

SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents

Wolfgang Wahlster Norbert Reithinger Anselm Blocher
DFKI GmbH, D-66123 Saarbrücken, Germany
{wahlster,reithinger,blocher}@dfki.de

Abstract

SmartKom is a multimodal dialogue system that combines speech, gesture, and facial expressions for input and output. SmartKom provides an anthropomorphic and affective user interface through its personification of an interface agent. Understanding of spontaneous speech is combined with video-based recognition of natural gestures and facial expressions. 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 dialogue paradigm, in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display. SmartKom is a multilingual system that analyses and generates German and English utterances. We describe the SmartKom architecture, the use of an XML-based mark-up language for multimodal content, and the most distinguishing features of the fully operational SmartKom 2.0 system.

1. Introduction

More effective, efficient, and natural interfaces to support the location-sensitive access to information, applications, and people are increasingly relevant in our information society which is plagued by information overload, increasing system complexity, and shrinking task time lines [cf. 2]. SmartKom (www.smartkom.org) is a multimodal dialogue system (see Fig. 1) that combines speech, gesture, and facial expressions for input and output [8]. SmartKom features the situated understanding of possibly imprecise, ambiguous, or incomplete multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations [2]. SmartKom's interaction management is based on representing, reasoning, and exploiting models of the user, domain, task, context and the media itself. SmartKom provides an anthropomorphic and affective user interface through its personification of an interface agent. 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 semantic and pragmatic levels.

SmartKom is the follow-up project to Verbmobil (1993-2000) and reuses some of Verbmobil's components for the understanding of spontaneous dialogues [7]. 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 current versions of the system.

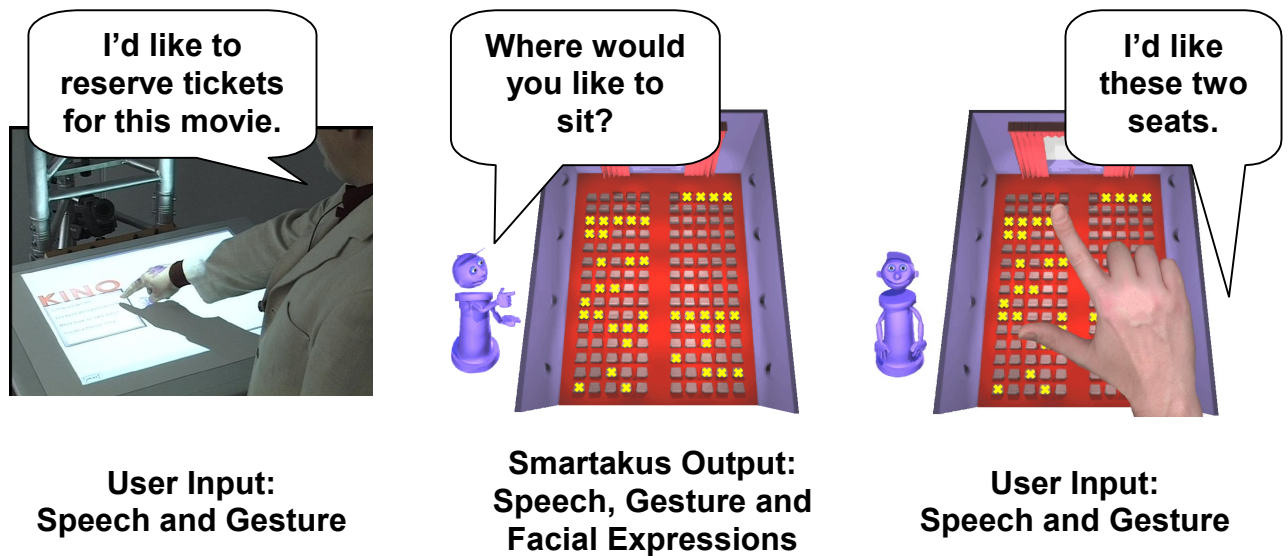


Figure 1: Multimodal Interaction with SmartKom

2. SmartKom's Multimodal Dialogue Paradigm

SmartKom aims to exploit one of the major characteristics of human-human interactions: the coordinated use of different code systems such as language, gesture, and facial expressions for interaction in complex environments (cf. [4]). SmartKom's multimodal interaction style eschews mouse and keyboard. SmartKom employs a mixed-initiative approach to allow intuitive access to knowledge-rich services.

SmartKom merges three user interface paradigms - spoken dialogue, graphical user interfaces, and gestural interaction - 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 SmartKom does not use a traditional WIMP (windows, icons, menus, pointer) interface; instead, it supports natural gestural interaction combined with facial expressions. Technically, gestural interaction is made possible by an extended version of SIVIT (Siemens Virtual Touchscreen), a realtime gesture recognition hardware and software system. The gesture module consists of a box containing an infrared camera and transmitter and is set to point at the projection area of a LCD video projector. The gestures can range from pointing with a finger to pushing a virtual button.

SmartKom's interaction style breaks radically with the traditional desktop metaphor. SmartKom is based on the situated delegation-oriented dialogue paradigm (SDDP): 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 dialogue 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 dialogues, in which the user carries out a task with the help of the system, with SDDP the user delegates a task to an agent and helps

the agent, where necessary, in the execution of the task (see Fig. 2). The interaction agent accesses various IT services on behalf of the user, collates the results, and presents them to the user.

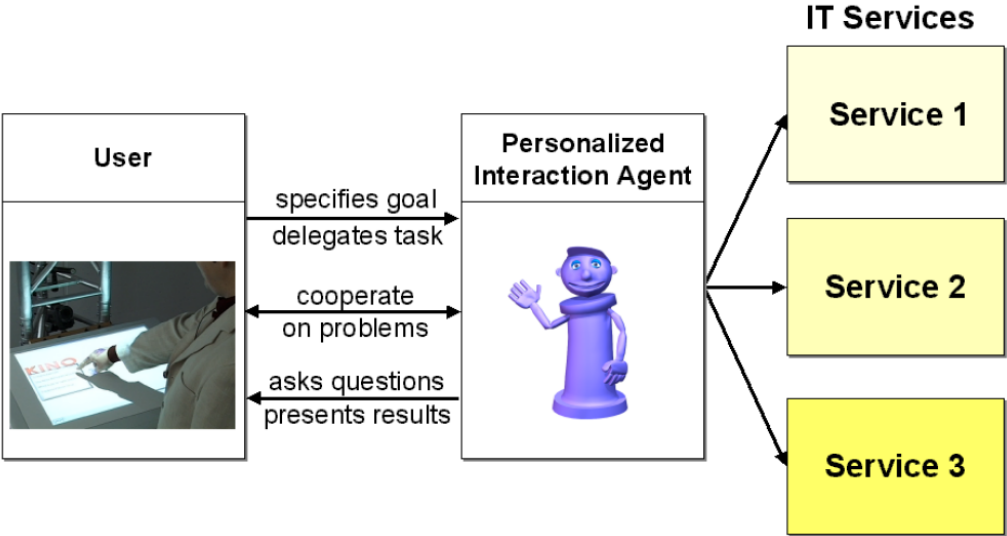


Figure 2: SmartKom’s Situated Delegation-oriented Dialogue Paradigm

The life-like character designed for the SmartKom system is called “Smartakus”. The “i”-shape of Smartakus reminds one of the “i” often used as a sign that directs people to information kiosks. The display of the 3D character Smartakus is adapted to the user’s viewing angle. Spotlight effects on the display are used to guide the user’s attention (see figure 3).

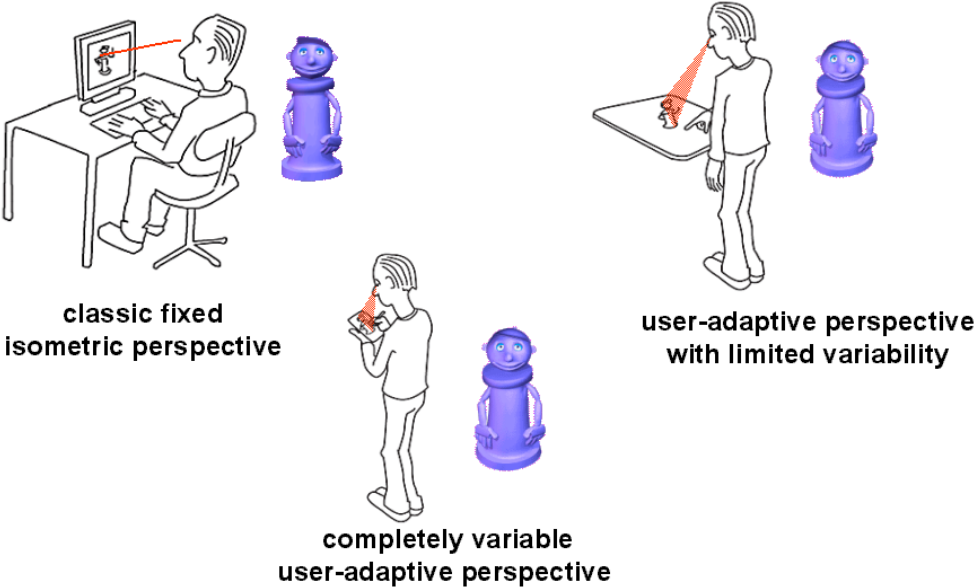


Fig. 3: User-adaptive Perspectives for the Life-like Character

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 dialogue systems based on SDDP (cf. [6]). Multi-channel audio and video data from Wizard-of-Oz (WOZ) experiments

are 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 facial expressions of emotion. A typical WOZ session lasts 4.5 minutes. The QuickTime file format is used for the integration of the multimodal and multi-channel data from the experiments. The annotated SmartKom corpora are distributed to all project partners via DVD-Rs and used as a basis for the functional and ergonomic design of the demonstrators (cf. [3]) as well as for the training of the various SmartKom components that are based on machine learning methods.

3. SmartKom as a Transportable Multimodal Dialogue Model

SmartKom's ultimate goal is a multimodal dialogue model that spans across a number of different platforms and application scenarios. One of the key design goals of SmartKom was the portability of the kernel functionality to a wide range of hardware platforms.

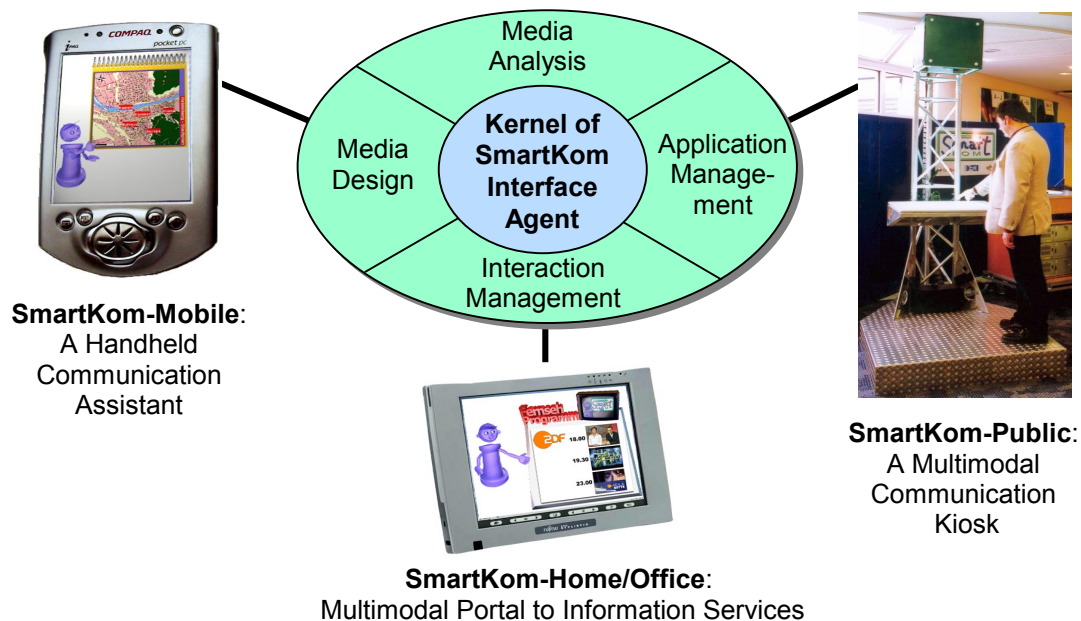


Fig. 4: Three Versions of SmartKom 2.0

Three versions of SmartKom are currently available (see figure 4):

- SmartKom-Public is a multimodal communication kiosk for airports, train stations, or other public places where people may seek information on facilities such as hotels, restaurants, and theaters. Users can also access their personalized standard applications via wideband channels. The user's speech input is captured with a directional microphone. The user's facial expressions of emotion are captured with a DV camera and his gestures are tracked with an infrared camera. A video projector is used for the

projection of SmartKom's graphical output onto a horizontal surface. Two speakers under the projection surface provide the speech output of the life-like character.

- SmartKom-Mobile uses a PDA as a front end. Currently, the iPAQ Pocket PC with a dual-slot PC card expansion pack is used as a hardware platform. It can be added to a car navigation system or carried by a pedestrian. SmartKom-Mobile provides personalized mobile services. Examples of value-added services include route planning and interactive navigation through a city via GPS and GSM, GPRS or UMTS connectivity.
- SmartKom-Home/Office realizes a multimodal portal to information services. It uses the Fujitsu Stylistic 3500X portable webpad as a hardware platform. SmartKom-Home/office provides electronic programme guides (EPG) for TV, controls consumer electronics devices like VCRs and DVD players, and accessed standard applications like phone and e-mail. The user operates SmartKom either in lean-forward mode, with coordinated speech and gestural interaction, or in lean-back mode, with voice input alone.

4. SmartKom's Architecture

Figure 5 shows the control GUI of the fully operational SmartKom 2.0 system. It reflects the modular software structure of SmartKom. The modules can be grouped as follows:

- input devices: audio input, gesture input, pen input, face camera input, and document camera input
- media analysis: speech recognition and analysis, prosodic analysis, face interpretation, gesture recognition and analysis, biometrics, and media fusion
- interaction management: context modeling, intention recognition, discourse modeling, lexicon management, dynamic help, interaction modeling, and action planning
- application management: the function modeling, interfaces to car navigation, external information services, consumer electronics, and standard applications like email
- media design: presentation planning, language generation, character animation, speech synthesis, display management, and audio output

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 on distributed computers, that run either Windows NT or Linux. Each module is implemented in C, C++, Java, or Prolog. The underlying integration software is based Verbmobil's testbed software framework [7].

SmartKom 2.0 is a multilingual system with speech recognition and speech synthesis modules for German and English. The GUI is used for the visualization of the data and control flow during processing multimodal input. Currently active modules are graphically highlighted, so that one can trace the processing steps.

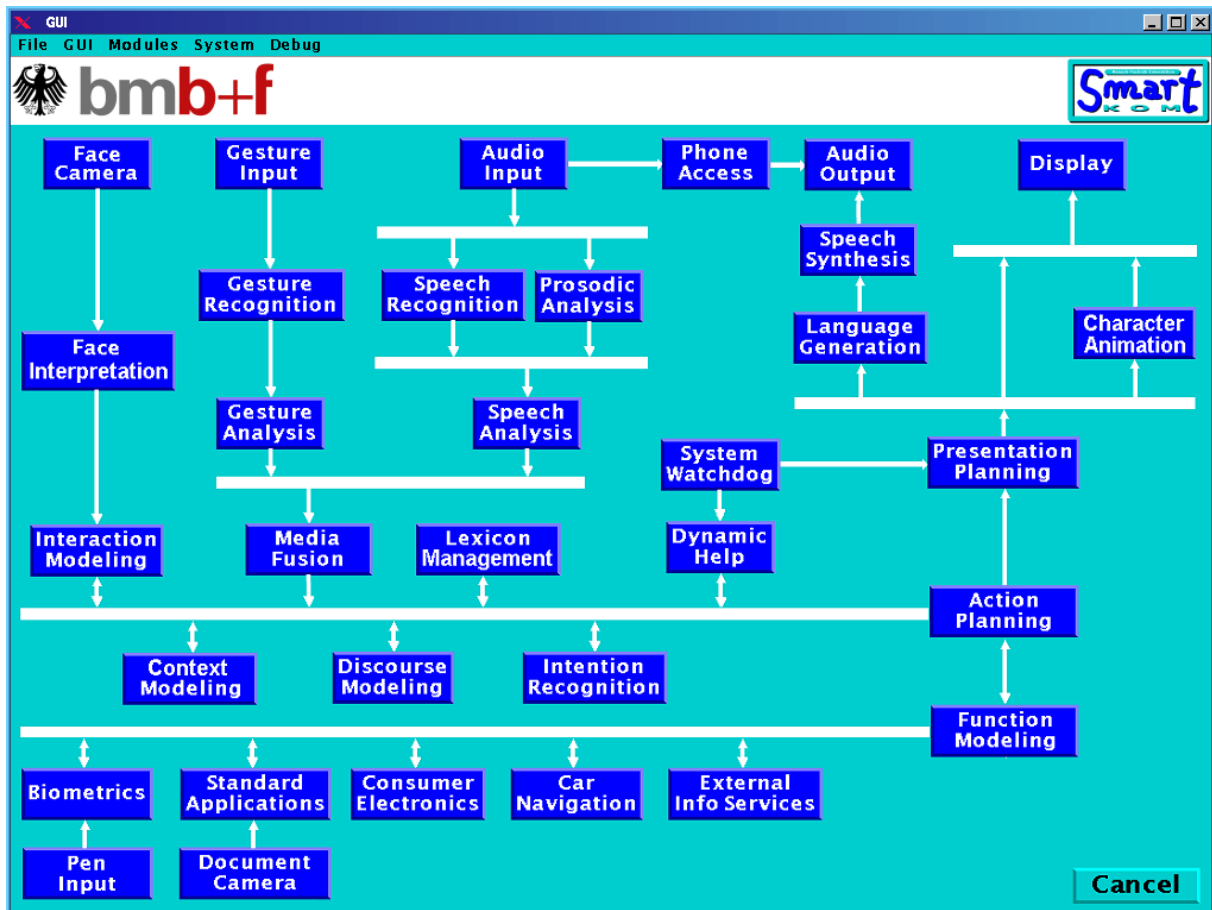


Figure 5: GUI for Tracing the Data and Control Flow in SmartKom 2.0

5. Multimodal Information Integration with M3L

A key design decision concerned the development of an XML-based markup language called M3L (MultiModal Markup Language) for the representation of all of the information that flows between the various processing components of SmartKom. For example, the word lattice and gesture lattice, the hypotheses about facial expressions, 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, of information about segmentation, and synchronization, and of information about the confidence in processing results. For each communication blackboard, XML schemas allow for automatic data and type checking during information exchange. The XML schemas can be viewed as typed feature structures. SmartKom uses unification and a new operation called overlay (cf. [1]) of typed feature structures encoded in M3L for media fusion and discourse processing.

In figure 6 the life-like character Smartakus presents a map of Heidelberg highlighting the location of cinemas. 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 surface¹. 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>

```

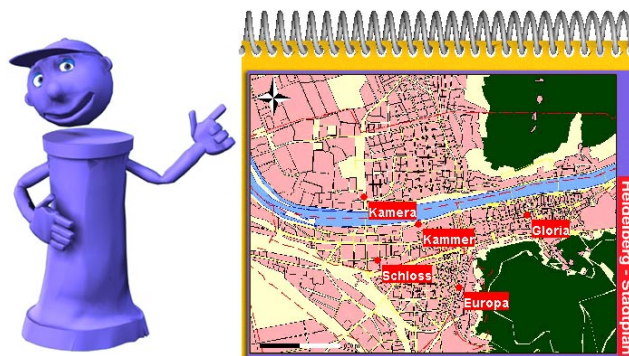


Figure 6: The Life-like Character's Pointing Gesture onto a Map Display Generated by SmartKom

Let's assume an utterance of the user like

User: I 'd like to reserve tickets for this [\uparrow] cinema.

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

[↑] denotes the user's pointing gesture to the cinema "Europa". Both recognizers for gesture and speech produce 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>
[...]
```

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

This XML structure is passed 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.

The face interpretation module (see Fig. 5) is trained to detect signs of annoyance. When a user showing a facial expression of annoyance says "That's a great suggestion", the interaction and intention recognition components can interpret this as an ironic comment and understand the user's dissatisfaction with a preceding proposal by Smartakus. The action planner can then try to find an alternative proposal and pass it on to the presentation planner.

The presentation planner selects the appropriate output modalities. It allocates information to the available media and activates the language generator, the character animation module and the graphics design component. The resulting graphics and animations are finally rendered by the display component. The language generator drives the concept-to-speech synthesizer. Since in the situated delegation-oriented dialogue paradigm a life-like character

interacts with the user, the output modalities have to be synchronized so as 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 the pointing gestures, facial expressions, and the lip movements of Smartakus with the speech signal.

6. Conclusions and Future Directions

The current version of the multimodal dialogue system SmartKom was presented. 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 dialogue paradigm (SDDP), in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display.

Important extensions of the current SmartKom version include work on increased robustness of the media-specific analyzers, the expansion of the domain of discourse to weather, restaurant and hotel information, improved dynamic media allocation for output generation, and metacommunicative subdialogues between the user and his Smartakus assistant.

Application domains ripe for SmartKom's multimodal interaction technology include in-home access to digital information, advanced web portals and information kiosks, mobile and wearable computing, in-car telematics, edutainment, and cultural tourism (see [5]).

7. References

- [1] Alexandersson, J., Becker, T.: Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System. In: Proceedings of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, August 2001, Seattle.
- [2] Maybury, M. , Wahlster, W.(eds.): Readings in Intelligent User Interfaces. San Francisco: Morgan Kaufmann, 1998.
- [3] Oppermann, D., Schiel, F., Steininger, S., Beringer, N.: Off-Talk - a Problem for Human-Machine Interaction? In: Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001, Vol. 3, p. 2197 – 2200.
- [4] Oviatt, S, Cohen, P.. Multimodal Interfaces That Process What Comes Naturally. In: CACM, 43, 3 2000, p. 45-53.
- [5] Stock, O.: Language-based Interfaces and Their Application for Cultural Tourism. In: AI Magazine, Vo. 22, No. 1, Spring 2001, p. 85 – 97.
- [6] Türk, U.: The Technical Processing in SmartKom Data Collection: a Case Study. In: Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, September 2001, Vol. 3, p. 1541 – 1544

- [7] Wahlster, W. (ed.): *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin, Heidelberg, New York, Springer, 2000.
- [8] Wahlster, W., Reithinger N., Blocher, A.: *SmartKom: Multimodal Communication with a Life-Like Character*. In: *Proceedings of Eurospeech 2001, 7th European Conference Speech Communication and Technology*, Aalborg, Denmark, September 2001, Vol. 3, p. 1547 – 1550