In: Proc. of the 1st International Workshop on Man-Machine Symbiotic Systems, pp. 213--225, Kyoto, Japan, 2002.

SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions

Wolfgang Wahlster

German Research Center for Artificial Intelligence (DFKI)
D-66123 Saarbrücken, Germany
e-mail: wahlster@dfki.de
www.dfki.de/∼wahlster

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. Various types of unification, overlay, constraint solving, and planning are the fundamental computational processes involved in Smartkom's modality fusion and fission components. The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results based on a three-tiered representation of multimodal discourse. We show that a multimodal dialogue system must not only understand and represent the user's multimodal input, but also its own multimodal output.

1 Introduction

More effective, efficient, and intuitive interfaces to support the location-sensitive access to IT services are increasingly relevant in our information society which is plagued by information overload, increasing system complexity, and shrinking task time lines (cf. [3]). SmartKom (www.smartkom.org) is a multimodal dialogue system (see Fig. 1) that combines speech, gesture, and facial expressions for input and output [12]. SmartKom features the situated understanding of possibly imprecise, ambiguous, or incomplete multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations [3]. 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 a mixed-initiative dialogue system, which integrates the fusion of the user's simultaneous input

modalities with the fission of coordinated output modalities of a life-like character that serves as an interface agent.

SmartKom is the follow-up project to Verbmobil (1993-2000) and reuses some of Verbmobil's components for the understanding of spontaneous dialogues [11]. In this paper, we will first present SmartKom's multimodal dialogue paradigm and discuss the portability of its dialogue backbone. In the core of the paper, the architecture, computational mechanisms, and discourse representations underlying modality fusion and modality fission are described.

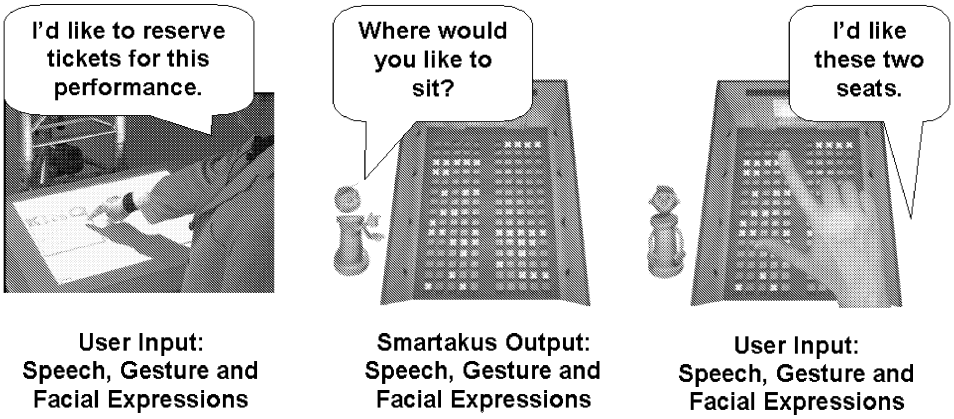

Figure 1: Multimodal Interaction with SmartKom

2 SmartKom's 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. [6], [7]). 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.

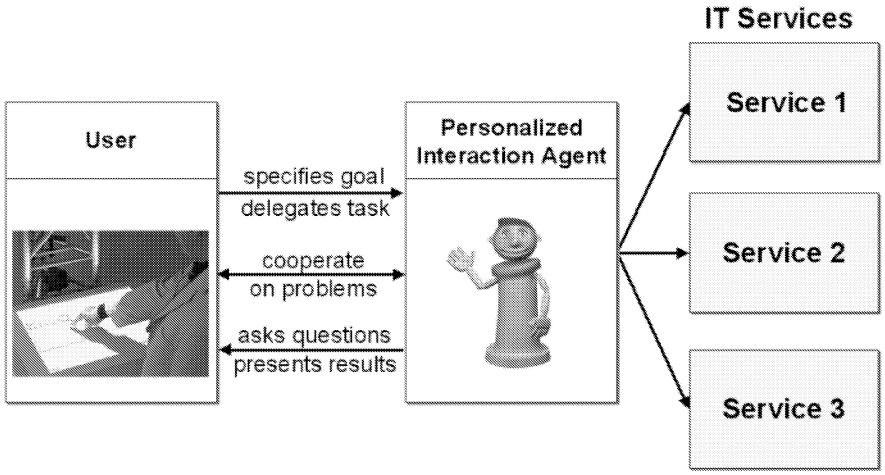

Figure 2: SmartKom's Situated Delegation-oriented Dialogue Paradigm

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. Smartakus is modeled in 3D Studio Max. It is a self-animated interface agent with a large repertoire of gestures, postures and facial expressions. Smartakus uses body language to notify the user that it is waiting for his input, that it is listening to him, that is has problems to understand his input, or that it is trying hard to find an answer to his question.

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. [8]). 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. [5]) 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.

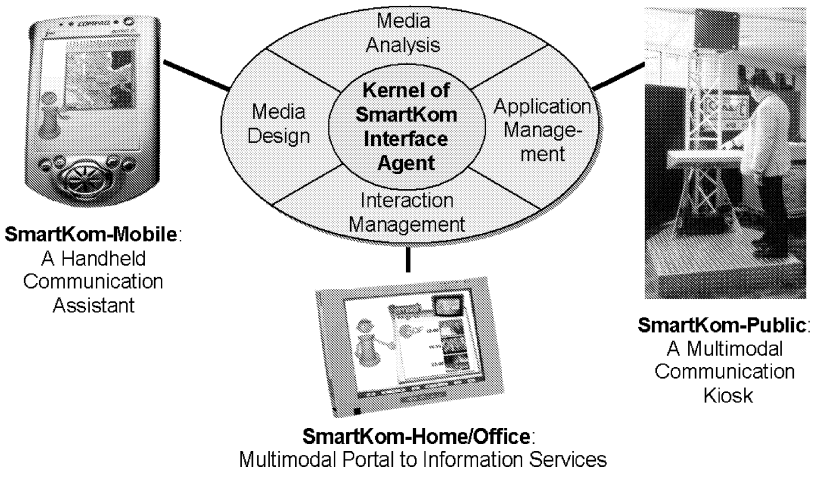

Figure 3: Three Versions of SmartKom 3.1

Three versions of SmartKom are currently available (see figure 3):

- *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 movie theaters. Users can also access their personalized webservices. 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. An additional camera is used to capture images of documents or 3D objects that the user wants to include in multimedia messages composed with the help of SmartKom.

- *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. Speech input can be combined with pen-based pointing and a simplified version of the Smartakus interface agent combines speech output, gestures and facial expressions.

- *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 TVs, VCRs and DVD players, and accesses 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 Processing Multimodal Discourse: From Modality Fusion to Modality Fission

Figure 4 shows the control GUI of the fully operational SmartKom 3.1 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

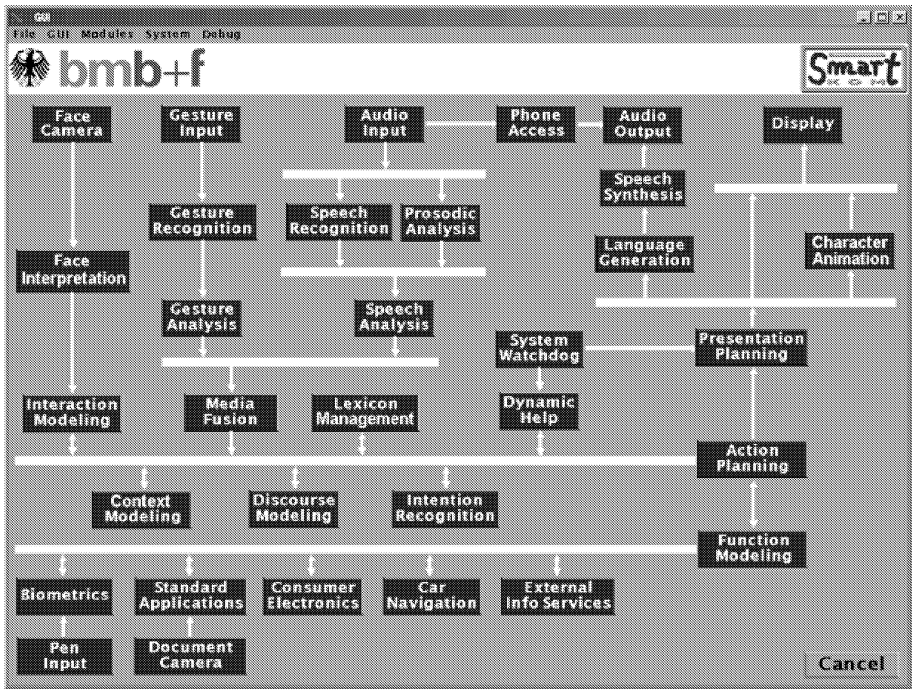

Figure 4: GUI for Tracing the Data and Control Flow in SmartKom 3.1

SmartKom 3.1 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.

SmartKom is based on a multi-blackboard architecture with parallel processing threads that support the multimodal fusion and fission processes. All modules shown in figure 4 are realized as separate processes on distributed computers, that run either Windows or Linux. Each module is implemented in C, C++, Java, or Prolog. The underlying integration software is based on Verbmobil's testbed software framework [11].

The information structures exchanged via the various blackboards are encoded in XML using the Multimodal Markup Language (M3L). M3L is defined by a set of XML schemas. For example, the word hypothesis graph and the gesture hypothesis graph, the hypotheses about facial expressions, the media fusion results, and the presentation goal 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 discourse processing.

Various types of unification, overlay, constraint solving, and planning are the fundamental computational processes involved in Smartkom's modality fusion and fission components. Overlay is a binary operation over two typed feature structures that is used for the refinement and validation of intentional hypotheses. Unlike unification, it never fails, and is a nonmonotonic and noncommutative operation. Overlay operations are also used in SmartKom for anaphora and ellipsis resolution. It overwrites conflicting information when applied to a new user input that is overlayed to the representation of the previous discourse. Using the type hierarchy of the domain model it simply computes the least upper bound of the two feature structures and inherits parts of the previous discourse representation despite type clashes (cf. [2]).

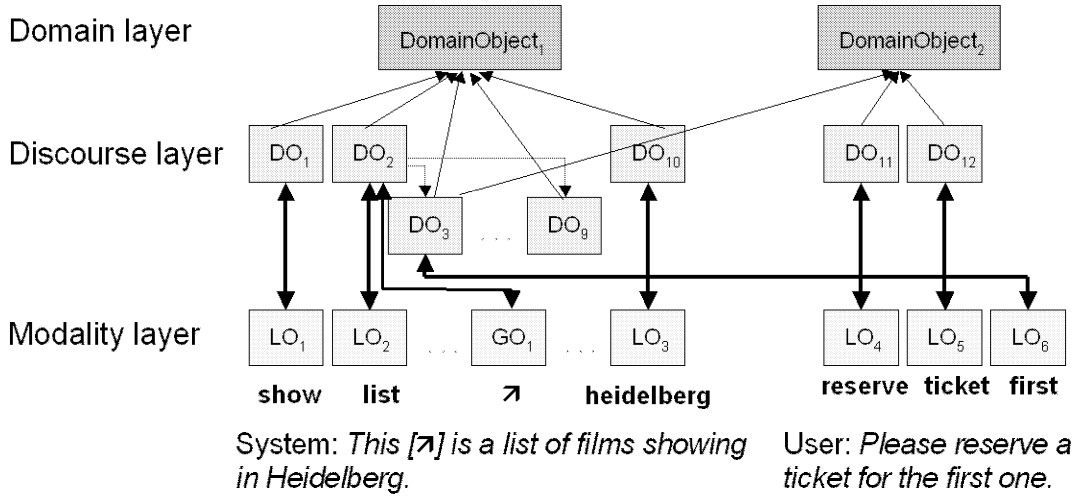

Figure 5: SmartKom's Three-tiered Representation of Multimodal Discourse

SmartKom uses a three-tiered representation of multimodal discourse, consisting of a domain layer, a discourse layer, and a modality layer. The modality layer consists of linguistic, visual, and gestural objects, that are linked to the corresponding discourse objects. Each discourse object can have various surface realizations on the modality layer. Finally, the domain layer links discourse objects with instances of the ontology-based domain model of SmartKom (cf. [2]). SmartKom's three-tiered discourse representation

makes it possible to resolve anaphora with non-linguistic antecedents. Let's consider the following dialogue segment:

User: *I would like to go to the movies tonight.*
Smartakus: *This [↗] is a list of films showing in Heidelberg.*
User: *Please reserve a ticket for the first one.*

The anaphoric reference "the first one" has no verbal antecedent, so that it can only be resolved by the visual context. SmartKom is a perceptive interface, since it takes the visual context of the user into account during multimodal communication. SmartKom's multimodal discourse representation keeps a record of all objects visible on the screen and the spatial relationships between them.

The action planner is a central component of SmartKom. It is based on a non-linear regression planning mechanism. It combines task planning to reach the user's goal with dialogue planning and response planning. It can trigger the function modelling component, which generates data access plans (eg. EPG) or device control plans (eg. TV or VCR). The action planner generates modality-free presentation goals that lead to modality fission in the presentation planner.

5 Modality Fusion in SmartKom

The analysis of the various input modalities of SmartKom is typically plagued by uncertainty and ambiguity. The speech recognition system produces a word hypothesis graph with acoustic scores, stating which word might have been spoken in a certain time frame. The prosody component generates a graph of hypotheses about clause and sentence boundaries with prosodic scores. The gesture analysis component produces a set of scored hypotheses about possible reference objects in the visual context. Finally, the interpretation of facial expressions leads to various scored hypotheses about the emotional state of the user. The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results.

One of the fundamental mechanisms implemented in SmartKom's modality fusion component is the unification of all scored hypothesis graphs and the application of mutual constraints in order to reduce the ambiguity and uncertainty of the combined analysis results. This approach was pioneered in our XTRA system, an early multimodal dialogue system, which assisted the user in filling out a tax form with a combination of typed natural language input and pointing (cf. [9]).

In SmartKom, the intention recognizer has the task to finally rank the remaining interpretation hypotheses and to select the most likely one, which is then passed on to the action planner. The modality fusion process is augmented by SmartKom's multimodal discourse model, so that the final ranking of the intention recognizer becomes highly context sensitive. The discourse component produces an additional score which states how good an interpretation hypothesis fits with the previous discourse. As soon as the modality fusion component finds a referential expression that is not combined with an unambiguous deictic gesture, it sends a request to the discourse component asking for reference resolution. If the resolution succeeds, the discourse components returns a fully instantiated domain object.

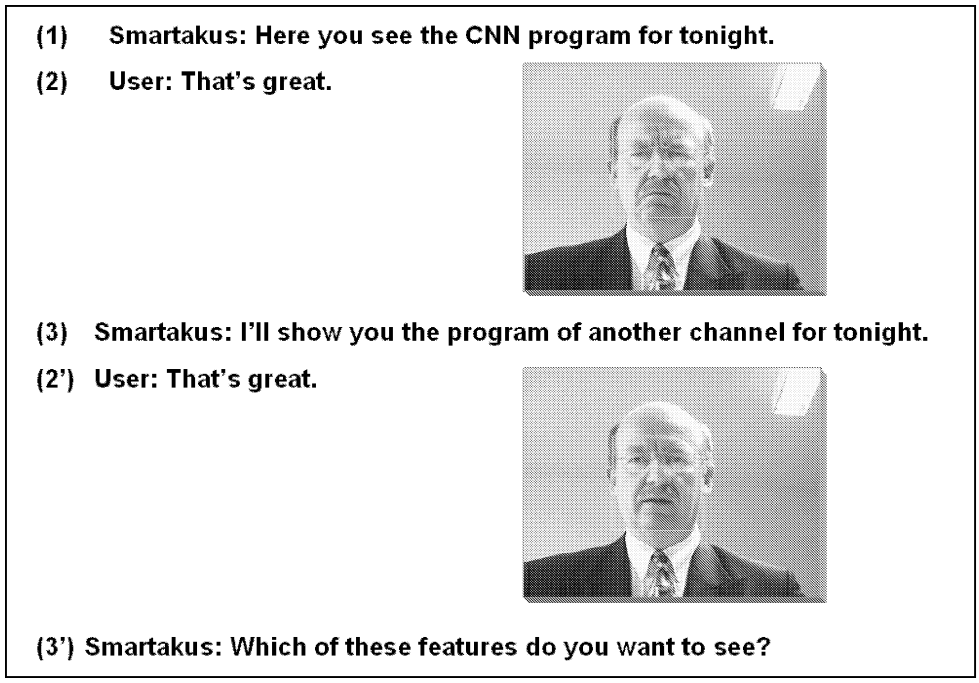

Figure 6: The Role of Facial Expressions in Multimodal Discourse

An exciting new possibility of multimodal fusion, that is being persued in SmartKom, is the generation of non-standard interpretations of a user's utterance, when the unification of the original input hypotheses fails. The example in figure shows that the default interpretation of the user's comment "That's great" as a positive remark cannot be unified with the very negative facial expression recognized by the face interpretation component. In such a case, an exceptional interpretation of "That's great" as an ironic or sarcastic comment may be triggered. Thus, the system interprets the user's multimodal input as negative feedback about the proposed TV channel, so that SmartKom's action planner has to check for an alternative channel. The same verbal input (2') is interpreted in a

positive way, if the user's utterance is combined with a neutral facial expression. Then the action planner can trigger a follow-up question like (3').

6 Modality Fission in SmartKom

In SmartKom, modality fission is controlled by a presentation planner. The input to the presentation planner is a presentation goal encoded in M3L as a modality-free representation of the system's intended communicative act. This M3L structure is either generated by the action planner or the dynamic help component, that can initiate clarification subdialogues. The presentation planning process can be adapted to various application scenarios via presentation parameters that encode user preferences (eg. spoken output is preferred by a car driver), specs of output devices (eg. size of the display), or the user's mother tongue (eg. German vs. English). A set of XSLT stylesheets is used to transform the M3L representation of the presentation goal according to the actual presentation parameter setting.

The presentation planner (cf. [10]) decomposes the presentation goal recursively into primitive presentation tasks using presentation strategies that vary with the discourse context, the user model, and ambient conditions. The presentation planner allocates different output modalities to primitive presentation tasks and decides whether specific media objects and presentation styles should be used by the media-specific generators for the visual and verbal elements of the multimodal output.

The presentation planner specifies presentation goals for the text generator, the graphics generator, and the animation generator. The animation generator selects appropriate elements from a large catalogue of basic behavioral patterns to synthesize fluid and believable actions of the Smartakus agent.

All planned deictic gestures of Smartakus must be synchronized with the graphical display of the corresponding media objects, so that Smartakus points to the intended graphical elements at the right moment. In addition, SmartKom's facial animation must be synchronized with the planned speech output.

SmartKom's lip synchronization approach is based on a simple mapping between phonemes and visemes. A viseme is a picture of a particular mouth position of Smartakus characterized by a specific jaw opening and lip rounding. Figure 7 shows the eight visemes used for the facial animation in SmartKom. The first row shows the visemes with unrounded lips and four different opening degrees, the second row the corresponding rounded lips. Only plosives and diphthongs are mapped to more than one viseme.

The speech synthesis module in SmartKom does not only produce audio data but also

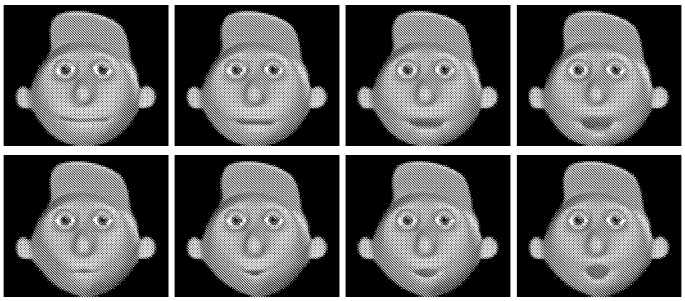

Figure 7: The Eight Visemes Used in SmartKom

a detailed representation of the phonemes and their exact time frames. Based on this representation and the phoneme-to-viseme mapping mentioned above, the presentation planner generates a lip animation script for Smartakus that is then executed by the display manager during speech output (for more details cf. [4]).

One of the distinguishing features of SmartKom's modality fission is the explicit representation of the generated multimodal presentation in M3L. This means that SmartKom follows the design principle "no presentation without representation" that ensures dialogue coherence in multimodal communication. The text generator provides a list of referential items that were mentioned in the last turn of the system. The display component generates an M3L representation of the current screen content, so that the discourse modeler can add the corresponding linguistic and visual objects to the discourse representation. Without such a representation of the generated multimodal presentation anaphoric, crossmodal, and gestural references of the user could not be resolved. Thus, it is an important insight of the SmartKom project that a multimodal dialogue system must not only understand and represent the user's multimodal input, but also its own multimodal output.

7 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. 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. Various types of unification, overlay, constraint solv-

ing, and planning are the fundamental computational processes involved in Smartkom's modality fusion and fission components.

Important extensions of the current SmartKom version include work on increased robustness of the media-specific analyzers, the expansion of the domains of discourse, the integration of multiple biometrics, and metacommunicative subdialogues between the user and his Smartakus assistant.

Acknowledgements

The SmartKom project is funded by the German Federal Ministry of Education and Research (BMBF) under grant 01 IL 905 K7. I would like to thank my SmartKom team at DFKI: Jan Alexandersson, Tilman Becker, Anselm Blocher, Ralf Engel, Markus Löckelt, Jochen Müller, Norbert Pfleger, Peter Poller, Norbert Reithinger, Michael Streit, Valentin Tschernomas, the SmartKom systems group headed by Gerd Herzog, and our numerous academic and industrial partners in the SmartKom project consortium.

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] Löckelt, M., Becker, T., Pfleger, N., Alexandersson, J.: *Making Sense of Partial.* In: Bos, J., Foster, M., Matheson, C. (eds.): Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002), Edinburgh, p. 101-107

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

[4] Müller, J., Poller, P., Tschernomas, V.: *Situated Delegation-Oriented Multimodal Presentation in SmartKom.* In: Krüger, A., Malaka, R. (eds.): Proceedings of the AAAI-2002 Workshop on Intelligent Situation-Aware Media and Presentations (ISAMP 2002), Edmonton, 2002, AAAI Press, Technical Report WS-02-08, p. 1-8.

[5] 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.

[6] Oviatt, S, Cohen, P.: *Multimodal Interfaces That Process What Comes Naturally.* In: CACM, 43, 3 2000, p. 45-53.

[7] Stock, O.: *Language-based Interfaces and Their Application for Cultural Tourism.* In: AI Magazine, Vol. 22, No. 1, Spring 2001, p. 85-97.

[8] 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

[9] Wahlster, W.: *User and Discourse Models for Multimodal Communication.* In: Sullivan, J., Tyler, S. (eds.): Intelligent User Interfaces, New York: ACM Press, 1991, p. 45-67

[10] Wahlster, W., André, E., Finkler, W., H.-J. Profitlich, H.-J., Rist, T.: *Plan-Based Integration of Natural Language and Graphics Generation.* In: Artificial Intelligence, 63, 1993, p. 387-427

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

[12] 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