In: Günter, A., Kruse, R., Neumann, B. (eds.): KI 2003: Advances in Artificial Intelligence.
Proceedings of the 26th German Conference on Artificial Intelligence, September 2003, Hamburg, Germany,
Pages 1 - 18, Berlin, Heidelberg: Springer, LNAI 2821, 2003.

Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture and Facial Expression

Wolfgang Wahlster

German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
www.dfki.de/~wahlster

Abstract. We introduce the notion of symmetric multimodality for dialogue systems in which all input modes (eg. speech, gesture, facial expression) are also available for output, and vice versa. A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output. We present the SmartKom system, that provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent. SmartKom represents a new generation of multimodal dialogue systems, that deal not only with simple modality integration and synchronization, but cover the full spectrum of dialogue phenomena that are associated with symmetric multimodality (including crossmodal references, one-anaphora, and backchannelling). We show that SmartKom's plug-and-play architecture supports multiple recognizers for a single modality, eg. the user's speech signal can be processed by three unimodal recognizers in parallel (speech recognition, emotional prosody, boundary prosody). Finally, we detail SmartKom's three-tiered representation of multimodal discourse, consisting of a domain layer, a discourse layer, and a modality layer.

1. Introduction

In-car electronics, dashboard computers, mobile devices (eg. PDAs, smartphones, wearables), and remote control systems for infotainment appliances are providing ever more functionality. However, along with greater functionality, the user must also come to terms with the greater complexity and a steeper learning curve. This complexity is compounded by the sheer proliferation of different devices lacking a standard user interface. Our SmartKom system (www.smartkom.org) is designed to support a wide range of collaborative and multimodal help dialogues, that allow users to intuitively and efficiently access the functionalities needed for their task. The application of the SmartKom technology is especially motivated in non-desktop scenarios, such as smart rooms, kiosks, or mobile environments. SmartKom features the situated understanding of possibly imprecise, ambiguous or incomplete multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations. SmartKom's interaction management is based on representing, reasoning, and exploiting models of the user, domain, task, context, and modalities. The system is capable

of real-time dialogue processing, including flexible multimodal turn-taking, back-channelling, and metacommunicative interaction.

Four major scientific goals of SmartKom were to:

- explore and design new symbolic and statistical methods for the seamless fusion and mutual disambiguation of multimodal input on semantic and pragmatic levels
- generalize advanced discourse models for spoken dialogue systems so that they can capture a broad spectrum of multimodal discourse phenomena
- explore and design new constraint-based and plan-based methods for multimodal fission and adaptive presentation layout
- integrate all these multimodal capabilities in a reusable, efficient and robust dialogue shell, that guarantees flexible configuration, domain independence and plug-and-play functionality

We begin by describing the notion of symmetric multimodality in section 2. In section 3, we introduce SmartKom as a flexible and adaptive multimodal dialogue shell and show in section 4 that SmartKom bridges the full loop from multimodal perception to physical action. SmartKom's distributed component architecture, realizing a multi-blackboard system, is described in section 5. Then in section 6 and 7, we describe SmartKom's methods for multimodal fusion and fission. Section 8 discusses the role of the three-tiered multimodal discourse model in SmartKom.

2. Towards Symmetric Multimodality

SmartKom provides full symmetric multimodality in a mixed-initiative dialogue system. Symmetric multimodality means that all input modes (speech, gesture, facial expression) are also available for output, and vice versa. A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output.

In this sense, SmartKom's modality fission component provides the inverse functionality of its modality fusion component, since it maps a communicative intention of the system onto a coordinated multimodal presentation (Wahlster 2002). SmartKom provides an anthropomorphic and affective user interface through an embodied conversational agent called Smartakus. This life-like character uses coordinated speech, gesture and facial expression for its dialogue contributions.

Thus, SmartKom supports face-to-face dialogic interaction between two agents that share a common visual environment: the human user and Smartakus, an autonomous embodied conversational agent. The "i"-shape of Smartakus is analogous to that used for information kiosks (see Fig. 1). 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 users that it is waiting for their input, that it is listening to them, that is has problems in understanding their input, or that it is trying hard to find an answer to their questions.

Most of the previous multimodal interfaces do not support symmetric multimodality, since they focus either on multimodal fusion (eg. QuickSet, see Cohen et al. 1977, or MATCH, see Johnston et al. 2002) or multimodal fission (eg. WIP, see Wahlster et al. 1993). But only true multimodal dialogue systems like SmartKom create a natural experience for the user in the form of daily human-to-human communication, by allowing both the user and the system to combine the same spectrum of modalities.

Fig. 1. Speech and Gesture for Input and Output

SmartKom is based on the situated delegation-oriented dialogue paradigm (SDDP): The user delegates a task to a virtual communication assistant (Wahlster et al. 2001). This cannot however be done in a simple command-and-control style for more complex tasks. Instead, a collaborative dialogue between the user and the agent elaborates the specification of the delegated task and possible plans of the agent to achieve the

user's intentional goal. The user delegates a task to Smartakus and helps the agent, where necessary, in the execution of the task. Smartakus accesses various digital services and appliances on behalf of the user, collates the results, and presents them to the user.

SmartKom represents a new generation of multimodal dialogue systems, that deal not only with simple modality integration and synchronization, but cover the full spectrum of dialogue phenomena that are associated with symmetric multimodality.

One of the technical goals of our research in the SmartKom project was to address the following important discourse phenomena that arise in multimodal dialogues:

- mutual disambiguation of modalities
- multimodal deixis resolution and generation
- crossmodal reference resolution and generation
- multimodal anaphora resolution and generation
- multimodal ellipsis resolution and generation
- multimodal turn-taking and backchannelling

Symmetric multimodality is a prerequisite for a principled study of these discourse phenomena.

3. Towards a Flexible and Adaptive Shell for Multimodal Dialogues

SmartKom was designed with a clear focus on flexibility, as a transmutable system that can engage in many different types of tasks in different usage contexts. The same software architecture and components are used in various roles that Smartakus can play in the following three fully operational experimental application scenarios:

- a communication companion that helps with phone, fax, email, and authentcation tasks
- an infotainment companion that helps to select media content and to operate various TV appliances (using a tablet computer as a mobile client)
- a mobile travel companion that helps with navigation and point-of-interest information retrieval in location-based services (using a PDA as a mobile client)

Currently, the user can delegate 43 types of complex tasks to Smartakus in multimodal dialogues. The SmartKom architecture supports not only simple multimodal command-and-control interfaces, but also coherent and cooperative dialogues with mixed initiative and a synergistic use of multiple modalities. SmartKom's plug-and-play architecture supports easy addition of new application services.

Fig. 2 shows a three-camera configuration of SmartKom that can be used as 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 theatres. 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 CCD camera and their 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 that can automatically tilt and pan, is used to capture images of documents or 3D objects that the user would like to include in multimedia messages composed with the help of SmartKom.

As a resource-adaptive multimodal system, the SmartKom architecture supports a flexible embodiment of the life-like character, that is used as a conversational partner

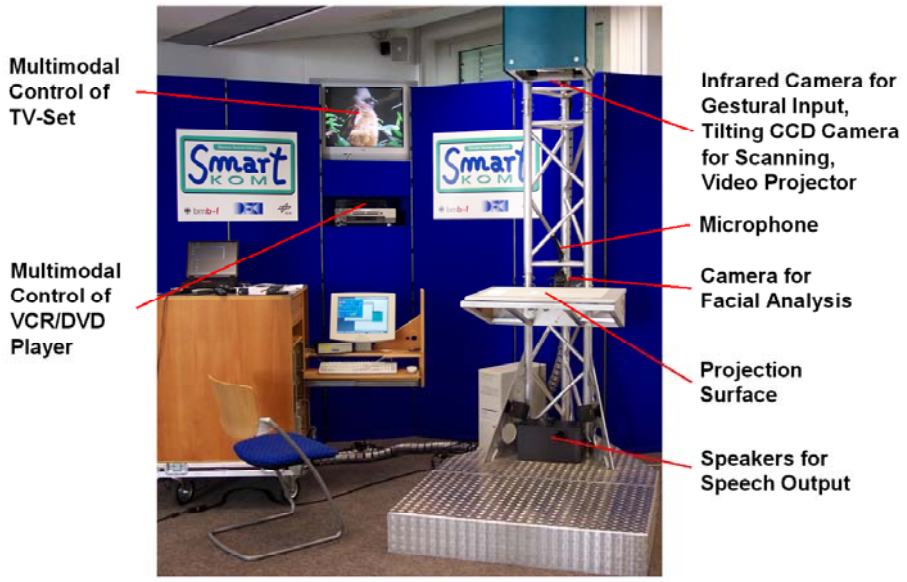

Fig. 2. SmartKom's Multimodal Input and Output Devices

in multimodal dialogue. The Smartakus agent is visualized either simply as a talking head together with an animated hand, when screen space is scarce, or as a full-body character, when enough screen space is available (see Fig. 4). Thus, Smartakus is embodied on a PDA differently than on a tablet computer or on the large top-projected screen used in the public information kiosk.

4. Perception and Action under Multimodal Conditions

SmartKom bridges the full loop from multimodal perception to physical action. Since the multimodal interaction with Smartakus covers both communicative and physical acts, the mutual understanding of the user and the system can often be validated by checking whether the user and the system "do the right thing" for completing the task at hand.

In a multimodal dialogue about the TV program, the user may browse a TV show database, create a personalized TV listing, and finally ask Smartakus to switch on the TV and tune to a specific program. Smartakus can also carry out more complex ac-

tions like programming a VCR to record the user's favourite TV show. Moreover, it can scan a document or a 3D object with its camera and then send the captured image to another person as an email attachment. Fig. 3 shows Dr. Johannes Rau, the German Federal President, using SmartKom's multimodal dialogue capabilities to scan the "German Future Award" trophy and send the scanned image via email to a colleague. This example shows that on the one hand, multimodal dialogue contributions can trigger certain actions of Smartakus. On the other hand, Smartakus may also ask the user to carry out certain physical actions during the multimodal dialogue.

Fig. 3. The German Federal President E-mailing a Scanned Image with SmartKom's Help

For example, Smartakus will ask the user to place their hand with spread fingers on a virtual scanning device, or to use a write-in field projected on the screen for their signature, when biometric authentication by hand contour recognition or signature verification is requested by a security-critical application. Fig. 3 shows a situation in which Smartakus has found an address book entry for the user, after they have introduced themself by name. Since the address book entry, which is partially visualized by SmartKom on the left part of the display, requests hand contour authentication for this particular user, Smartakus asks the user to place their hand on the marked area of the projected display, so that the hand contour can be scanned by its camera (see Fig. 4).

Since quite complex tasks can be delegated to Smartakus, there may be considerable delays in replying to a request. Our WOZ experiments and user tests with earlier prototypes of SmartKom, showed clearly that users want a simple and fast feedback on the state of the system in such situations. Therefore, a variety of adaptive perceptual feedback mechanisms have been realized in SmartKom.

In the upper left corner of a presentation, SmartKom can display a "magic eye" icon, that lights up while the processing of the user's multimodal input is proceeding (see the left part of Fig. 5). "Magic eye" is the common name applied to the green-glow tubes used in 1930's radio equipment to visually assist the listener in tuning a radio station to the point of greatest signal strength. Although SmartKom works in real-time, there may be some processing delays caused by corrupted input or complex disambiguation processes.

Fig. 4. Interactive Biometric Authentication by Hand Contour Recognition

An animated dashed line (see the left part of Fig. 5) circles the Smartakus character, while the system is engaged in an information retrieval task (e.g. access to maps, EPG, web sites). This type of feedback is used when screen space is scarce. When more screen space is available, an animation sequence that shows Smartakus working on a laptop is used for the same kind of feedback. When Smartakus is downloading a large file, it can show a progress bar to indicate to the user how the data transfer is going (see the right part of Fig. 5).

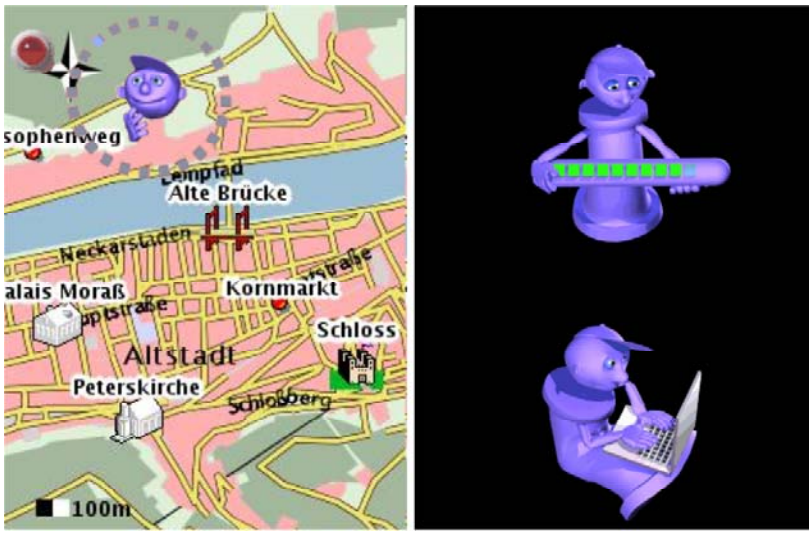

Fig. 5. Adaptive Perceptual Feedback on the System State

5. A Multi-Blackboard Platform with Ontology-Based Messaging

SmartKom is based on a distributed component architecture, realizing a multi-blackboard system. The integration platform is called MULTIPLATFORM (Multiple Language Target Integration Platform for Modules, see Herzog et al. 2003) and is built on top of open source software. The natural choice to realize an open, flexible and scalable software architecture, is that of a distributed system, which is able to integrate heterogeneous software modules implemented in diverse programming languages and running on different operating systems. SmartKom includes more than 40 asynchronously running modules coded in four different programming languages: C, C++, Java, and Prolog.

The MULTIPLATFORM testbed includes a message-oriented middleware. The implementation is based on PVM, which stands for parallel virtual machine. On top of PVM, a message-based communication framework is implemented based on the so-called publish/subscribe approach. In contrast to unicast routing known from multi-agent frameworks, that realize a direct connection between a message sender and a known receiver, MULTIPLATFORM is based on the more efficient multicast addressing scheme. Instead of addressing one or several receivers directly, the sender publishes a notification on a named message queue, so that the message can be forwarded to a list of subscribers. This kind of distributed event notification makes the communication framework very flexible as it focuses on the data to be exchanged and it decouples data producers and data consumers. Compared with point-to-point messaging used in multi-agent frameworks like OAA (Martin et al. 1999), the publish/subscribe scheme helps to reduce the number and complexity of interfaces significantly.

GCSI, the Galaxy Communicator Software Infrastructure (Seneff et al. 1999) architecture is also fundamentally different from our approach. The key component of GCSI is a central hub, which mediates the interaction among various servers that realize different dialog system components. Within MULTIPLATFORM there exists no such centralized controller component, since this could become a bottleneck for more complex multimodal dialogue architectures.

In order to provide publish/subscribe messaging on top of PVM, we have added another software layer called PCA (Pool Communication Architecture). In MULTIPLATFORM, the term data pool is used to refer to named message queues. Every single pool can be linked with a pool data format specification in order to define admissible message contents. The messaging system is able to transfer arbitrary data contents, and provides excellent performance characteristics (see Herzog et al. 2003).

In SMARTKOM, we have developed M3L (Multimodal Markup Language) as a complete XML language that covers all data interfaces within this complex multimodal dialog system. Instead of using several quite different XML languages for the various data pools, we aimed at an integrated and coherent language specification, which includes all sub-structures that may occur on the different pools. In order to make the specification process manageable and to provide a thematic organization, the M3L language definition has been decomposed into about 40 schema specifications. The basic data flow from user input to system output continuously adds further processing results so that the representational structure will be refined, step-by-step.

The ontology that is used as a foundation for representing domain and application knowledge is coded in the ontology language OIL. Our tool OIL2XSD (Gurevych et al. 2003) transforms an ontology written in OIL (Fensel et al. 2001) into an M3L compatible XML Schema definition. The information structures exchanged via the various blackboards are encoded in 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. It provides information about segmentation, synchronization, and 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. Alexandersson and Becker 2003) of typed feature structures encoded in M3L for discourse processing.

Application developers can generate their own multimodal dialogue system by creating knowledge bases with application-specific interfaces, and plugging them into the reusable SmartKom shell. It is particularly easy to add or remove modality analyzers or renderers, even dynamically while the system is running. This plug and play of modalities can be used to adjust the system's capability to handle different demands of the users, and the situative context they are currently in. Since SmartKom's modality analyzers are independent from the respective device-specific recognizers, the system can switch in real-time, for example, between video-based, pen-based or touch-based gesture recognition. SmartKom's architecture, its dialogue backbone, and its fusion and fission modules are reusable across applications, domains, and modalities.

MULTIPLATFORM is running on the SmartKom server that consists of 3 dual Xeon 2.8 GHz processors. Each processor uses 1.5 GB of main memory. One processor is running under Windows 2000, and the other two under Linux. The mobile clients (an iPAQ Pocket PC for the mobile travel companion and a Fujitsu Stylistic 3500X webpad for the infotainment companion) are linked to the SmartKom server via WaveLAN.

6. Reducing Uncertainty and Ambiguity by Modality Fusion

The analysis of the various input modalities by 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. All the recognizers produce time-stamped hypotheses, so that the fusion process can consider various temporal constraints. The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results. By fusing symbolic and statistical

information derived from the recognition and analysis components for speech, prosody, facial expression and gesture, SmartKom can correct various recognition errors of its unimodal input components and thus provide a more robust dialogue than a unimodal system.

In principle, modality fusion can be realized during various processing stages like multimodal signal processing, multimodal parsing, or multimodal semantic processing. In SmartKom, we prefer the latter approach, since for the robust interpretation of possibly incomplete and inconsistent multimodal input, more knowledge sources become available on later processing stages. An early integration on the signal level allows no backtracking and reinterpretation, whereas the multimodal parsing approach has to pre-specify all varieties of crossmodal references, and is thus unable to cope robustly with unusual or novel uses of multimodality. However, some early fusion is also used in SmartKom, since the scored results from a recognizer for emotional prosody (see Batliner et al. 2000) are merged with the results of a recognizer for affective facial expression. The classification results are combined in a synergistic fashion, so that a hypothesis about the affective state of the user can be computed.

In SmartKom, the user state is used for example, in the dialogue-processing backbone to check whether the user is satisfied or not with the information provided by Smartakus. It is interesting to note that SmartKom's architecture supports multiple recognizers for a single modality. In the current system, prosody is evaluated by one recognizer for clause boundaries and another recognizer for emotional speech. This means that the user's speech signal is processed by three unimodal recognizers in parallel (speech recognition, emotional prosody, boundary prosody).

The time stamps for all recognition results are extremely important since the confidence values for the classification results may depend on the temporal relations between input modalities. For example, experiments in SmartKom have shown that the results from recognizing various facial regions (like eye, nose, and mouth area) can be merged to improve recognition results for affective states like anger or joy. However, while the user is speaking, the mouth area does not predict emotions reliably, so that the confidence value of the mouth area recognizer must be decreased. Thus, SmartKom's modality fusion is based on adaptive confidence measures, that can be dynamically updated depending on the synchronization of input modalities.

One of the fundamental mechanisms implemented in SmartKom's modality fusion component is the extended 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 that assisted the user in filling out a tax form with a combination of typed natural language input and pointing gestures (Wahlster 1991). QuickSet uses a similar approach (Cohen et al. 1997).

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 that states how good an interpretation hypothesis fits to the previous discourse (Pfleger et. al. 2002).

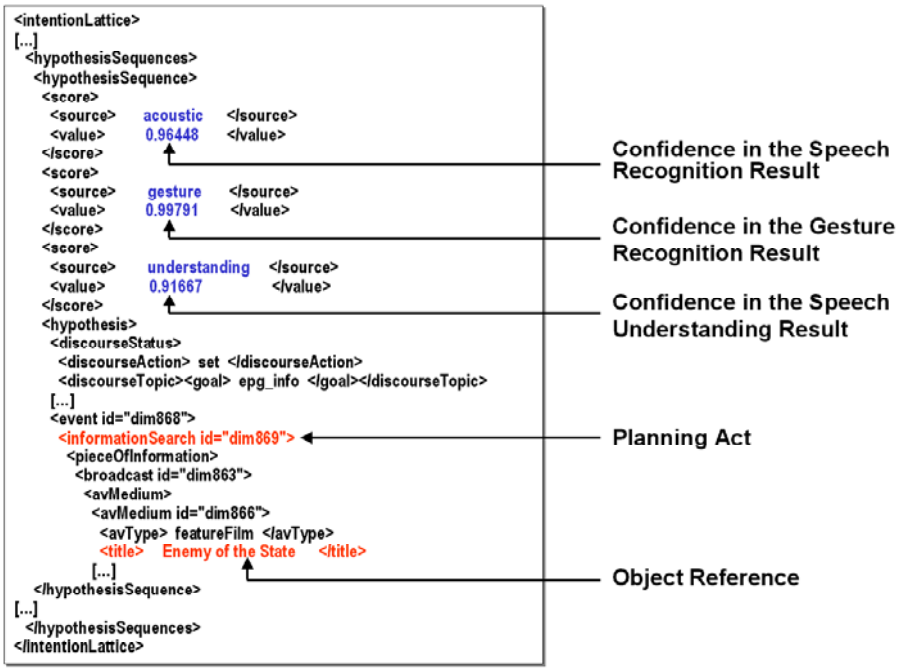

Fig. 6. M3L Representation of an Intention Lattice Fragment

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 component returns a completely instantiated domain object.

Fig. 6 shows an excerpt from the intention lattice for the user's input "I would like to know more about this [deictic pointing gesture]". It shows one hypothesis sequence with high scores from speech and gesture recognition. A potential reference object for the deictic gesture (the movie title "Enemy of the State") has been found in the visual context. SmartKom assumes that the discourse topic relates to an electronic program guide and the intended action of Smartakus refers to the retrieval of information about a particular broadcast.

7. Plan-based 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 generated by either an action planner or the dynamic help component, which 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), output devices (eg. size of the display), or the user's native language (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 recursively decomposes the presentation goal into primitive presentation tasks using 121 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 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. Only plosives and diphthongs are mapped to more than one viseme.

```
<presentationTask>
 <presentationGoal>
  <inform> <informFocus> <RealizationType>list </RealizationType> </informFocus> </inform>
  <abstractPresentationContent>
      <discourseTopic> <goal>epg_browse</goal> </discourseTopic>
      <informationSearch id="dim24"><tvProgram id="dim23">
          <broadcast><timeDeictic id="dim16">now</timeDeictic>
                <between>2003-03-20T19:42:32 2003-03-20T22:00:00</between>
            <channel><channel id="dim13"/> </channel>
           </broadcast></tvProgram>
       </informationSearch>
   <result> <event>
      <pieceOfInformation>
        <tvProgram id="ap_3">
          <broadcast> <beginTime>2003-03-20T19:50:00</beginTime>
                       <endTime>2003-03-20T19:55:00</endTime>
                      <avMedium> <title>Today's Stock News</title></avMedium>
                       <channel>ARD</channel>
          </broadcast>……..</event>
   </result>
</presentationGoal>
</presentationTask>
```

Fig. 7. A Fragment of a Presentation Goal, as specified in M3L

One of the distinguishing features of SmartKom's modality fission is the explicit representation of generated multimodal presentations in M3L. This means that SmartKom ensures dialogue coherence in multimodal communication by following the design principle "no presentation without representation". 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.

Fig. 7 shows the modality-free presentation goal that is transformed into the multimodal presentation shown in Fig. 8, by SmartKom's media fission component and unimodal generators and renderers. Please note that all the graphics and layout shown in Fig. 8 are generated on the fly and uniquely tailored to the dialogue situation, ie. nothing is canned or pre-programmed. The presentation goal shown in Fig. 7 is coded in M3L and indicates that a list of broadcasts should be presented to the user. Since there is enough screen space available and there are no active constraints on using graphical output, the strategy operators applied by the presentation planner lead to a graphical layout of the list of broadcasts. In an eyes-busy situation (eg. when the user is driving a car), SmartKom would decide that Smartakus should read the list of retrieved broadcasts to the user. This shows that SmartKom's modality fission process is highly context-aware and produces tailored multimodal presentations.

The presentation planner decides that the channel should be rendered as an icon, and that only the starting time and the title of the individual TV item should be mentioned in the final presentation.

Fig. 8. . A Dynamically Generated Multimodal Presentation based on a Presentation Goal

In the next section, we show how the visual, gestural and linguistic context stored in a multimodal discourse model can be used to resolve crossmodal anaphora. We will use the following dialogue excerpt as an example:

(1) User: I would like to go to the movies tonight.
(2) Smartakus: [displays a list of movie titles] This is a list of films showing in Heidelberg.
(3) User: Hmm, none of these films seem to be interesting... Please show me the TV program.
(4) Smartakus: [displays a TV listing] Here [points to the listing] is a listing of tonight's TV broadcasts. (see Fig. 7)
(5) User: Please tape the third one!

8. A Three-Tiered Multimodal Discourse Model

Discourse models for spoken dialogue systems store information about previously mentioned discourse referents for reference resolution. However, in a multimodal dialogue system like SmartKom, reference resolution relies not only on verbalized, but also on visualized information. A multimodal discourse model must account for entities not explicitly mentioned (but understood) in a discourse, by exploiting the verbal, the visual and the conceptual context. Thus, SmartKom's multimodal discourse representation keeps a record of all objects visible on the screen and the spatial relationships between them.

An important task for a multimodal discourse model is the support of crossmodal reference resolution. 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 SmartKom's ontology-based domain model (cf. Loeckelt et al. 2002). SmartKom's three-tiered discourse representation makes it possible to resolve anaphora with non-linguistic antecedents. SmartKom is able to deal with multimodal one-anaphora (eg. "the third one") and multimodal ordinals ("the third broadcast in the list").

SmartKom's multimodal discourse model extends the three-tiered context representation of (Luperfoy, 1991) by generalizing the linguistic layer to that of a modality layer (see Fig. 9). An object at the modality layer, encapsulates information about the concrete realization of a referential object depending on the modality of presentation (eg. linguistic, gestural, visual). Another extension is that objects at the discourse layer may be complex compositions that consist of several other discourse objects (cf. Salmon-Alt 2001). For example, the user may refer to an itemized list shown on SmartKom's screen as a whole, or they may refer to specific items displayed in the list. In sum, Smartkom's multimodal discourse model provides a unified representation of discourse objects introduced by different modalities, as a sound basis for crossmodal reference resolution.

The modality layer of SmartKom's multimodal discourse model contains three types of modality objects:

- Linguistic Objects (LOs): For each occurrence of a referring expression in SmartKom's input or output, one LO is added .
- Visual Objects (VOs): For each visual presentation of a referrable entity, one VO is added.
- Gesture Objects (GOs) For each gesture performed either by the user or the system, a GO is added.

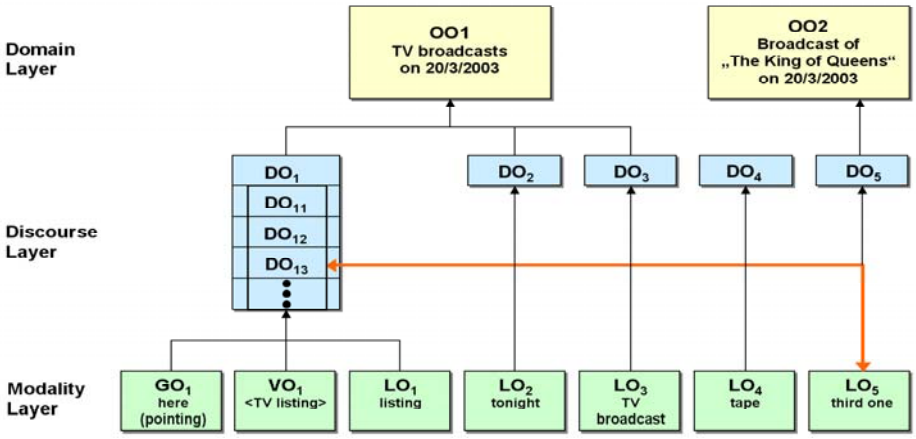

Fig. 9. An Excerpt from SmartKom's Multimodal Discourse Model

Each modality object is linked to a corresponding discourse object. The central layer of the discourse model is the discourse object layer. A Discourse Object (DO) represents a concept that can serve as a candidate for referring expressions, including objects, events, states and collections of objects. When a concept is newly introduced by a multimodal communicative act of the user or the system, a DO is created. For each concept introduced during a dialogue, there exists only one DO, regardless of how many modality objects mention this concept.

The compositional information for the particular DOs that represent collections of objects, is provided by partitions (Salmon-Alt, 2001). A partition provides information about possible decompositions of a domain object. Such partitions are based either on perceptual information (eg. a set of movie titles visible on the screen) or discourse information (eg. "Do you have more information about the first and the second movie" in the context of a list of movie titles presented on the screen). Each element of a partition is a pointer to another DO, representing a member of the collection. The elements of a partition are distinguishable from one another by at least one differentiation criterion like their relative position on the screen, their size, or color. For instance, the TV listing shown in Fig. 8 is one DO that introduces 13 new DOs corresponding to particular broadcasts.

The domain object layer provides a mapping between a DO and instances of the domain model. The instances in the domain model are Ontological Objects (OO) that provide a semantic representation of actions, processes, and objects. SmartKom's domain model is described in the ontology language OIL (Fensel et al. 2001).

Let us discuss an example of SmartKom's methodology for multimodal discourse modeling. The combination of a gesture, an utterance, and a graphical display that is generated by SmartKom's presentation planner (see Fig. 8) creates the gestural object GO1, the visual object VO1 and the linguistic object LO1 (see Fig. 9). These three objects at the modality layer, are all linked to the same discourse object DO1, that refers to the ontological object OO1 at the domain layer. Note that DO1 is composed of 13 subobjects. One of these subobjects is DO13, that refers to OO2, the broadcast of "The King of Queens" on 20 March 2003 on the ZDF channel. Although there is no linguistic antecedent for the one-anaphora "the third one", SmartKom can resolve the reference with the help of its multimodal discourse model. It exploits the information, that the spatial layout component has rendered OO1 into a horizontal list, using the temporal order of the broadcasts as a sorting criterion. The third item in this list is DO13, which refers to OO2. Thus, the crossmodal one-anaphora "the third one" is correctly resolved and linked to the broadcast of "The King of Queens" (see Fig. 9).

During the analysis of turn (3) in the dialogue excerpt above, the discourse modeler receives a set of hypotheses. These hypotheses are compared and enriched with previous discourse information, in this example stemming from (1). Although (3) has a different topic to (1) (it requests information about the cinema program, whereas (3) concerns the TV program), the temporal restriction (tonight) of the first request is propagated to the interpretation of the second request. In general, this propagation of information from one discourse state to another is obtained by comparing a current intention hypothesis with previous discourse states, and by enriching it (if possible) with consistent information. For each comparison, a score has to be computed reflecting how well this hypothesis fits in the current discourse state. For this purpose, the non-monotonic overlay operation (an extended probabilistic unification-like scheme, see Alexandersson and Becker 2003) has been integrated into SmartKom as a central computational method for multimodal discourse processing.

9. Conclusion

We have introduced the notion of symmetric multimodality for dialogue systems in which all input modes (eg. speech, gesture, facial expression) are also available for output, and vice versa. We have shown that a dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output. We presented the SmartKom system, that provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent.

The industrial and economic impact of the SmartKom project is remarkable. Up to now, 51 patents concerning SmartKom technologies have been filed by members of the SmartKom consortium, in areas such as speech recognition (13), dialogue man-

agement (10), biometrics (6), video-based interaction (3), multimodal analysis (2), and emotion recognition (2).

In the context of SmartKom, 59 new product releases and prototypes have been surfacing during the project's life span. 29 spin-off products have been developed by the industrial partners of the SmartKom consortium at their own expense.

SmartKom's MULTIPLATFORM software framework (see section 5) is being used at more than 15 industrial and academic sites all over Europe and has been selected as the integration framework for the COMIC (COnversational Multimodal Interaction with Computers) project funded by the EU (Catizone et al. 2003).

The sharable multimodal resources collected and distributed during the SmartKom project will be useful beyond the project's life span, since these richly annotated corpora will be used for training, building, and evaluating components of multimodal dialogue systems in coming years. 448 multimodal Wizard-of-OZ sessions resulting in 1.6 terabytes of data have been processed and annotated (Schiel et al. 2002). The annotations contain audio transcriptions combined with gesture and emotion labeling.

Acknowledgements

The SmartKom project has been made possible by funding from the German Federal Ministry of Education and Research (BMBF) under grant 01 IL 905. I would like to thank my SmartKom team at DFKI: Jan Alexandersson, Tilman Becker, Anselm Blocher (project management), Ralf Engel, Gerd Herzog (system integration), Heinz Kirchmann, Markus Löckelt, Stefan Merten, Jochen Müller, Alassane Ndiaye, Rainer Peukert, Norbert Pfleger, Peter Poller, Norbert Reithinger (module coordination), Michael Streit, Valentin Tschernomas, and our academic and industrial partners in the SmartKom project consortium: DaimlerChrysler AG, European Media Laboratory GmbH, Friedrich-Alexander University Erlangen-Nuremberg, International Computer Science Institute, Ludwig-Maximilians University Munich, MediaInterface GmbH, Philips GmbH, Siemens AG, Sony International (Europe) GmbH, Stuttgart University for the excellent and very successful cooperation.

References

Alexandersson, A., Becker, T. (2003): The Formal Foundations Underlying Overlay. In: Proc. of the Fifth International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands, 2003, p. 22-36.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K. (2000): The Recognition of Emotion. In: W. Wahlster (ed.): Verbmobil: Foundations of Speech-to-Speech Translations, Berlin, New York: Springer, p. 122-130.

Catizone, R., Setzer, A., Wilks, Y. (2003): Multimodal Dialogue Management in the COMIC Project, In: Workshop on 'Dialogue Systems: interaction, adaptation and styles of management', European Chapter of the Association for Computational Linguistics (EACL), Budapest, Hungary, April 2003.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J. (1997): QuickSet: Multimodal interaction for distributed applications, In: Proc. of the Fifth International Multimedia Conference (Multimedia '97), ACM Press, p. 31-40.

Fensel, D., van Harmelen, F., Horrocks, I. McGuinness, D., Patel-Schneider, P. (2001): OIL: An Ontology Infrastructure for the Semantic Web. In: IEEE Intelligent Systems, 16(2), 2001. p. 38-45.

Gurevych, I., Merten, S., Porzel, R. (2003): Automatic Creation of Interface Specifications from Ontologies. In: Proc. of the HLT-NAACL'03 Workshop on the Software Engineering and Architecture of Language Technology Systems (SEALTS), Edmonton, Canada.

Herzog, G., Kirchmann, H., Merten S., Ndiaye, A. Poller, P. (2003): MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems. In: Proc. of the HLT-NAACL'03 Workshop on the Software Engineering and Architecture of Language Technology Systems (SEALTS), Edmonton, Canada.

Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S, Maloor, P. (2002): MATCH: An Architecture for Multimodal Dialogue Systems. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics., Philadelphia, p. 376-383.

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

Luperfoy, S. (1991): Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions. Ph.D. thesis, University of Texas at Austin.

Martin, D. L. , Cheyer, A. J., Moran, D.B. (1999): The Open Agent Architecture: A Framework for Building Distributed Software Systems. Applied Artificial Intelligence, 13(1–2), p. 91–128.

Pfleger, N., Alexandersson,J., Becker, T. (2002): Scoring Functions for Overlay and their Application in Discourse Processing. In: Proc. of KONVENS 2002, Saarbrücken, Germany, 2002, p. 139-146.

Salmon-Alt S. (2001): Reference Resolution within the Framework of Cognitive Grammar. In: Proc. of International Colloquium on Cognitive Science, San Sebastian, Spain, May 2001, p. 1-15.

Schiel, F., Steininger, S., Türk. U. (2002): The SmartKom Multimodal Corpus at BAS. In: Proc. of the 3rd Language Resources & Evaluation Conference (LREC) 2002, Las Palmas, Gran Canaria, Spain, p. 35-41.

Seneff, S., Lau, R., Polifroni, J. (1999): Organization, Communication, and Control in the Galaxy-II Conversational System. In: Proc. of Eurospeech' 99, Budapest, Hungary, p. 1271–1274.

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

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

Wahlster, W., Reithinger N., Blocher, A. (2001): SmartKom: Multimodal Communication with a Life-Like Character. In: Proc. of Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001, Vol. 3, p. 1547–1550.

Wahlster, W. (2002): SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions. In: Proc. of the 1st International Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan, 2002. p. 213–225.