

Flexible Multimodal Human-Machine Interaction in Mobile Environments

Dirk Bühler, Wolfgang Minker¹ and Jochen Häußler, Sven Krüger²

Abstract. This article describes requirements and a prototype system for a flexible multimodal human-machine interaction in two substantially different mobile environments, namely pedestrian and car. The system allows an integrated trip planning using multimodal input and output. Motivated by the specific safety and privacy requirements in both environments, we present a framework for flexible modality control. A characteristic feature of our framework is the insight that both user and system may independently and asynchronously initiate a modality transition. We conclude with a brief discussion of further issues and research questions.

1 Introduction

SmartKom [5] is a long-term research effort funded by the German Federal Ministry for Education and Research (BMBF). The main contractor of the project consortium is the German Research Center for Artificial Intelligence (DFKI). The major industrial project partners are DaimlerChrysler, Philips, Siemens and Sony. As the follow-up project to Verbmobil (1993-2000), SmartKom reuses some Verbmobil's components for the automatic processing of spontaneous dialogues [3].

SmartKom aims at developing advanced concepts for intuitive human-centred computer interfaces. In this particular context, the "intuitive interface" should

- a) be self-explaining, thus requiring from a user only a basic understanding of the underlying technology and mechanisms
- b) use a mixed-initiative approach
- c) be able to recognise and interpret a broad range of both spoken natural language utterances and natural gestures in the context of the ongoing dialogue
- d) use a similar multimodal and personalised way of presentation in response
- e) be built around an integrated and interrelated set of tasks that the user and system can work on in a collaborative way

One of the major scientific goals of SmartKom is to explore and design new computational methods for the integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level [1, 2, 5]. This implies an exploitation of one of the major characteristics of human interactions: the co-ordinated use of different code systems, like language, gesture, and mimics, to interact in complex environments.

The most important modalities for human-machine interaction considered in the SmartKom project are

¹ DaimlerChrysler Research and Technology, Ulm, Germany, email: {dirk.buehler,wolfgang.minker}@daimlerchrysler.com

² European Media Lab, Heidelberg, Germany, email: {jochen.haeussler,sven.krueger}@eml.villa-bosch.de

- *Speech recognition and synthesis*
- *Gesture recognition*, such as pointing on the screen
- *Recognition of facial expressions* using a face tracker
- *Graphical display* of text, animations, and maps

Natural language interaction in SmartKom is based on speaker-independent speech understanding technology.

Figure 1 shows the control graphical user interface of the SmartKom system [5]. It reflects the modular software structure. The modules can be grouped into

- *Interface modules* with the audio module on the input and the display manager on the output side
- *Recognisers and synthesisers* with gesture, prosody and speech recognition modules on the input and speech synthesis and display management on the output side
- *Semantic processing modules* that create or transform meaning representations and include gesture and speech analysis, media fusion, intention recognition, discourse and domain modelling, action planning, presentation planning, and concept-to-speech generation
- *External services* that include EPG databases, a module for external info services such as map services and pedestrian navigation, and a car navigation module

SmartKom is based on a multi-blackboard architecture with parallel processing threads that support the media fusion and design process. All modules shown in Figure 1 are realised 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 on the Verbmobil testbed [3].

The concepts for intuitive interfaces are currently being tested and demonstrated as running systems in three major application scenarios:

- 1) *SmartKom Home/Office*: The system will serve as a personal device comparable to a traditional desktop computer, enhanced with multimodal interaction.
- 2) *SmartKom Public*: The system will be made available as an advanced multi-media communication center in central public places such as airports, train stations, or administrative buildings.
- 3) *SmartKom Mobile*: Throughout the past few years, natural interfaces to support the location-sensitive access to information, services, and people have become increasingly relevant in mobile environments [4]. SmartKom Mobile, a small yet powerful portable device (cf. Figure 4) will act as a permanent digital companion that provides the system's functionality and services to the user in mobile environments, namely pedestrian and car.

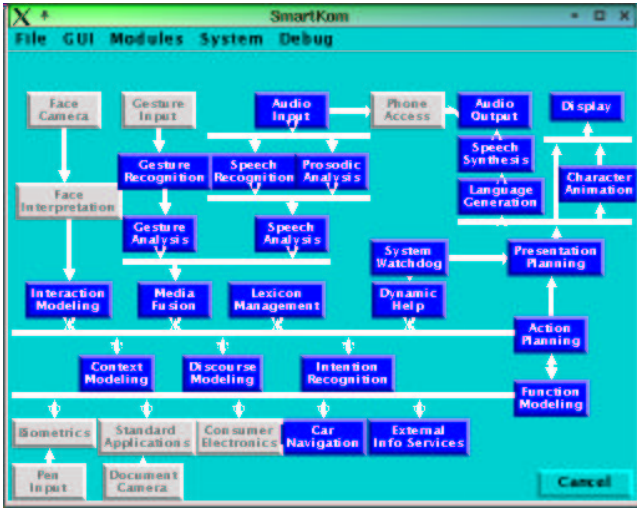


Figure 1. Graphical user interface for tracing the flow of information in SmartKom. The interface reflects the modular software architecture

The three SmartKom scenarios share a core functionality. A set of “standard applications” has therefore been defined. They include communication (via email and telephone) and personal assistance (address-book, agenda). In addition, for each application scenario there exists a number of exclusive applications (e.g., incremental and location-aware navigation in SmartKom Mobile).

In this article, we will focus on the SmartKom Mobile scenario and present a first prototype system. Section 2 describes this application scenario and its differences with respect to the other SmartKom scenarios. Section 3 discusses a framework for flexible modality control motivated by the identification of certain characteristic interaction modes in SmartKom Mobile. Section 4 presents the implementation platform and a first version of the SmartKom Mobile prototype. Finally, section 5 summarises the presented concepts and designs and provides an outlook to further development and research issues.

2 The SmartKom Mobile Application Scenario

The SmartKom Mobile application scenario applies to *two* related yet significantly different user environments:

- 1) *Pedestrian environment*: The user (walks around and) carries the SmartKom Mobile device as a permanent and location-aware digital companion.
- 2) *Driver environment*: By installing the SmartKom Mobile device into the docking station inside the car the system can be used as a car-based digital assistant that takes advantage of a in-car hands free audio system and car-specific functionality, such as real-time information about the car state.

As outlined above the general goal of the SmartKom project is to make available to the user some standard applications and common way of interaction, while flexibly enriching the system with functionality specific to a particular scenario.

For illustration, we outline the navigation functionality as the salient application domain in SmartKom Mobile. Implemented as modules for car navigation and external info services (cf. Figure 1) this functionality includes:

- *Trip planning*: A trip may consist of multiple route segments to be covered either by walking or driving. This requires specifying

start and destination locations, as well as optional stops. The user may also wish to specify different route types (e.g., shortest distance) and other properties of the trip. He may be interested in some information on the trip in advance, e.g. distance, estimated duration, details of points of interest nearby.

- *Trip execution*: This functionality includes *incremental* guidance and monitoring. It involves timed presentation of driving directions as well as processing incoming positioning information. In the pedestrian environment, the functionality also involves proactiveness, e.g. by presenting additional navigation-unrelated information available from digital maps.

While the first part of the functionality may also be useful in other scenarios (e.g., for planning and making reservations at home using SmartKom Home/Office), trip execution is only available within the mobile scenario.

The SmartKom system aims at providing a common and uniform dialogue-based interaction model in the three application scenarios. Therefore, one of the major challenges is to accommodate the various peculiar restrictions that each of these application scenarios pose on the system. This concerns in particular SmartKom Mobile since the user interacts with the system in two substantially different environments. In many situations, the user may also be concerned with tasks other than communicating with the system (i.e., driving) and he is not immobilised in front of the system (i.e., the user frequently moves his head and body). Finally, due to traffic safety considerations, pointing gestures are prohibitive if the system is used whilst driving.

Consequently, for a given application the different dialogue flows need to be adapted across the different application scenarios and also even within SmartKom Mobile itself.

3 A Framework for Flexible Modality Control

SmartKom Mobile is different from the Home/Office and Public scenarios in that the control of the interaction modality is a major concern.

We have therefore identified five major combinations of modalities – from now on referred to as *interaction modes* – that seem characteristic to the SmartKom Mobile scenario (cf. Figure 2):

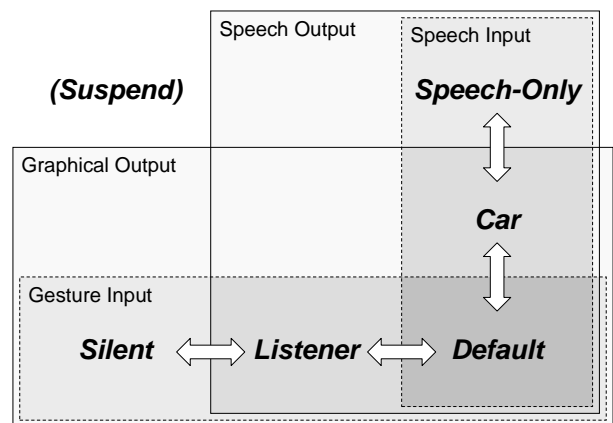


Figure 2. Interaction modes as combinations of communication modalities. The inter-modal transitions are described in the text.

- *Default*: Mainly used in the pedestrian environment and when privacy or disturbing others is not an issue, all modalities are enabled

in this mode. Stemming from safety considerations (discussed below), this mode is available in the driver environment only when the car is not moving.

- *Listener*: In this mode the user accepts spoken system output but he does not use speech as an input modality (i.e., the user is the listener here). Spoken output is useful for quickly providing concise summaries as well as extended background information to the user whose focus of attention is placed on tasks others than interacting with the system. The user may employ ear phones in order to protect his privacy. The listener mode may also prove useful for avoiding confusion of the system by off-talk.

In the listener mode, the system does not try to excite spoken input from the user. Thus, it should generate prompts like “Show me ...” rather than “Tell me ...”.

- *Silent*: This mode is useful when spoken language human-machine dialogue is problematic, for instance in certain public places or in public transportation. The interface should be similar to traditional graphical user interfaces.
- *Car*: This mode is a restriction version of the default mode for use mainly in the driver environment. In this mode, speech is the dominant communication modality, while graphical displays are used only for presenting additional (“non-vital”) information, such as maps for presenting a route while driving or the locations of parking options. Consequently, the display needs not to be touch sensitive.
- *Speech-Only*: This mode is mainly used in the driver environment notably when the car is moving. Any additional safety risk (such as driver distraction from the driving task by gesture input or graphical output) must be avoided. System interaction is restricted to the audio channel, i.e. the use of a speech-only dialogue is mandatory.
- (*Suspend*): In the driver environment an emergency requires the dialogue with the system to be temporarily suspended. This situation occurs when the driving situation becomes dangerous and the entire user attention is required. Although the system suspends any interactivity, it should be prepared to resume the dialogue, especially after very short interruptions. Temporarily suspending interaction may also be useful in the pedestrian environment when the user is engaged in tasks others than communicating with the system, such as crossing a street or talking to another person.

Having identified these five interaction modes we need to determine, under which circumstances transitions between these modes are possible. Informally speaking, a transition is realised by “switching” on or off one or more modalities.

We distinguish between *user-* and *system-driven* transitions. A user-driven transition is the result of an explicit command to use or not to use a specific modality. In turn, system-driven transitions are triggered by the system without any user interaction.

The user should be able to switch between modalities anytime, although some modalities may temporarily be disabled for safety reasons. A user-driven modality transition may be performed by the following actions (cf. Table 1):

- The user may suspend/resume the operation with the system. In the driver environment this could be realised by a Push-to-Activate button or lever.
- In the pedestrian environment the user may switch off speech input and output, for instance by using one of the buttons of the portable device.
- In the pedestrian environment, the user may toggle display opera-

tion. A request for turning on the display could be recognised by the touch sensitive screen of the portable device.

- Re-opening the speech channel by uttering spoken commands re-enables spoken language dialogue, e.g. when the system interaction is suspended or when graphical output is exclusively used. Requesting graphical output by spoken commands (like “Show me the map”) enables graphical output, unless this modality is disabled for safety reasons (cf. speech-only mode).

The system may initiate a modality transition in one of the following ways (cf. Table 1):

- As outlined above a specialised software running on the car PC may detect an emergency in the driver environment. To this end, the software has access to the necessary pieces of information about the state of the car (e.g. driving speed, state of brakes, etc.) through the CAN bus.
- Analogously, starting the car is detected by the car PC and is interpreted as a transition into speech-only mode. Likewise, stopping the car leads to the default mode.
- In the pedestrian environment, i.e. when graphical output is available, repeatedly failing to understand spoken input from the user should lead the system to infer that the current situation is not suitable for speech recognition (e.g., though strong background noise). Thus, the system switches to listener mode in order not to rely on spoken input.

Table 1. Summary of modality transitions in SmartKom Mobile.

Signal	Effect
<i>User-driven transitions</i>	
Push-to-activate	Suspend/Resume operation
Command/button	Turn on/off speech
Command/button	Turn on/off graphics
<i>System-driven transitions</i>	
Install in car	Turn off gesture input
Take out from car	Switch to default mode
Start car	Turn off graphical displays
Stop car	Turn on graphical displays
Detect emergency	Suspend operation
Misrecognition/noise	Turn off speech input

A major concern will be the adaption of the presentation modules in SmartKom in order to reflect the modality restrictions. At the same time the general way of interaction (SmartKom Look&Feel) needs to be coherent and recognisable. Table 2 illustrates how a system presentation could be adapted to the available modalities. In addition, modality transitions may need to “grounded” with the user in order to avoid user confusion.

Table 2. A sample output presentation depending on the interaction mode.

Mode	Speech Output	Display
Default:	“Please select a parking place”	Map
Listener:	“Please select a parking place from the map”	Map
Silent:	—	Map + text (*)
Car:	“There are 5 parking places. Number 1 is ...”	Map
Speech-Only:	“There are 5 parking places. Number 1 is ...”	—

(*) The text could be displayed as illustrated in Figure 3.

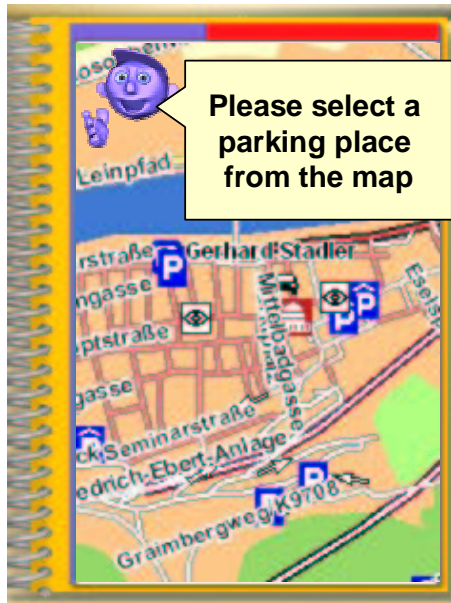


Figure 3. A potential screen design for presenting text instead of synthesised speech in the silent mode.

4 The SmartKom Mobile Prototype

A prototype version of the SmartKom mobile system has been designed and implemented. The Compaq iPAQ has been chosen as the primary interface because it allows innovative modes of interaction and flexibility. For the driver environment, a hands-free audio system and a colour display is available in the car.

Our primary concern relies on connecting different mobile environments rather than specific technological issues (such as, for instance, speech recognition constrained by the limited resources and computing power of current PDAs). Therefore, all the data processing is performed on the PCs whereas the portable device and the car interior hardware are used as mere periphery clients responsible for recording and displaying the input and output data, respectively. The communication of the data is done via a high-speed WaveLAN connection, as shown in Figure 4. In the car, the iPAQ will be installed in a docking station. There, it will take advantage of the car interior infrastructure, such as built-in speakers, a microphone array, and a GPS device.

With the current prototype, the SmartKom Mobile system for the pedestrian environment, the user is able to perform integrated trip planning (pedestrian and car) using a simple multimodal interaction (i.e. speech and pointing gesture). The system displays graphical output in the form of maps (route information) or slide shows (sight information). Using synthesized speech an animated agent provides additional information. Figure 5 shows an example interaction and a possible systems output. For now limited to the default mode, we are currently adapting the multimodal dialogue strategy to the framework described in section 3.

5 Summary and Outlook

With the SmartKom Mobile prototype system we have developed and described a new architecture and a conceptual framework for flexible multimodal human-machine interaction in mobile environments. For research purposes we have combined the flexibility of a digital handheld device with the computing power of standard personal

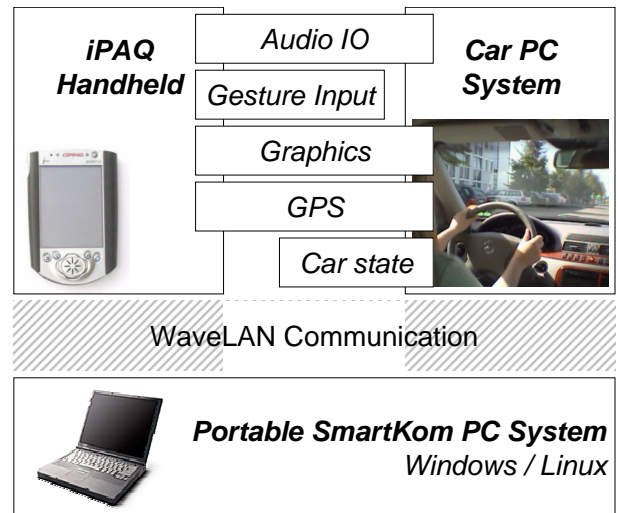


Figure 4. Hardware and software components used in the SmartKom Mobile scenario.

computers. The system is also novel in that it allows integrated trip planning (pedestrian and car) using multimodal interaction.

Stemming mainly from different safety and privacy requirements in mobile environments, we have discussed a framework for flexible modality control. It consists of a state model for modality combination, describing possible interaction modes in different situations. We conclude that both the user and the system should be able to initiate transitions across interaction modes.

A number of interesting issues and challenges remains to be discussed. One question is how the graphical displays and the system in general can convey a uniform “Look & Feel” whether or not spoken language is available. Another question is the relation between interaction modes and dialogue initiative.

The system should be adaptive (or adaptable) to user preferences and to his experience with the system. Novice users often require more system guidance (e.g. in the form of a tutorial mode) than experienced ones who may wish to avoid unnecessary verbosity by graphical or spoken short cuts. In addition, memorisation and statistical models could be used to automatically update the user profile or infer preferences.

We are currently adapting the multimodal dialogue strategy to the framework of flexible modality control.

One of the next steps also consists in the integration of SmartKom Mobile in the car and to enable incremental route planning for pedestrians. The system will also undergo performance tests with real users.

ACKNOWLEDGEMENTS

This paper is based on research carried out within the German Federal Ministry for Education and Research (BMBF) Project SmartKom. The work at the European Media Lab (EML) is partially funded by the Klaus Tschira foundation (KTF).

SMA: *Displays initial display*
USR: I would like to travel to Heidelberg
SMA: *Displays initial display*
 Where would you like to start?
USR: Saarbrücken
SMA: *Displays initial display*
 Would you like the fastest or shortest way?
USR: The fastest
SMA: *Displays initial display*
 The car route is being calculated
SMA: *Displays car route*
 Here you can see a map with the fastest way from Saarbrücken to Heidelberg
SMA: Do you wish to continue the trip planning?
USR: Yes
SMA: *Displays map with points of interest*
 Here you can see a map of Heidelberg
USR: <Points to Peterskirche>
 I would like to know more about this
SMA: *Slide show about Peterskirche*
 Here you have some information about the Peterskirche
USR: How can I get to the Peterskirche?
SMA: *Displays pedestrian route*
 The route is displayed in the map
 (a)



(b)

Figure 5. Example interaction using the SmartKom mobile prototype system. It allows an integrated planning of car and pedestrian routes in a single dialogue. (a) Multimodal interaction between the user and SmartKom in the default mode; (b) handheld device showing a pedestrian route.

REFERENCES

- [1] M. Maybury and W. Wahlster, *Readings in Intelligent User Interfaces*, Morgan Kaufmann, 1998.
- [2] S. Oviatt and P. Cohen, 'Multimodal interfaces that process what comes naturally', in *CACM*, volume 43, pp. 45–53, (2000).
- [3] W. Wahlster, *VERBMOBIL: Foundations of Speech-to-Speech Translation*, Springer-Verlag, 2000.
- [4] W. Wahlster, *Pervasive Speech and Language Technology*, Springer-Verlag, 2001.
- [5] W. Wahlster, N. Reithinger, and A. Blocher, 'SmartKom: Multimodal communication with a life-like character', in *EUROSPEECH*, (2001).