

Situated Multimodal Interaction in SmartKom

Norbert Reithinger, Gerd Herzog, Alassane Ndiaye

German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{reithinger,herzog,ndiaye}@dfki.de

Abstract

In this paper, we give a short overview of the SMARTKOM system, a flexible and adaptive multimodal dialog system. The system uses a common dialog backbone to realize three interaction scenarios. Through the use of generic methods and knowledge sources for multimodal fusion and fission, we were able to develop a generic dialog system for flexible multimodal interaction.

Key words: Dialog systems, Multimodality, Situated Interaction

1 Introduction

Intelligent multimodal interfaces support users to navigate in information spaces or to control their technical environment (Johnston et al., 2002). As shown by Oviatt and Cohen (2000) the combination of multiple modalities on the input and output side of an interactive system reduces communication errors; Wahlster (2003) speaks in this context of full symmetric multimodality. With SMARTKOM (www.smartkom.org) we look into the combination of multiple modalities for input and output to reach these goals (Wahlster et al., 2001; Wahlster, 2003).

This paper presents the basic ideas of the situated interaction metaphor we developed, the scenarios covered by SMARTKOM, and the principles behind the development of the system (Section 2). In Section 3, we present a sketch of the multimodal processing in the dialog backbone, which is responsible for the processing of the ongoing interaction with the user, with an emphasis on the modality processing and a short example.

2 SmartKom's Situated Delegation-Oriented Dialog Paradigm

SMARTKOM is a mixed-initiative dialog system that provides full symmetric multimodality by combining speech, gesture, and facial expressions for both user input and system output (Wahlster et al., 2001; Wahlster, 2003). The system aims to provide an anthropomorphic and affective user interface through its personification of an embodied conversational agent, called Smartakus. The interaction metaphor is based on the so-called *situated, delegation-oriented dialog paradigm*. The basic idea is that the user delegates a task to a virtual communication assistant which is visualized as a life-like character. The interface agent recognizes the user's intentions and goals, asks the user for feedback if necessary, accesses the various services on behalf of the user, and presents the results in an adequate manner. The reactions of the user for a presented solution are extracted from his facial expression or the prosodic features of the next utterance. If the user is not satisfied, the system tries to reassess its solution to better accommodate it to the user's needs.

As it is depicted in Figure 1, SMARTKOM realizes a flexible and adaptive shell for multimodal dialogs and addresses three different application scenarios:

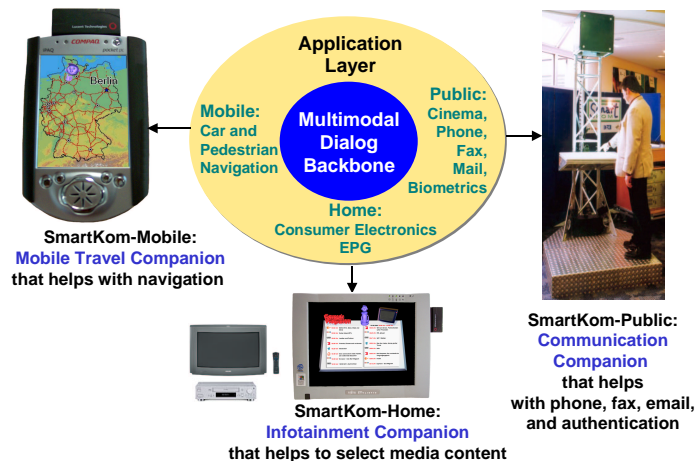


Fig. 1. SMARTKOM kernel and application scenarios.

SMARTKOM PUBLIC realizes an advanced multimodal information and communication kiosk for airports, train stations, or other public places. It supports users seeking for information concerning movie programs, offers reservation facilities, and provides personalized communication services using telephone, fax, or electronic mail. Before the system grants access to personal data, e.g. an address book, the user has to authenticate himself using either hand contour recognition, signature or voice verification. Speech input is captured with a directional microphone, facial expressions with a DV camera and gestures are tracked with an infrared camera. A video projector is used for the projection of graphical output onto a horizontal surface.

SMARTKOM HOME serves as a multimodal infotainment companion that helps to select media content and to operate various appliances. Using a portable webpad, the user is able to utilize the system as an electronic program guide or to easily control consumer electronics devices like a TV set or a VCR. Similar to the kiosk application, the user may also use communication services at home.

SMARTKOM MOBILE realizes a mobile travel companion for navigation and location-based services. It uses a PDA as a front end, which can be added to a car navigation system or is carried by a pedestrian. This application scenario comprises services like integrated trip planning and incremental route guidance. In the mobile scenario speech input can be combined with pen-based pointing. It also uses a simplified version of the Smartakus interface agent.

All functionalities, modality combinations and technical realizations including a wide variety of hardware options for the periphery are addressed by the same core dialog system with common shared knowledge sources. Its main underlying principles are the following:

- The processing uses a knowledge based, configurable approach: We do not implement special solutions for specific problems. There are no shortcuts or application specific procedural processing steps within the dialog core of the system.
- No processing and presentation without representation: For all multimodal inputs and outputs we use a common representation approach which allows us to use generic interaction models. The interaction processing is based on M3L (**M**ultimodal **M**arkup **L**anguage), a complete XML language designed in the context of SMARTKOM that covers all data interfaces within the complex multimodal dialog system (Gurevych et al., 2003).

The technical realization is based on the MULTIPLATFORM testbed (Herzog et al., 2003), an integration platform that provides a distributed component architecture. MULTIPLATFORM is implemented on the basis of the scalable and efficient publish/subscribe approach that decouples data producers and data consumers. Software modules communicate via so-called data pools that correspond to named message queues. Every data pool can be linked to an individual data type specification in order to define admissible message contents.

3 Integrated Modality Processing

SMARTKOM can be divided in two major parts: a transmutable multimodal dialog backbone that can engage in many different types of tasks in different

contexts, and the application functionalities that communicate with the backbone using a separate function modeling interface. Currently, 12 applications with over 50 functionalities can be freely addressed by the user. The analysis, understanding, and generation modules use a single semantic representation formalism, based on the W3C Ontology Interchange Language OIL (Gurevych et al., 2003). However, they do not use the knowledge source directly, but an XML schema which is automatically generated from the ontology and is part of M3L. Therefore, the modules can use the multitude of tools to process XML documents that are available for many programming languages. If two modules of the system exchange information, they use M3L documents using this XML-schema to represent the content. It is of particular importance that content and layout information for all visualized objects is also described using this knowledge source.



Fig. 2. Screenshot of the system reaction to the input “Show me tonight’s movie program”. Smartakus, the SMARTKOM life-like character is shown in the middle.

The analysis of various input modalities in SMARTKOM has to cope with uncertainty from speech and gesture recognition as well as with ambiguity when interpreting the meaning of the user’s input. For example, the gesture analyzer tracks the user’s hand or the stylus movement and passes the geometry data to the gesture analysis. It compares the geometry of the gestures with the representations of the current objects visible on the screen and produces a set of scored hypotheses about possible reference objects. Figure 2 shows a dynamically generated multimodal presentation based on the presentation content depicted partially in Figure 3 (see below). After the user asked for tonight’s movie program (for this case in Heidelberg), the system presents the program on the left-hand side and a map of Heidelberg with the cinema’s locations on the right-hand side. The user now can address items on the screen using speech and gestures, either unimodal or multimodal.

The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results. Since all recognizers and analyzers produce time-stamped hypotheses, the fusion process considers various temporal constraints. By fusing symbolic and statistical information SMARTKOM can correct recognition errors of the unimodal input components and provide a more robust interpretation. The modality fusion and interpretation process is augmented by SMARTKOM's multimodal discourse model. The discourse component adds contextual information and scores that state how well an interpretation hypothesis fits to the previous discourse (Pfleger et al., 2002). As soon as, e.g., the modality fusion component processes a referring expression that can not be combined with an unambiguous deictic gesture, it sends a request to the discourse component asking for reference resolution. If the resolution succeeds, the discourse component returns a completely instantiated domain object. The intention recognizer finally ranks the remaining interpretation hypotheses. The action planner takes this list and processes the best one fitting the current context to identify the user's intention, to select the appropriate request to an application, or to select items that the user has to be asked about (Löckelt et al., 2000).

The modality fission module in SMARTKOM receives abstract, modality independent presentation goals from the action planner. The multimodal output depends on several constraints, e.g., the current scenario, the display size, and user preferences like the currently applicable modality mix (Müller et al., 2002). The fission module applies presentation strategies that decompose the complex presentation goal into presentation tasks. It also decides whether an object description is to be uttered verbally or graphically. The result is a presentation script that is passed to the display manager. A special sub-task of this planning process is the generation of animation scripts for the presentation agent including movements of the agent and its gestures. In the fission step the template-oriented natural language generation component gets all content descriptions based on the ontology of SMARTKOM that have to be uttered verbally. The output consists of text annotated with conceptual structures for a concept-to-speech synthesizer. The synthesizer not only generates the audio signal, but also a phoneme transcription that is passed back to the animation of Smartakus. This allows for a lip-synchronous visualization while it is talking.

Finally, the display manager creates the visual appearance for the presentation content using predefined graphical layout elements, e.g. for a cinema program, or for maps. They are filled dynamically with content elements from the action planner's output, e.g., movie titles, or maps with locations and names of the cinemas. Since many objects can be shown at the same time on the display, the manager re-arranges the objects on the screen and removes objects, if necessary. It also integrates the animated agent into the presentation and coordinates the animations with the graphical and the speech output. The

```

<?xml- version="1.0"?>
<presentationContent>
[...]
<abstractPresentationContent>
<performance id="PP325">
<beginTime> 2003-06-04T20:15:00 </beginTime>
<avMedium>
<avType>action</avType> <avType>scienceFiction</avType>
<title> Matrix: Reloaded </title>
</avMedium>
<cinema>
<name>Saal 1</name>
<partOf>
<movieTheater><name>Lux/Harmonie</name></movieTheater>
[...]
</cinema>
</performance>
</abstractPresentationContent>
[...]
<panelElement>
<label id="PM14">
<boudingShape>
<leftTop><x>0.0546675</x><y>0.6484375</y> </leftTop>
<rightBottom><x>0.40625</x><y>6888020834</y></rightBottom>
</boudingShape>
<contentReference>PP325</contentReference>
<text> Matrix: Reloaded; Action, Science Fiction;
(Lux/Harmonie, 20:15 Uhr) </text>
</label> </panelElement> [...]
</presentationContent>

<abstractPresentationContent>
<mapLocation id="PP717">
<locationName> Lux/Harmonie </locationName>
<objectTypes>
<objectType> cinema </objectType>
</objectTypes>
<geometries>
<point>
<x> 3478601.0 </x> <y> 5474947.0 </y>
</point>
</geometries>
</mapLocation>
[...]
</abstractPresentationContent>
[...]
<label>
<boudingShape>
<leftTop>
<x>0.732421875</x> <y>0.39322916666666663</y>
</leftTop>
<rightBottom>
<x> 0.7421875 </x> <y> 0.40625 </y>
</rightBottom>
</boudingShape>
<contentReference>PP717</contentReference>
<text> Lux/Harmonie </text>
</label>
[...]
</presentationContent>

```

Fig. 3. Simplified, partial M3L structure with visualization representations.

last activity of the display manager in an interaction cycle is finally to update the visualization description: for each element visible on the screen it posts the geometry and the link to SMARTKOM's object description using the XML based knowledge representation. Figure 3 shows parts of the representation corresponding to the entry for the movie *Matrix: Reloaded* in the left part of Figure 2 and for the label for the cinema *Lux/Harmonie* as marked in the map on the right part. The bounding shapes are defined using the top left and lower right corners. The measurements are percentages of the screen width and height: this enables for a representation independent from the actual output device. The geometries-entry for the cinema contains Gauss-Krueger coordinates of its location and can be used later on for other activities like, e.g., route planning.

4 Conclusion

We presented an overview of SMARTKOM, its scenarios and the approach for multimodal processing. The system was developed in the framework of the German collaborative Human Technology Interaction program over the last 4 years, starting in 1999. We now have a fully functional demonstrator installed at various sites. The knowledge based component approach to multimodal dialog interaction proved to be a sound basis for the integration of new applications. Scientific results and software components are re-used in other research projects such as MIAMM (www.miamm.org), COMIC (www.hcrc.ed.ac.uk/comic), and VirtualHuman (www.virtual-human.org).

Acknowledgements

We would like to thank the SMARTKOM developers at DFKI, and the partners in the project consortium: DaimlerChrysler AG, European Media Laboratory GmbH, Friedrich-Alexander University Erlangen-Nürnberg, International Computer Science Institute, Ludwig-Maximilian University München, MediaInterface GmbH, Philips GmbH, Siemens AG, Sony International (Europe) GmbH, and Stuttgart University. This work was funded by the German Federal Ministry for Education and Research (BMBF) under grant 01 IL 905 K7.

References

- Gurevych, I., Porzel, R., Slinko, E., Pfeleger, N., Alexandersson, J., Merten, S., 2003. Less is more: Using a single knowledge representation in dialogue systems. In: Hirst, G., Nirenburg, S. (Eds.), Proc. of the HLT-NAACL 2003 Workshop on Text Meaning. Edmonton, Canada.
- Herzog, G., Kirchmann, H., Merten, S., Ndiaye, A., Poller, P., Becker, T., 2003. MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems. In: Cunningham, H., Patrick, J. (Eds.), Proc. of the HLT-NAACL 2003 Workshop: Software Engineering and Architecture of Language Technology Systems (SEALTS). Edmonton, Canada.
- Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., Maloor, P., 2002. Match: An architecture for multimodal dialogue systems. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).
- Löckelt, M., Becker, T., Pfeleger, N., Alexandersson, J., 2000. Making Sense in Partial. In: Bos, J. (Ed.), Proc. of the Sixth Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002). Edinburgh, Scotland.
- Müller, J., Poller, P., Tschernomas, V., 2002. Situated Delegation-Oriented Multimodal Presentation. In: Krüger, A., Malaka, R. (Eds.), Proc. of the AAAI-2002 Workshop on Intelligent Situation-Aware Media And Presentations (ISAMP 2002). Edmonton, Canada.
- Oviatt, S. L., Cohen, P. R., 2000. Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM* 43 (3), 45–53.
- Pfeleger, N., Alexandersson, J., Becker, T., 2002. Scoring Functions for Overlay and their Application in Discourse Processing. In: Proc. of KONVENS 2002. Saarbrücken, Germany.
- Wahlster, W., 2003. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In: Krahl, R., Günther, D. (Eds.), Proc. of the Human Computer Interaction Status Conference 2003. Berlin, Germany.
- Wahlster, W., Reithinger, N., Blocher, A., 2001. SmartKom: Multimodal Communication with a Life-Like Character. In: Proc. of Eurospeech'01. Aalborg, Denmark.