In: Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001, Vol. 3, pp. 1547 - 1550.

SmartKom: Multimodal Communication with a Life-Like Character

Wolfgang Wahlster Norbert Reithinger Anselm Blocher

DFKI GmbH, D-66123 Saarbrücken, Germany

{wahlster,reithinger,blocher}@dfki.de

Abstract

SmartKom is a multimodal dialog system that combines speech, gesture, and mimics input and output. Spontaneous speech understanding is combined with the video-based recognition of natural gestures. One of the major scientific goals of SmartKom is to design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level. SmartKom is based on the situated delegation-oriented dialog paradigm, in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display. We describe the SmartKom architecture, the use of an XML-based mark-up language for multimodal content, and some of the distinguishing features of the first fully operational SmartKom demonstrator.

1. Introduction

More effective, efficient, and natural interfaces to support the location-sensitive access to information, applications, and people are increasingly relevant in mobile and time-critical situations [4]. SmartKom (www.smartkom.org) is a multimodal dialog system that combines speech, gesture, and mimics input and output [5]. It supports the situated understanding of possibly imprecise, ambiguous, or partial multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations [1]. SmartKom's interaction management is based on representing, reasoning, and exploiting models of the user, domain, task, context and the media itself. One of the major scientific goals of SmartKom is to explore and design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level [2].

The main contractor of the SmartKom consortium is the German Research Center for Artificial Intelligence (DFKI) and W. Wahlster serves as the scientific project director. The major industrial partners involved in SmartKom are DaimlerChrysler, Philips, Siemens and Sony. The project was started in 1999 and will last for four years. SmartKom is funded by the German Federal Ministry for Education and Research (BMBF) and the industrial partners. The total budget is roughly 25 Million Euros. SmartKom is the follow-up project to Verbmobil (1993-2000) and reuses some of Verbmobil's components for the understanding of spontaneous dialogs [3]. In this paper we will present the main objectives of SmartKom, introduce the basic architecture and XML-based knowledge and interface descriptions, and present the first demonstrator system (see figure 1).

Figure 1: Interacting with the SmartKom Demonstrator System

2. Objectives of SmartKom

The main goal of SmartKom is to exploit one of the major characteristics of human interactions: the coordinated use of different code systems, like language, gesture, and mimics, to interact in complex environments. It is our goal to support the intuitive access to knowledge-rich services, using a mixed-initiative approach.

SmartKom merges three user interface paradigms, namely spoken dialogs, graphical interfaces, and gestural interactions, to achieve truly multimodal communication. Natural language interaction in SmartKom is based on speaker-independent speech understanding technology. For the graphical user interface and the gestural interaction we do not use a traditional WIMP (windows, icons, mouse pointer) interface, as we try to support natural gestural interaction. Technically, this is made possible by the SIVIT® virtual touch screen, a gesture recognition hardware and software system.

SmartKom's interaction metaphor breaks radically with the traditional desktop. SmartKom is based on the situated delegation-oriented dialog paradigm (SDDP), in which the user delegates a task to a virtual communication assistant, visible on the graphical display. Since for more complex tasks

this cannot be done in a simple command-and-control style, a collaborative dialog between the user and the agent, visualized as a life-like character, elaborates the specification of the delegated task and possible plans of the agent to achieve the user's intentional goal. In contrast to task-oriented dialogs, where the user carries out a task with the help of the system, in SDDP the user delegates a task to an agent and helps him to carry out this task, if necessary.

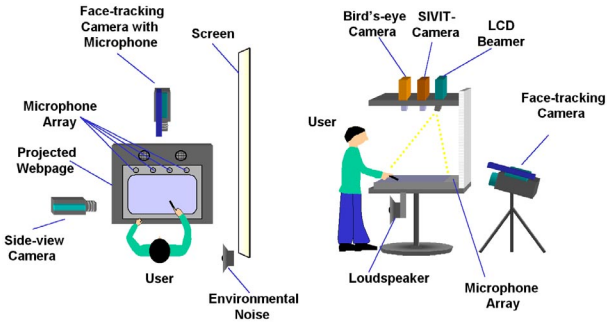

Figure 2: SmartKom's Multi-Channel Audio and Video Data Collection

An important research area of SmartKom is a massive data collection effort in order to get realistic data of the spontaneous use of advanced multimodal dialog systems. Multi-channel audio and video data from WOZ experiments have been transliterated, segmented and annotated, so that systematic conversation analysis becomes possible and statistical properties can be extracted from large corpora of coordinated speech, gestures, and mimics (see figure 2). The annotated SmartKom corpora are distributed to all project partners via DVDs and used as a basis for the functional and ergonomic design of the demonstrators as well as for the training of the various SmartKom components that are based on machine learning methods.

Three versions of SmartKom are defined for various application scenarios (see figure 3):

- SmartKom-Public: a multimodal communication booth for airport or train launches where travelers can get information on e.g. hotels, restaurants, entertainment facilities for the city they are visiting, and can access their personalized standard applications via broadband channels.
- SmartKom-Mobile: a PDA version for web access in a car as an add-on to a car navigation system and for pedestrians. The additional services are route planning and interactive navigation through a city, using GPS and GSM/UMTS connectivity.
- SmartKom-Home/Office: the system acts as a portal to information services, especially in connection with electronic programme guides (EPG) for TV, control of consumer electronics and access service to standard applications like phone, e-mail etc. The system can be used e.g. in the living room, where the user can operate the system in the lean-back mode merely by voice input or in the lean-forward mode through a portable WEB-pad with coordinated speech and gestures.

3. The SmartKom Demonstrator

In the first fully operational SmartKom demonstrator, that was released in December 2000 (see figures 1 and 4), the user can combine spontaneous speech and natural hand gestures for input. SmartKom will react with coordinated speech, gestures, graphics, and mimics of an autoanimated interface agent. The demonstrator consists of the SIVIT ® unit on top, which contains the LCD projector and the gesture recognition hardware, and a projection panel below, where the graphical user interface is projected. The system is controlled by three dual Pentium PCs, running Linux and Windows NT®. For the first demonstrator, we use a close-speaking microphone for speech input. The user is located in front of the projection panel and can use her hands and fingers to point onto the the visualized objects. There is no need to touch the panel, since the video-based gesture recognition unit tracks the location of the hand and fingers.

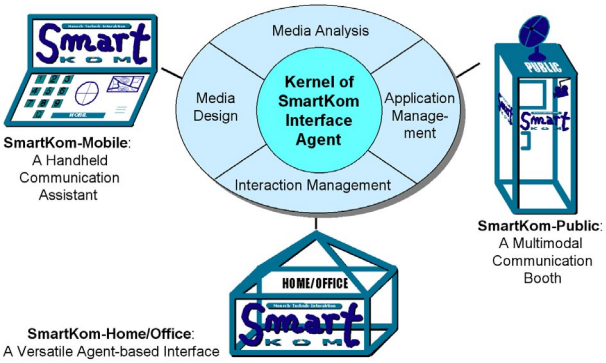

Figure 3: Three Instances of SmartKom

The demonstrator works in the web domains of electronic programme guides and online ticket reservation for movie theatres.

Figure 4 shows an example for a part of the current demonstrator's interface. The ongoing sub-dialog is about cinema seat reservation. The life-like character, called Smartakus, presents the seating layout of a cinema. The crosses mark all seats, that have been reserved. Referring to the graphical presentation, the system asks the user:

System: Please show me where your preferred seats are.[1]

The user then points to her preferred seats, accompanied by an utterance like:

User: I want these [↑] two seats

where [↑] stands for the gesture to the seats. Note that the user does not have to localize the seats by "clicking" on them, but she can use natural pointing gestures. The pointing with the index finger is ambiguous, since the demonstratum is an area between two seats. But the system disambiguates the gesture exploiting the semantics of the natural language input. Thus, the system recognizes the intended seats' positions,

[1] The current interaction language of SmartKom is German. The dialog examples are literal translations.

recognizes the intention of the user to reserve the seats and contacts the reservation agent through an external service. If the reservation is successful, Smartakus presents the floor plan with the seats marked as reserved.

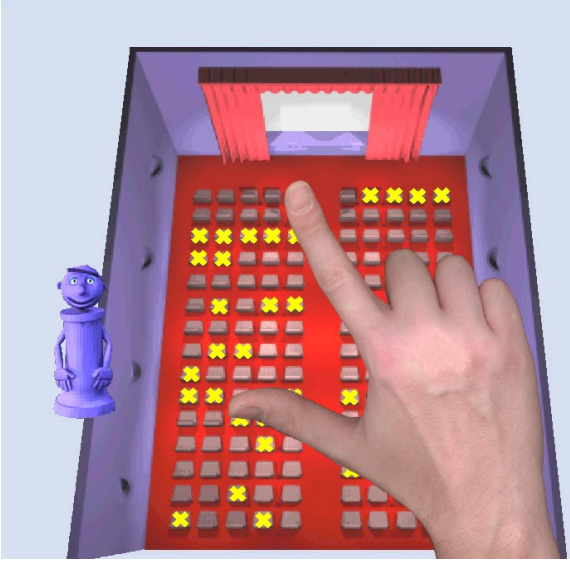

Figure 4: Natural Pointing Gesture onto a Projected Seating Layout

4. SmartKom's Architecture

Figure 5 shows the control GUI of the fully operational demonstrator system. It reflects the modular software structure of SmartKom . The modules can be grouped into

- Interface modules: on the input side we have the audio module, on the output side the display manager.
- Recognizers and synthesizers: on the input side, we have gesture recognition, prosody and speech recognition modules , on the output side speech synthesis and the display manager.
- Semantic processing modules: this group of modules comprises create meaning representations or transform them: gesture and speech analysis, media fusion, intention recognition, discourse and domain modeling, action planning, presentation planning, and concept-to-speech generation.
- External services: the function modeling module is the interface to external services, e.g. EPG databases, map services and information extraction from the Web.

SmartKom is based on a multi-blackboard architecture with parallel processing threads that support the media fusion and media design processes. All modules shown in figure 5 are realized as separate processes, running either on Linux or Windows NT®. They are implemented in C, C++, Java or Prolog. The underlying integration software is based Verbmobil's testbed [3] .

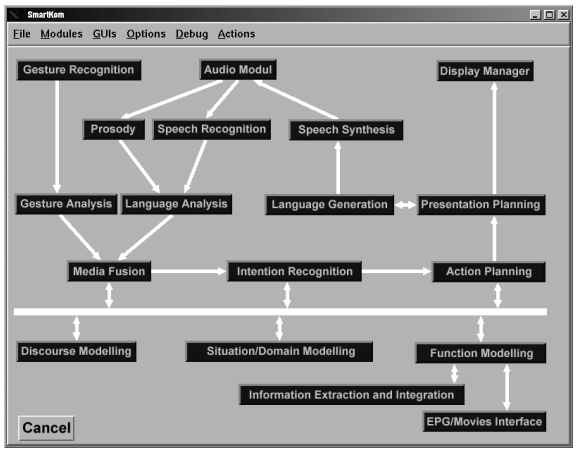

Figure 5: GUI for Tracing the Flow of Information in SmartKom

5. M3L as an XML-based Interface Language

A key design decision was the use of an XML-based markup language called M3L for representing all the information, which flows between the various processing components of SmartKom. For example, the word lattice and gesture lattice, the media fusion results, the presentation plan and the discourse context are all represented in M3L. M3L is designed for the representation and exchange of complex multimodal content as well as information about segmentation, synchronization and the confidence into processing results. For each communication pool, XML schemas are designed which allows for automatic data checking during transfer.

In figure 6 the presentation agent Smartakus presents a map of Heidelberg highlighting the location of cinemas. The discourse context, represented in M3L is

```
<presentationContent>
[...]
  <abstractPresentationContent>
    <movieTheater structId=pid3072>
      <entityKey> cinema_17a </entityKey>
      <name> Europa </name>
        <geoCoordinate>
          <x> 225 </x> <y> 230 </y>
        </geoCoordinate>
    </movieTheater>
  </abstractPresentationContent>
[...]
  <panelElement>
    <map structId="PM23">
      <boundingShape>
        <leftTop>
          <x> 0.5542 </x> <y> 0.1950 </y>
        </leftTop>
        <rightBottom>
          <x> 0.9892 </x> <y> 0.7068 </y>
        </rightBottom>
      </boundingShape>
      <contentRef>pid3072</contentRef>
    </map>
  </panelElement>
[...]
</presentationContent>
```

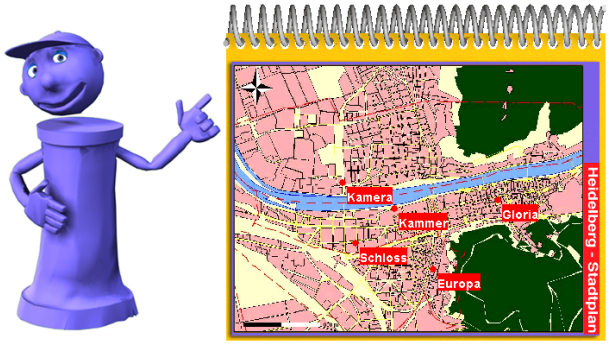

Figure 6: The Life-like Character's Pointing Gesture onto a Map Display Generated by SmartKom

The first element in the XML structure describes the cinema "Europa" with its real-world geo coordinates. The identifier **pid3072** links it to the description of the panel element, which also contains the relative coordinates on the presentation surface[1]. Let's assume an utterance of the user like

User: I want to make a reservation here [↑]

where [↑] stands for the gesture to the cinema "Europa". Then, both recognizers for gesture and speech deliver interpretation lattices. The gesture analysis processes coordinates from the SIVIT® unit and from the content representation. The resulting lattice contains hypotheses about possible referents of the gesture. In our example the result of the analysis is

```
<gestureAnalysis>
[...]
 <type>   tarrying  </type>
 <referencedObjects>
  <object>
   <displayObject>
    <contentReference>dynId30 </contentReference>
   </displayObject>
   <priority>      1     </priority>
  </object>
  <object>
   <displayObject>
     <contentReference>dynId28 </contentReference>
   </displayObject>
   <priority>      2     </priority>
  </object>
 </referencedObjects>
 <presentationContent>
[...]
       <movieTheater structId=dynId30>
         <entityKey> cinema_17a </entityKey>
         <name> Europa  </name>
           <geoCoordinate>
             <x> 225 </x> <y> 230 </y>
           </geoCoordinate>
[...]
</gestureAnalysis>
```

[1] The display coordinates are relative and thus independent of a certain type of display and resolution.

where the entry with the highest priority (1) is the one for the cinema "Europa".

This XML structure is passed on to the media fusion component, which merges it with the output from the speech analyzer, that is, as all other data in SmartKom, represented in M3L. After various inferences in the intention recognition module and augmentation through discourse and situative knowledge, the action planner contacts external services via the function modeling module and finally provides the presentation planner with a presentation task for the system's reaction.

It is then the task of the presentation planner to select the appropriate output modalities. The presentation planner activates the language generator and the speech synthesizer for speech output. Since the situated delegation-oriented dialog paradigm is based on a life-like character communicating with the user, we have to synchronize his actions with the other output devices to ensure a coherent and natural communication experience. For synchronization with the speech output, the synthesizer sends time-stamped word information back to the presentation planner, which uses it to synchronize e.g. the lips of Smartakus to the speech signal.

6. Conclusions

We presented the first demonstrator of the multimodal dialog system SmartKom. The massive data collection effort from multimodal WOZ dialogs was described. We sketched the multi-blackboard architecture and the XML-based mark-up of semantic structures as a basis for media fusion and media design. We introduced the situated delegation-oriented dialog paradigm (SDDP), in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display.

.

7. References

[1] Maybury, M. , Wahlster, W.(eds.): Readings in Intelligent User Interfaces. San Francisco: Morgan Kaufmann, 1998.

[2] Oviatt, Sharon & Cohen, Philip. "Multimodal Interfaces That Process What Comes Naturally" CACM, 43, 3, 2000, pp. 45-53.

[3] Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, Heidelberg, New York: Springer, 2000.

[4] Wahlster, W.: Pervasive Speech and Language Technology. In: Wilhelm, R. (Ed.): Informatics - 10 Years Back 10 Years Ahead. Berlin, Heidelberg, New York: Springer, (Lecture notes in computer science Vol. 2000), p. 274-293, 2001

[5] Wahlster, W.: SmartKom: Multimodal Dialogs with Mobile Web Users. In: Proceedings of the International Cyber Assist Symposium, Tokyo International Forum, p. 33-40, 2001.