SmartKom–Home — and advanced multi-modal interface to home entertainment

Thomas Portele[1], *Silke Goronzy*[2], *Martin Emele*[2],
Andreas Kellner[1], *Sunna Torge*[2], *Jürgen te Vrugt*[1]

[1]Philips Research Aachen, Aachen, Germany
[2]Sony International (Europe) GmbH, Stuttgart, Germany
{thomas.portele,andreas.kellner,juergen.te.vrugt}@philips.com
{goronzy,emele,torge}@sony.de

Abstract

This paper describes the SmartKom-Home system realized within the SmartKom project. It assists the user by means of a multi-modal dialogue system in the home environment. This involves the control of various devices and the access to services. SmartKom-Home is supposed to serve as a uniform interface to all these devices and services so the user is freed from the necessity to understand which of the devices to consult how and when to fulfill complex wishes. We describe the setting of this scenario together with the hardware used. We furthermore discuss the specific requirements that evolve in a home environment, and how they are handled in the project.

1. Introduction

SmartKom is a joint research project funded by the German federal ministry of Education and Research. Lead by the Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI), 10 partners collaborate in developing a multi-modal interface to several applications in three application domains with distinct usage scenarios [1]. For each scenario one dedicated demonstrator is developed. However, one key concept of SmartKom is using the same core engine (the "backbone") in all three scenarios. Whenever possible, this concept is extended to the applications as well.

The scenarios are called SmartKom-Public, SmartKom-Mobile, and SmartKom-Home. SmartKom-Public [2] contains an advanced information and communication booth with phone, FAX, tourist and movie information, and biometric access control. Gesture and mimic recognition is included. SmartKom-Mobile [3] supports hand-held devices for route planning and travel information. SmartKom-Home uses a tablet PC with a touch screen, controls entertainment devices, and serves as an electronic program guide (EPG). The SmartKom-Home system is jointly defined and monitored by Philips Research in Aachen and Sony International (Europe) GmbH in Stuttgart.

We start with a description of the features of the SmartKom-Home system in Section 2. The hardware used is described in more detail in Section 2.1. The applications are discussed in Section 2.2. Hardware and applications pose some particular constraints on the dialogue system. On one hand devices the user usually uses at home such as TV, video cassette recorder (VCR), MP3 juke-box, etc. need to be controlled by SmartKom. As a consequence, the many functionalities of these devices and services need to be modeled within the system to be able to control them and offer their full functionality to the user. In our system this is handled by the function modeling component that is described in Section 2.3. Furthermore, services such as the EPG are characterized by their highly dynamic content, i.e. the daily changing TV program and information for current shows. In order to handle dynamic content appropriately, various modules in the system need to be aware of it. Therefore, a dynamic lexicon serves as a central knowledge source. It is described in Section 2.4. Also, the interaction modes are different from the other scenarios. In the home scenario we explicitly distinguish between learn-forward and lean-backward mode (Section 2.5). Feedback strategies as an important part of the interaction are also discussed. In Section 3 we mention the evaluations that were conducted to evaluate the SmartKom-Home demonstrator. Finally, Section 4 summarizes this paper.

2. Features

The environment of a SmartKom-Home system is, not surprising, the living room of the user. The global application is accessing entertainment functionality. Key features are multi-modal interaction using speech and a tablet PC, switching between a multi-modal lean–forward and a speech–only lean–back mode, and an animated interaction agent called "Smartakus". One important point is emphasizing the functional aspect. The user does not need to know device-specific features or service idiosyncrasies.

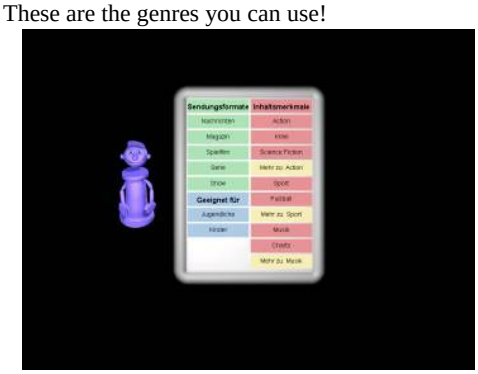
Figure 1 shows a typical interaction (translated from German). After initial greeting a TV show is selected using dynamic help. The video recorder is programmed to record that show, and the TV is switched on to a currently running program. The system is set to lean–back mode while the user watches TV.

2.1. Hardware

The application hardware is a high-end digital Sony VCR DHR-1000VC, and a standard Philips TV set. This represents a sample set of devices. However, the layout of the modules that control these devices is designed to support other devices as well, such as HiFi components etc. For choosing the interaction device we had to consider that it should reflect the necessities of the home environment. A platform that is mounted at a fixed place in the room will not be accepted by people used to remote–controls, but the constraints of a mobile scenario on device size do not apply either. A portable easy-to-use device with a sufficiently large interaction display to allow showing several information items (EPG info and device status) and the SmartKom persona "Smartakus" simultaneously was judged to

User Hello!
System Welcome to SmartKom. My name is Smartakus. How may I help you?
User What's on TV this noon?
System This is the TV program at noon!

User What else is on this channel (points to a program from provider ARD)?
System This is the TV program on ARD in the afternoon (shows only program items from ARD)!
User Which genres do you know?
System These are the genres you can use!



User Is there sports on any channel?
System These are the sports programs this afternoon (shows all sports program items)!
User Record that (points to a tennis show)!
System The show "US Open Semifinals" today at eight p.m. on RTL will be recorded!
User Switch the TV to ARD!
System (switches the TV on and selects channel ARD). The TV is now showing the station ARD!

User Turn off the display!
System (turns display off) The display is now deactivated!

Figure 1: Typical interaction with SmartKom-Home (translated from German). After initial greeting a TV show is selected using dynamic help. The video recorder is programmed to record that show, and the TV is switched on to a currently running program. The system is set to lean–back mode while the user watches TV.

Figure 2: Tablet PC as interaction device of SmartKom-Home.

be the optimal interaction device. At the time of the definition project phase early 2000 the Fujitsu Stylistic 3500 Tablet PC (a full-fledged PC with a touch screen) (see Figure 2) was the best choice — nowadays several similar web pads are available. The device is sufficiently powerful to run the SmartKom modules for audio I/O and display generation locally.

2.2. Applications

The main applications are the control of consumer electronics (CE) devices and access to an EPG. Access to an MP3 jukebox is not yet fully integrated. Furthermore, the application 'personal calendar' is used to store and retrieve shows to be recorded. For each application a description for the SmartKom action planning module was created. The action planning module is responsible for coordinating/triggering the different modules when necessary. Links between these descriptions allow switching from one application description (also termed *action plan*) to another, e.g. if the user indicates in the EPG application that she wants to record a certain program with the VCR. This process is transparent for the user, but facilitates development and integration of new applications.

An important difference between device control and EPG is the way of interaction. EPG is performed in a *browsing mode*, where information is always displayed, and missing information is substituted by meaningful default values (e.g. 'today,now' for time, 'all' for channels and genres). These values are changed and refined during the browsing process [6]. On the other hand, device control needs to have all values determined before any action can be carried out. This is in line with traditional mixed-initiative slot-filling dialogues [8]. The SmartKom backbone supports both styles of interaction.

Each application has a visual representation (see screen shots at Figure 1). The EPG application usually takes the full screen (top and bottom image) while TV and VCR information are displayed in the bottom (bottom image). When delivering user assistance in a meta-dialogue the screen changes again (middle image).

The EPG application supports preferences for genres and channels. These preferences, however, are not used as a filter

to eliminate hypothetically unwanted information, but help resorting the results - preferred items are displayed (or spoken in lean-back mode) first. The preference values are extracted by the dynamic help component [7] according to emotive user reactions. For instance, if information about a thriller is displayed and the user likes it (she says something like "fine" or smiles), the preference value for the genre 'thriller' is increased.

2.3. The Function Model

Since we consider SmartKom-Home as a uniform interface to various services and devices, all of these are modeled by the function model. This was motivated by the fact that users will have complex wishes involving several devices and/or services. However, the user should not have to care about which device needs to do what at a certain point in time. Therefore the function model formally describes the functionalities of all services and devices in the system. The action planner triggers the requests to the function model. A planning component within the model generates a plan that includes all devices/services needed to fulfill the user wish. Furthermore this plan includes the necessary steps to be taken for each of the applications. A user wish could e.g. be 'Record the thriller with Julia Roberts tonight'. This involves querying the EPG services for the information needed (in this case channel and starting time) to then program the VCR accordingly. The function model is described in more detail in [4, 5].

2.4. The Dynamic Lexicon

Most of the applications in the home scenario are characterized by their highly dynamic content. This holds in particular for the EPG, since the TV program changes daily or even more frequently. Many modules are affected by this dynamic content. The dynamic lexicon handles this dynamic content and serves as a knowledge source for those modules that need lexical information. Concretely, these are the speech recognizer, the prosody recognizer, the speech analysis and the speech synthesis for obvious reasons: Of course the user should be able to speak any name of a show, actor etc., genre, time channel and so on when browsing the EPG. These new words must be recognizable and interpretable, and also the speech synthesis on the output side uses the new words when presenting the EPG search results. It is the task of the lexicon to keep track of words that need to be added/removed to/from the lexicon when the user request requires this. Words can also be removed from the lexicon when not needed any longer. This is important for keeping the speech recognizer's vocabulary as small as possible to maintain high speech recognition rates.

Also, the pronunciations for new words are necessary for the modules named above. These are automatically generated by a decision tree-based grapheme-to-phoneme conversion. Pronunciations are mainly generated for German. However, some lexical entries, such as movie titles or actor names, are often not German but rather English. As a consequence, English pronunciations are generated simultaneously for those entries.

2.5. Interaction

The interaction with SmartKom-Home is governed by the principle that the user should have control over the system. The system is designed rather like a butler than like an autonomous agent. The user must be able to understand *what* the system is doing, and *why* it is doing it, and it is one important task of the system to support this.

The home scenario with the tablet PC demands some specific interaction features. Most importantly, users should be able to switch between two modes:

- The *lean-forward* mode supports using the display for touch input and visual output. This is the 'normal' mode for focused interaction with the system.
- The *lean-back* mode uses only the acoustic channel (i.e. speech) for input and output. This mode is assumed to be used when the user is either not willing to leave the sofa and reach for the device, or when only short commands like channel switching are necessary. An example corresponding to the second system turn of Figure 1 would be read as: *"Some of the broadcasted programs are the news show 'Tagesschau' on ARD at twelve thirty, the sitcom 'Alles unter einem Dach' at twelve thirty, and the news show 'Tagesschau' on ZDF, also at twelve thirty.* Only a limited number of items (currently three) is read.

Switching between the modes must be initiated by the user. The switch to lean-back mode is given via one of several speech commands (e.g. *"Go to sleep!"*). The lean-forward mode is turned on either by a speech command (e.g. *"Wake up!"*) or by touching the display. The system can suggest a mode switch to lean-forward mode if a list with more than three items has to be conveyed to the user.

Feedback is an important issue in a complex dialogue system [9]. Three levels of feedback are distinguished in SmartKom-Home, and each level has its own modality in the lean-forward mode:

System State: The state of the system is conveyed by the Smartakus persona. When a user command can be accepted, it makes a listening gesture. Processing stages (recognition, analysis, processing, presentation) are symbolized by different persona gestures. Thus, the user knows when the system is working (and even how long it may take), and when it is ready to accept commands.

System Belief: The results of the analysis are presented by speech, e.g. EPG constraints ("Here you see this evening's sports shows!"), or device states ("The TV is now switched to ARD!"). Thus, the user can detect misinterpretations. The visual presentation also contains information about the system belief, e.g. the current constraints in the EPG application.

System Answer: The results of the user query (EPG information, device states) are displayed on the screen.

In the lean-back mode, only spoken feedback is possible. The system state is not conveyed to the user, but the other two items are (at least partially for result lists) spoken.

Another issue for complex systems is user assistance. Three cases are considered in SmartKom-Home:

Help: The user can ask for help. Supported questions contain formulations like "What can I say?" for general help, "Which channels/genres/... do you know?" for the possible EPG constraints, and "Which commands do you know?" for the device control. The possible alternatives are then displayed (see central screen shot in Figure 1).

Problem detection: The system can detect problems during the interaction. These problems can be caused by failure of recognizing or interpreting user input. The system

then gives assistance by describing current input possibilities (similar to the response to "What can I say?"), or by asking for missing information,

Constraining and Relaxing: The EPG database access can lead to no or too many results. The system then displays the available information and gives hints to constrain or relax the query - it will not do that automatically, because the user should remain in control. The same holds for failures during VCR operation (e.g. missing recording media).

Modality-switching, feedback, and user assistance, are designed according to the butler paradigm: the user has control, the system gives assistance.

3. Integration and Evaluation

The complete system contains 28 modules communicating via a multi-blackboard architecture [10]. The interface format is an XML dialect which allows using online validation and several XML tools. The modules were developed by 10 partners and span the whole range from device handlers for audio and display, recognizers for gestures and speech, analyzers, dialog, function, lexicon and presentation managers, service and hardware interface modules, to generation and synthesis modules. The complexity of the system is, therefore, substantial, and intensive testing and checking against the specifications was performed by the integration group at the DFKI and the scenario managers (Philips and Sony). To avoid a mixture of incompatible module versions fixed integration dates were defined by the system group, and several versions of SmartKom-Home exist, the current one being version 4.1.

A formal evaluation of an earlier version (SmartKom-Home 2.1) had been carried out by project partners from the university of Munich using their PROMISE framework [11]. The evaluation established several problematic aspects such as unacceptable system response times that could be resolved by module optimization and new hardware. Furthermore the evaluation resulted in an improved feedback mechanism and optimized module behavior. The current prototype will also be subject to a similar evaluation in the near future, and an improved performance can be expected.

4. Conclusion

This paper presented the SmartKom-Home system as a multimodal assistant for home environments. SmartKom-Home serves as a uniform interface to all service and devices by combining applications for EPG and device control in a transparent way. The interaction style adheres to the principle: the user has control, the system gives assistance. We described particular requirements for this scenario and how the involved modules take this into account. This included a description of the hardware, the function model that jointly controls all services and devices, and the dynamic lexicon which serves as central knowledge source for all modules that need to deal with dynamic content such as the entries in the EPG data. Also the interaction modes – lean-forward and lean-backward mode – that are specific for the home scenario were discussed together with strategies for feedback and user assistance. The system was iteratively improved by conducting user evaluations which confirmed the importance of feedback and user assistance.

5. Acknowledgments

This research was conducted within the SMARTKOM project and partly funded by the German Federal Ministry of Education and Research (BMBF) under grants 01L905G7 and 01IL905I7. Without all our partners from the SmartKom project even the definition phase would not have been possible. A most special "Thank You" to the system integrations group at DFKI Kaiserslautern.

6. References

[1] W. Wahlster, N. Reithinger, and A. Blocher, "Smartkom: Multimodal communication with a life-like character," in *Proceedings of the Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1547–1550.

[2] S. Koch, "Small talk with the pc," *New World Siemens Magazine*, vol. 4, 2001.

[3] D. Bühler, W. Minker, J. Häussler, and S. Krüger, "The smartkom mobile multi-modal dialogue system," in *Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, 2002.

[4] S. Torge, S. Rapp, and R. Kompe, "The planning component of an intelligent human machine interface in changing environments," in *Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, 2002.

[5] S. Torge, S. Rapp, and R. Kompe, "Serving Complex User wishes with an enhanced spoken dialogue system" in *Proceedings of the ICSLP2002*, Denver, USA, 2002.

[6] A. Kellner and T. Portele, "Spice - a multimodal conversational user interface to an electronic program guide," in *Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, 2002.

[7] M. Streit, "Dynamic help in the smartkom system (working title)," 2003, submitted to ISCA workshop on Error Handling in Spoken Dialogue systems.

[8] H. Aust, M. Oerder, F. Seide, and V. Steinbiss, "The Philips automatic train timetable information system," *Speech Communication*, vol. 17, pp. 249–262, 1995.

[9] S. E. Brennan and E. A. Hulteen, "Interaction and feedback in a spoken language system: A theoretical framework," *Knowledge-based Systems*, vol. 8, pp. 143–151, 1995.

[10] A. Klüter, A. Ndiaye, and H. Kirchmann, "Verbmobil from a software engineering point of view: System design and software integration," in *Verbmobil: Foundations of speech–to–speech translation*. Springer, 2000, pp. 659–670.

[11] N. Beringer, K. Louka, V. Penide-Lopez, and U. Türk, " PROMISE - A Procedure for Multimodal Interactive System Evaluation," in *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation'*, Las Palmas, Gran Canaria, 2002.