in an Extended Information State. This is an extension of the Information State Update approach (Traum and Larsson, 2003) to the multi-modal setting.

Novel aspects in turn planning and realization include the comprehensive modeling in a single, OWL-based ontology and an extended range of context-sensitive variation, including system alignment to the user on multiple levels.

5 Flexible Multi-modal Interaction

5.1 Extended Information State

The information state of a multimodal system needs to contain a representation of contextual information about discourse, but also a representation of modality-specific information and user-specific information which can be used to plan system output suited to a given context. The over-

[2] Short for (P)roduction rule system based on (A)ctivation and (T)yped feature structure (E)lements.

[3] `http://www.w3.org/TR/voicexml20`

all information state (IS) of the SAMMIE system is shown in Fig. 3.

The contextual information partition of the IS represents the multimodal discourse context. It contains a record of the latest user utterance and preceding discourse history representing in a uniform way the salient discourse entities introduced in the different modalities. We adopt the three-tiered multimodal context representation used in the SmartKom system (Pfleger et al., 2003). The contents of the task partition are explained in the next section.

5.2 Collaborative Problem Solving

Our dialogue manager is based on an agent-based model which views dialogue as collaborative problem-solving (CPS) (Blaylock and Allen, 2005). The basic building blocks of the formal CPS model are problem-solving (PS) objects, which we represent as typed feature structures. PS object types form a single-inheritance hierarchy. In our CPS model, we define types for the upper level of an ontology of PS objects, which we term *abstract PS objects*. There are six abstract PS objects in our model from which all other domain-specific PS objects inherit: objective, recipe, constraint, evaluation, situation, and resource. These are used to model problem-solving at a domain-independent level and are taken as arguments by all update operators of the dialogue manager which implement conversation acts (Blaylock and Allen, 2005). The model is then specialized to a domain by inheriting and instantiating domain-specific types and instances of the PS objects.

5.3 Adaptive Turn Planning

The *fission* component comprises detailed content planning, media allocation and coordination and synchronization. Turn planning takes a set of CPS-specific conversational acts generated by the dialogue manager and maps them to modality-specific communicative acts.

Information on how content should be distributed over the available modalities (speech or graphics) is obtained from *Pastis*, a module which stores discourse-specific information. Pastis provides information about (i) the modality on which the user is currently focused, derived by the current discourse context; (ii) the user's current cognitive load when system interaction becomes a secondary task (e.g., system interaction while driving); (iii) the user's expertise, which is represented as a state variable. Pastis also contains information about factors that influence the preparation of output rendering for a modality, like the currently used language (German or English) or the display capabilities (e.g., maximum number of displayable objects within a table). Together with the dialogue manager's embedded part of the information state, the information stored by Pastis forms the *Extended Information State* of the SAMMIE system (Fig. 3).

Planning is then executed through a set of production rules that determine which kind of information should be presented through which of the available modalities. The rule set is divided in two subsets, domain-specific and domain-independent rules which together form the system's multimodal plan library.

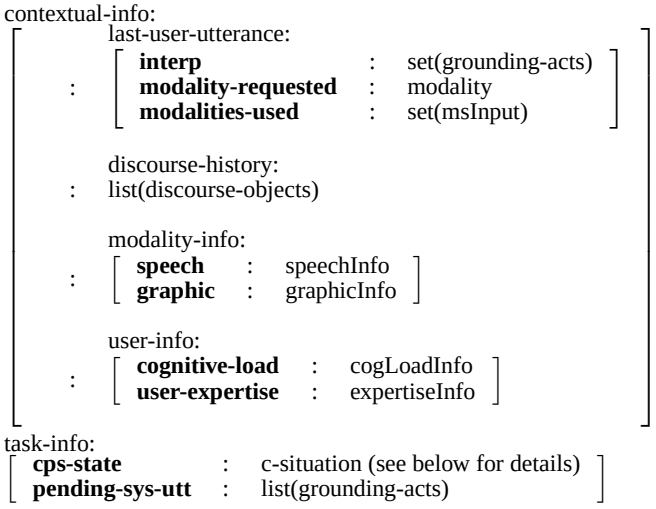

Figure 3: SAMMIE Information State structure.

5.4 Spoken Natural Language Output Generation

Our goal is to produce output that varies in the surface realization form and is adapted to the context. A template-based module has been developed and is sufficient for classes of system output that do not need fine-tuned context-driven variation. Our template-based generator can also deliver alternative realizations, e.g., alternative syntactic constructions, referring expressions, or lexical items. It is implemented by a set of straightforward sentence planning rules in the PATE system to build the templates, and a set of XSLT transformations to yield the output strings. Output in German and English is produced by accessing different dictionaries in a uniform way.

In order to facilitate incremental development of the whole system, our template-based module has a full coverage wrt. the classes of sys-

tem output that are needed. In parallel, we are experimenting with a linguistically more powerful grammar-based generator using OpenCCG[4], an open-source natural language processing environment (Baldridge and Kruijff, 2003). This allows for more fine-grained and controlled choices between linguistic expressions in order to achieve contextually appropriate output.

5.5 Modeling with an Ontology

We use a full model in OWL as the knowledge representation format in the dialogue manager, turn planner and sentence planner. This model includes the entities, properties and relations of the MP3 domain–including the player, data base and playlists. Also, all possible tasks that the user may perform are modeled explicitly. This task model is *user centered* and not simply a model of the application's API.The OWL-based model is transformed automatically to the internal format used in the PATE rule-interpreter.

We use multiple inheritance to model different views of concepts and the corresponding presentation possibilities; e.g., a *song* is a *browsable-object* as well as a *media-object* and thus allows for very different presentations, depending on context. Thereby PATE provides an efficient and elegant way to create more generic presentation planning rules.

6 Experiments and Evaluation

So far we conducted two WOZ *data collection* experiments and one *evaluation* experiment with a baseline version of the SAMMIE system. The SAMMIE-1 WOZ experiment involved only spoken interaction, SAMMIE-2 was multimodal, with speech and haptic input, and the subjects had to perform a primary driving task using a Lane Change simulator (Mattes, 2003) in a half of their experiment session. The wizard was simulating an MP3 player application with access to a large database of information (but not actual music) of more than 150,000 music albums (almost 1 million songs). In order to collect data with a variety of interaction strategies, we used multiple wizards and gave them freedom to decide about their response and its realization. In the multimodal setup in SAMMIE-2, the wizards could also freely decide between mono-modal and multimodal output. (See (Kruijff-Korbayová et al., 2005) for details.)

We have just completed a user evaluation to explore the user-acceptance, usability, and performance of the baseline implementation of the SAMMIE multimodal dialogue system. The users were asked to perform tasks which tested the system functionality. The evaluation analyzed the user's interaction with the baseline system and combined objective measurements like task completion (89%) and subjective ratings from the test subjects (80% positive).

Acknowledgments This work has been carried out in the TALK project, funded by the EU 6th Framework Program, project No. IST-507802.

[4] `http://openccg.sourceforge.net`

References

[Alexandersson and Becker2001] J. Alexandersson and T. Becker. 2001. Overlay as the basic operation for discourse processing in a multimodal dialogue system. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, Washington, August.

[Baldridge and Kruijff2003] J.M. Baldridge and G.J.M. Kruijff. 2003. Multi-Modal Combinatory Categorial Grammar. In *Proceedings of the 10th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, April.

[Blaylock and Allen2005] N. Blaylock and J. Allen. 2005. A collaborative problem-solving model of dialogue. In Laila Dybkjær and Wolfgang Minker, editors, *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 200–211, Lisbon, September 2–3.

[Bunt et al.2005] H. Bunt, M. Kipp, M. Maybury, and W. Wahlster. 2005. Fusion and coordination for multimodal interactive information presentation: Roadmap, architecture, tools, semantics. In O. Stock and M. Zancanaro, editors, *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology*, pages 325–340. Kluwer Academic.

[Kruijff-Korbayová et al.2005] I. Kruijff-Korbayová, T. Becker, N. Blaylock, C. Gerstenberger, M. Kaißer, P. Poller, J. Schehl, and V. Rieser. 2005. An experiment setup for collecting data for adaptive output planning in a multimodal dialogue system. In *Proc. of ENLG*, pages 191–196.

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

[Pfleger et al.2003] N. Pfleger, J. Alexandersson, and T. Becker. 2003. A robust and generic discourse model for multimodal dialogue. In *Proceedings of the 3rd Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco.

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

[Traum and Larsson2003] David R. Traum and Staffan Larsson. 2003. The information state approach to dialog management. In *Current and New Directions in Discourse and Dialog*. Kluwer.