Multimodality and Ambient Intelligence

Anton Nijholt
Centre of Telematics and Information Technology
University of Twente, PO Box 217
7500 AE Enschede, The Netherlands
anijholt@cs.utwente.nl

Abstract
In this report we discuss multimodal interface technology. We present examples of multimodal interfaces and show problems and opportunities. Fusion of modalities is discussed and some roadmap discussions on research in Multimodality are summarized. This report also discusses future developments where rather than communicating with a single computer, users communicate with their environment using multimodal interactions and where the environmental interface has perceptual competence that includes being able to interpret what is going on in the environment. We contribute roles to virtual humans in order to allow daily users of future computing environments to establish relationships with the environments, or more in particular, these virtual humans.

Multimodality and Ambient Intelligence[†]

Anton Nijholt
Centre of Telematics and Information Technology
University of Twente, PO Box 217
7500 AE Enschede, The Netherlands
anijholt@cs.utwente.nl

Abstract
In this report we discuss multimodal interface technology. We present examples of multimodal interfaces and show problems and opportunities. Fusion of modalities is discussed and some roadmap discussions on research in Multimodality are summarized. This report also discusses future developments where rather than communicating with a single computer, users communicate with their environment using multimodal interactions and where the environmental interface has perceptual competence that includes being able to interpret what is going on in the environment. We contribute roles to virtual humans in order to allow daily users of future computing environments to establish relationships with the environments, or more in particular, these virtual humans.

Keywords: multimodal interactions, language and speech technology, ambient intelligence, environmental interfaces, virtual reality, embodied conversational agents

1. Introduction

It is tempting to assume that language and speech are the most natural modalities for human-human communication. There are good reasons to attack this assumption. In communication between humans non-verbal aspects are important as well. When trying to establish efficient communication we may concentrate on the literal meaning of what is being said. However, even when efficiency is an aim, it is often useful to take other communication aspects into account. Whom are we attracted to and do we want to speak? What roles are played by facial expression and body posture? Gaze, eye contact or a smile may mean more for human-human communication than a right choice of words. A blink of the eye may give a completely different meaning to what was just said or it may even replace a long sequence of words. The same holds for avoiding eye contact. As observed by the linguist Robin Lakoff, in the majority of cases people don't say what they mean, but what they want to say. Nevertheless they can communicate with each other. Speech and language need to be interpreted in a context where other communication modalities are available (seeing, touching, smelling, tasting).

When looking at human-computer interaction we can certainly try to model speech and language in their context. Attempts are being made to model dialogues between humans and computers using natural speech in the context of a particular domain, a particular task and even taking into account knowledge of a particular user. However, there is a second reason to consider a different view on the importance of speech and language. When looking at computer technology that aims at assisting humans in their daily work and recreation we want non-intrusive technology. The computer should interpret what is going on in an environment where there are one or more occupants, assist when asked or take the initiative when useful, either for the occupants or for future use. This situation differs from traditional human-computer interaction where there is explicit addressing of *the* computer.

[†] This report will appear as a chapter in *Algorithms and Ambient Intelligence*. W.F.J. Verhaegh, E.H.L. Aarts & J. Korst (eds.), Kluwer Academic Publishers, Boston/Dordrecht/London, 2003.

In documents of the European Community we see the mentioning of "the real world being the interface". In particular the 'Ambient Intelligence' theme of the European Framework 6 Research Programme demands systems, which are capable of functioning within natural, unconstrained environments - within scenes. Hence notions of space, time and physical laws play a role and they are maybe more important than the immediate and conscious communication between a human and a computer screen. In a multi-sensory environment, maybe supported with embedded computer technology, the environment can capture and interpret what the user is doing, maybe anticipating what the user is doing or wanting, and therefore can be pro-active and re-active, just capturing what is going on for later use, or acting as an environment that assists the user in real-time or collaborates with the user in real-time. Hence, speech and language, or whatever modalities we want to consider, are part of a context where temporal and spatial issues play important roles.

All this certainly does not mean that traditional ways of interacting with computers will disappear, at least not in the next ten years. There are tasks that can be done efficiently with keyboard and mouse. Constructing a text, for example. Other tasks can profit from speech input, a head nod or shake detector, a touch screen, a data glove allowing manipulation of virtual objects, translation of body movements to interpretations, e.g., a choreography, force-feedback to help in the design of a gear-lever and, obviously, there will be tasks where combinations of these modalities can help to obtain efficient or entertaining communication with the computer. However, as just mentioned, this is about communication with a computer and about using the computer to perform a certain task. Ubiquitous computing technology will spread computing and communication power all around us. That is, in our daily work and home environment and we need computers with perceptual competence in order to profit from this technology.

Embedded computers will appear everywhere but at the same time they will become less visible as computers. Nevertheless, we need interfaces. For some applications these can be big, for example the walls of room or the surface of a tabletop. In other situations the interface may consist of invisible attentive sensors. Current interface technology works well in controlled and constrained conditions. Extending the existing technology to obtain multisensorial interfaces capable of functioning within natural environments is an important challenge.

In this report we look at and discuss examples of multimodal and environmental interfaces. In our research we have introduced multimodal interfaces for several applications. Some of them will be discussed here, showing problems and opportunities. Although most of the work has been done in the context of extensions of traditional interfaces, we think that designing these extensions as integrations of different interaction modalities, will allow us to extend the research to environmental interfaces and to build on the results that have been obtained. One reason to think so is the emphasis in our research on actions and interactions in the context of visualized worlds, either the graphical user interface to which references may be made using different modalities, or virtual worlds that have been used as a laboratory where real-world actions and interactions can be visualized.

In section 2 we discuss some of our multimodal interaction environments. In section 3 we look at multimodal interaction modeling in general. We shortly introduce the results of a roadmap discussion on multimodal research held a few years ago and explain that the views developed there were rather conservative. Nevertheless, the research advocated there is more than useful. In section 4 we discuss the role of computers as social actors. Should we continue to see computers as social actors or does new technology for ubiquitous computing invite us to consider other paradigms? Section 5 is about embodied conversational agents or virtual humans. In our view they can play the role of social actors, being able to build relationships

with humans working and recreating in computer-supported environments. Finally, in section 6 we repeat the main findings of this report.

2. Multimodality: Background and Examples

In this section we show examples of multimodal systems we have been working on. Generally our approach has been bottom-up. That is, we started with a specific practical problem rather than with a theoretical problem. The assumption was, and still is, that theory on syntactic, semantic and pragmatic aspects of multimodality is rather underdeveloped. On the other hand, interface technology allows the building of multimodal interfaces, the use of many nontraditional interaction devices (6-DOF navigation devices, haptic devices, eye trackers, camera's for tracking gestures and body movement, facial expression recognition, etc.) and that rather than waiting for a more mature theory the development of theory can profit from experiences and experiments with multimodal interfaces and environments. Below we introduce four environments. The virtual music center, a 3D environment inhabited by several agents, where, among others, the user or visitor can interact in a multimodal way with an agent called Karin that knows about performances and performers in this theatre. The second system we discuss is a multimodal tutoring environment. In this environment an agent called Jacob is able to monitor the actions of a student that tries to solve the problem of the Towers of Hanoi. The third research environment is meant to experiment with multimodal tools that support the visitor to navigate in a virtual environment. How does a visitor make clear where she wants to go? How does the system explain where a visitor can go or what can be found in the environment? Finally, our last example is on a system that tries to recognize and interpret events in a room and interactions between occupants of this room.

What do these examples have in common? In research on multimodality speech and language often take primate ship. This certainly shows in these examples. As a consequence, multimodal interaction modeling often starts with attainments from speech and language research and other modalities are considered as add-ons. Our view on what is important in human-computer communication is determined by our view on the role of speech and language in human-human interaction. However, but this will not be discussed in this section, we would prefer to have an emphasis on human behavior 'modalities' (or characteristics) rather than on human communication modalities when considering the future role of computing and communication technology. The environments we discuss show the importance of information fusion, taking into account synchronization issues, they show the importance of cross-modal reference resolution and the importance of being able to reason beyond modality borders, and they show the importance of coordination in multimodal information presentation. In the examples there is no explicit discussion about syntactic, semantic and pragmatic level representation of information obtained from different media sources. In section 3 we will shortly address this question.

2.1 Multimodal Information Presentation and Transaction

The first environment we discuss is the Virtual Music Center. This is a virtual reality environment modeled on the real Music Hall in Enschede and is built using VRML (Virtual Reality Modeling Language). It is inhabited by a number of agents that can provide the user with information. Each agent has its specific task and domain knowledge. One of the agents, Karin, provides information about the agenda of performances and is able to sell tickets to visitors. Embedding the dialogue system in a visual context made it possible to present

information through other modalities. Because the agent became embodied, this made us look into aspects of non-verbal communication that could now be modeled as well.

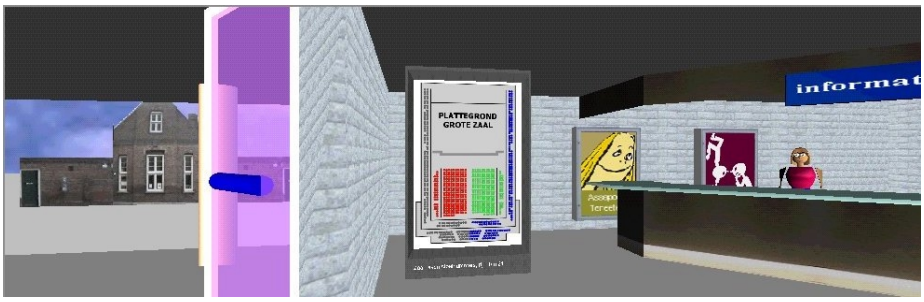

Figure 1: View from the entrance to the inside

In Figure 1 we see a screenshot taken from the entrance of the virtual theatre. With a virtual reality browser plugged into one of the usual web browsers, visitors can explore this virtual environment, walking from one location to another, looking at posters and other information sources and clicking on smart objects that set music and video tracks in motion. Behind the desk, the agent Karin is waiting to inform the visitor about performances, artists and available tickets. Karin can engage in a natural language dialogue with the visitor. She has access to an up-to-date database about all the performances in the various theaters. Karin has a cartoon-like 3D face that allows facial expressions and lip movements that are synchronized with a text-to-speech system to mouth the dialogue system's utterances to the user. Synchronization of sound and lip movements is currently implemented robustly because of the current limitations of web technology. Further design considerations that allow an embodied agent like Karin to display combinations of verbal and non-verbal behavior can be found in [Nijholt & Hulstijn, 2000].

All of this means that we have conversations and other forms of interaction between agents, between visitors, and between visitors and agents. Because communication is situated in a shared visible or otherwise observable virtual environment, communication partners are allowed to support their linguistic communicative acts by other means of interaction, referencing objects by gazing or pointing toward them, for instance. Using embodied

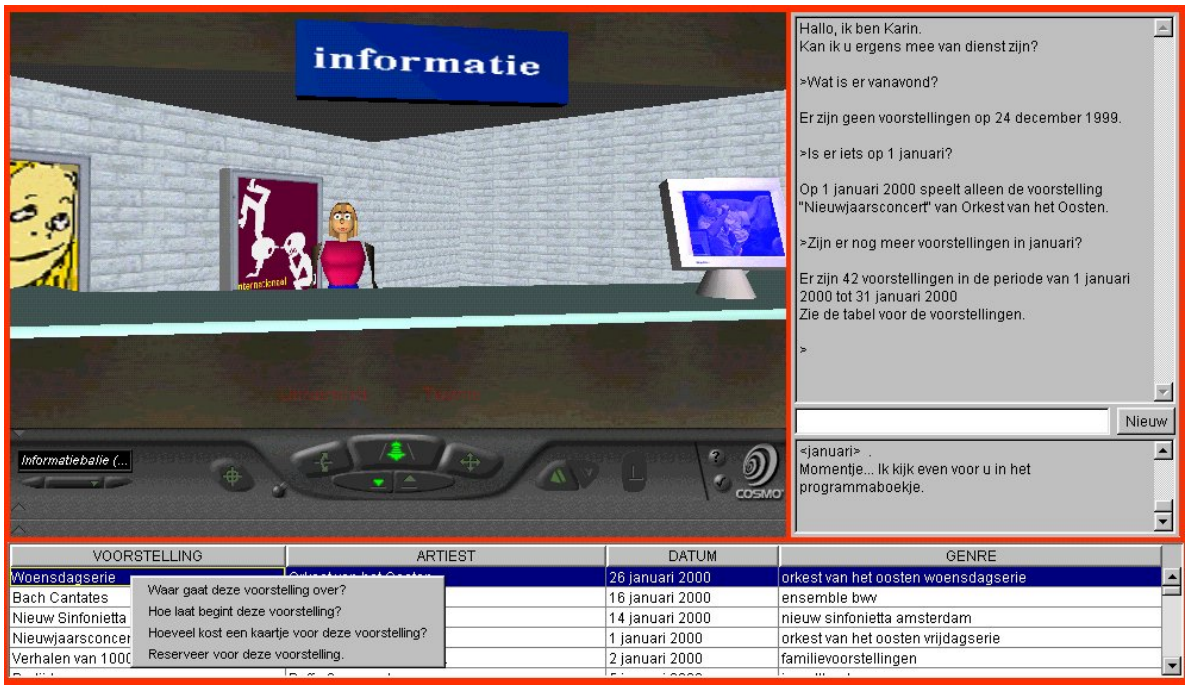

Figure 2: The dialogue windows for Karin

conversation agents as information sources in a virtual environment has allowed us to investigate a number of issues in natural and multimodal human-computer interactions.

Figure 2 shows a snapshot of a typical conversation between Karin and the user. The 3D environment with the typical buttons to navigate in 3D space shows Karin behind the information desk. It is just one frame in the window that also contains a dialogue frame in which the user types in questions and in which Karin's replies are shown. Note that Karin's answer does not just appear in the dialogue frame but that it is also read out loud using speech with proper lip-synchronization. The Dutch text-to-speech system Fluency was used for this.

A table containing information about performances takes up another part of the window. When the reply to an information request is a long list of performances, the system decides to present the reply in the table. Having Karin read out all the information would take too long. Therefore, Karin's reply will merely consist in directing the user to the information presented in the table. The user can click on the individual cells and menus will pop up with further information or with a suggestion of some frequently asked questions about performances (What is this performance about? What time does it start? Etc.). Karin can interpret and generate references to items in the table. A question like "Please give me more information about the third performance." will be understood correctly as making a reference to the third performance in the table.

The conversational agent Karin is essentially an embodied version of a spoken dialogue manager allowing natural language interactions with the user. The dialogue management system behind Karin is not very sophisticated from a linguistic point of view but reasonably intelligent from a practical and pragmatic point of view. The dialogues are assumed to be task-oriented dialogues about information and transaction. Karin knows information about performances only and expects the user to ask questions about these or to make a reservation for them. This assumption guides Karin's dialogues moves. The system prompts are designed in such a way that the users are gently adapted to Karin's way of thinking.

Linguistic analysis is performed by a rewriting system. Input sentences are processed by a series of transducers that rewrite the input form to a canonical form that can be interpreted and acted upon by the dialogue manager. Although the linguistic analysis is crude, the restrictions on the kind of information that can be queried and on the types of dialogues that are normally undertaken allow the system to produce adequate responses in the majority of cases assuming reasonable user behavior. The original specification for the dialogue system was based on a corpus of dialogues obtained by a Wizard of Oz experiment. This corpus provided insight in the kind of utterances and dialogues visitors would engage in.

The rewriting system that implements the natural language analysis robustly had important benefits in the initial development of the virtual environment in that it allowed us to build a reliable and fast tailor-made system without having to spend much time developing it. However, its simplicity also has its drawbacks. In the rewriting step, information is lost that is needed for the naturalness of the dialogues. For instance, the system may forget the exact phrases uttered by the visitor and use its own wording instead. This may give the impression that the system is correcting the user; an effect that was obviously not intended.

2.2 Multimodal Tutoring with Jacob

Our second environment concerns a virtual teacher (Jacob) that guides a student in solving the problem of the Towers of Hanoi [Evers & Nijholt, 2000] (see also Figure 3), a classic toy problem from the field of programming language teaching. In this game, the student has to move a stack of blocks from one peg to another using an auxiliary peg. The student can only

move single blocks from peg to peg; it is not allowed to place a larger block on top of a smaller one. During the moving of blocks Jacobs monitors the student, he can correct the student and when asked demonstrate a few next steps. Obviously, a more generic way to look at this application is to consider it as an environment where tasks have to be performed that involve moving objects, pressing buttons, pulling levers, etcetera, and where an intelligent agent provides assistance.

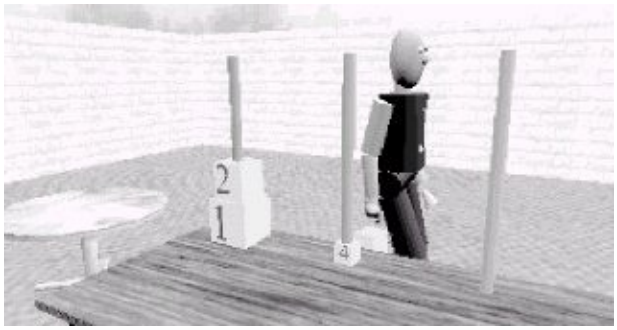

Figure 3: Jacob demonstrates the Towers of Hanoi

Before looking at the interaction with Jacob we shortly discuss the (generic) architecture of the system. The software architecture is layered in order to separate the concerns of the 3D visualization from the basic functionality of the system. The abstract 3D world contains objects representing e.g. blocks, pegs, and Jacob's physical manifestation. An avatar object represents a user. We have defined an interface layer between the abstract and concrete 3D world layers to make the system more robust to changes. The concrete 3D world layer consists of a hierarchical structure of VRML nodes, like transformation nodes, geometry nodes, and sensor nodes. The interface layer exposes only the essential properties of the nodes in the concrete 3D world, like the position of an object. All other details are hidden. The abstract 3D world layer provides simulation of physical properties. For the Towers of Hanoi, we have implemented basic collision avoidance so that objects cannot be moved through each other. Furthermore, we have implemented a simple gravity variant that makes objects fall at a constant speed. A task model and an instruction model together form Jacob's mind. They act as controllers that observe the abstract world and try to reach specific instruction objectives by manipulating Jacob's body.

The interaction between the user and Jacob is multimodal. An important issue in this project was how natural language dialogue and nonverbal actions are to be integrated and what knowledge of the virtual environment is needed for that purpose. The Jacob agent behaves in an intelligent way, helping the user proactively and learning from the interaction. Moreover, visualization of Jacob plays an important role, including natural animation of the body, generation of facial expressions, and synchronization of lip movement and speech. Unfortunately, apart from some simple eyebrow movements, this has not been implemented. Both the user and Jacob are however able to manipulate objects in the virtual environment. The student can use mouse and keyboard to manipulate objects. A data glove could have been used as well, without making changes to the architecture.

Figure 4: Student interacting with Jacob

The interaction between Jacob and student uses keyboard natural language and animations by Jacob. A sample dialogue between Student (S) and Jacob (J) is:

S: "What should I do?"

J: "The green block should be moved from the left to the right peg. Do you

want me to demonstrate?"

S: "Yes, please."

Jacob moves the green block

S: "Which block should I move now?"

J: "I will show you."

Jacob picks up the red block

The student accepts the red block and puts it on the right peg

J: "That was good. Please continue!"

As becomes clear from this sample dialogue, Jacob knows about the current situation, making it possible to react in an appropriate way on user utterances that deal with the current situation. In this system no attention has been paid to references made by the user to events or parts of the dialogue history that go back a few steps in time. The system has very limited intelligence and only observes the current situation. The system displays information by giving verbal responses to the student and by animating the Jacob agent and having him perform movements from blocks from one peg to the other, shaking his head to show his disagreement or, when the user is really stupid, walking away. The user can use speech input to communicate with Jacob (see Figure 4), but he cannot make detailed references to the objects in the virtual world or about the actions that are being performed, either by the student or by Jacob. Blocks can be moved using the mouse or the keyboard, not by using speech input.

2.3 Multimodal Navigation in 2D and 3D Intelligent Environments

In section 2.1 we met Karin. She knows how to access the theater database and to generate answers concerning performers and performances. Our world has been made accessible to the audience. More generally, as visitors of an unknown world, be it a physical space, a public space on the web, a smart room or home environment, or a 3D virtual world, we often need other types of information as well. To whom do we address our questions about the environment and its functions? To whom do we address our questions about how to continue when performing a certain task, where to find other visitors or where to find domain or task related information?

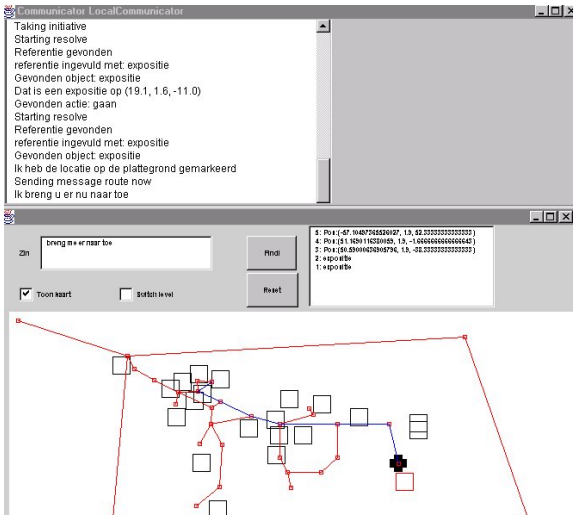

Figure 5: 2D Floor map (ground floor) of the VMC

At this moment we are following different approaches to solve this problem. These approaches are related and can be integrated since all of them are agent-oriented and based on a common framework of communicating agents. In addition, we have built this framework in such a way that different agents with different abilities can become part of it: Karin who knows about theater performances, tutors that monitor the student's actions (a virtual piano teacher [Broersen & Nijholt, 2002], Jacob that follows the student's progress while solving the Towers of Hanoi [Evers & Nijholt, 2000], a teacher that presents modules in nurse education [Hospers et al., 2003] and navigation agents that have knowledge

about a particular environment and that allows users to make references to a visualized environment while interacting with the agent.

Developing a navigation agent poses a number of questions. How can we build navigation intelligence into an agent? What does navigation intelligence mean? How can we connect this intelligence to language and vision intelligence? We know at least that visitors of an environment are language users and recognize and interpret what they see. There is a continuous interaction between verbal and nonverbal information when interpreting a situation in our virtual environment. Modeling this interaction and the representation and interpretation of these sources, together with generating multimedia information from these sources is a main issue in navigation research. For observations on the preferences that users have for navigation support, the reader is referred to [Darken & Silbert, 1996; Höök et al., 1988].

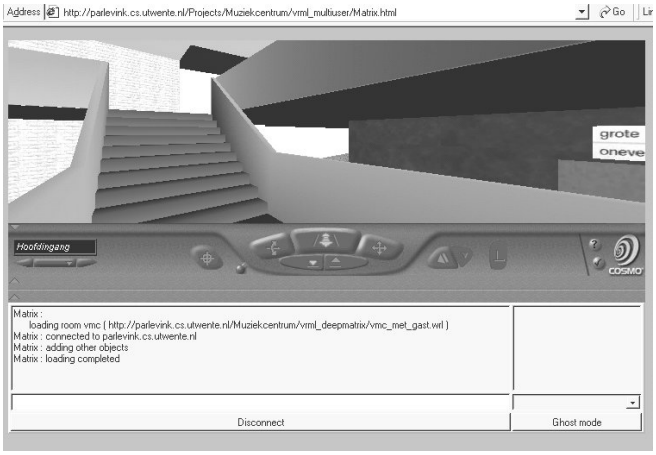

Figure 6: The visitor has been brought to the stairs

As a case study we introduced a 2D map and a multimodal dialogue agent to our VMC laboratory environment. A first version appeared in [van Luin et al., 2001]. The work is in progress, meaning that the system is there, but no effort has been made to add 'graphic sugar' to the layout and the integration of the different windows that are used.

In Figure 5 we display a floor map of the VMC. In Figure 6 a view on part of the virtual world is presented. In this view the user has been brought to the stairs leading to the next floor of the building. The visitor can ask questions, give commands and provide information when prompted by the agent. This is done by typing natural language utterances (in a more recent version we have added speech recognition [Hofs et al., 2003]) and by moving the mouse pointer over the map to locations and objects the user is interested in. On the map the user can find the performance halls, the lounges and bars, selling points, information desks and other interesting locations and objects. The current position of the visitor in the virtual environment is marked on the map. While moving in virtual reality the visitor can check her position on this floor map. When using the mouse to point at a position on the map, references can be made by both user (in natural language) and system to the object or location pointed at [van Luin et al., 2001].

As we mentioned, the navigation agent can be accessed using natural language. We have annotated a small corpus of user utterances that appear in navigation dialogues. On the one hand we find complete questions and commands. On the other hand we also have short phrases that are given by the user in reply to a clarifying question of the navigation agent. An example of a question is: "What is this?" while pointing at an object on the map, or "Is there an entrance for wheel chairs?" Examples of commands are "Bring me there." or "Bring me to the information desk." Examples of short phrases are "No, that one." or "Karin." From the annotated corpus a grammar was induced and a unification-type parser for Dutch was used to parse these utterances into feature structures.

Three agents communicate to fill in missing information in the feature structure (when the information given by the user in her question, answer or command is not yet complete) and to

determine the action that has to be undertaken (answering the question, prompting for clarification or missing information, displaying a route on the map or guiding the user in virtual reality to a certain position). The navigation agent, the dialogue manager agent and the browser agent do this in co-operation. The latter can communicate with the virtual reality browser (embedded in a standard web browser) to retrieve the current position of the visitor and to have a smooth transition from the user's viewpoint along a route that has been computed. Not yet implemented is the possibility that not only the position but also what is in the eyesight of the visitor can be retrieved. This will allow better resolution of references in the dialogue to objects that are visible for the visitor in the virtual environment.

An illustrative example of a navigation dialogue is given below:

Visitor:	[Clicks on an object on the map] What is this?
Agent:	That is an exposition.
Visitor:	Where is it?
Agent:	You can find it in the lounge.
Visitor:	Let's go there.
Agent:	I bring you there.

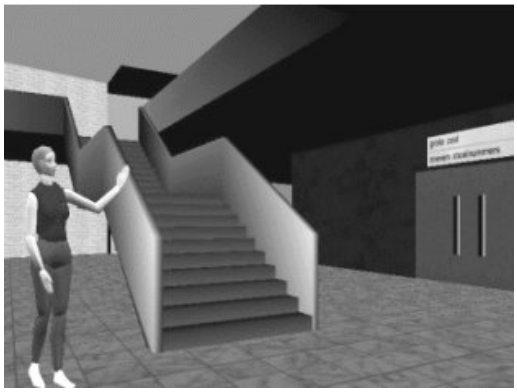

Figure 7: You go UP the STAIRS

The prototype navigation agent that we discussed here is certainly not our final solution to assisting visitors of our smart environment. There are several issues we are working on right now. The first issue concerns speech recognition. A web-based speech architecture has been introduced [Hofs et al., 2003]. The second issue concerns the modeling of speech dialogues. We introduced an approach to multimodal dialogue modeling for navigation that emphasized – and tried to find solutions – for cross-modal reference resolution and for sub dialogue modeling. In a next phase, we need to concentrate on the communication with other agents that are available in the environment. How can we ensure that a question reaches the appropriate agent? How can we model the history of interaction in such a way that different agents do not only know about their own role in this interaction but also about others? Unlike others, our environment allows the investigation of communication between active and passive agents that inform the visitor about the possibilities and the properties of an information-rich virtual environment (see also [Nijholt et al., 2003]). Embodiment of the navigation agent, as illustrated in Figure 7, adds an other dimension to navigation in a virtual, visualized, environment.

2.4 Multimodal Actions and Interactions in Smart Environments

While in the previous examples we discussed interactions between users and visitors of environments, where the environment was represented on the screen of a PC as a menu-based graphical user interface, a 2D or 3D world or a virtual reality environment, we can also look at the interaction with or taking place in physical environments, where these environments are equipped with sensors that capture audio- and visual information. For example, in the 5th framework IST project Multi-Modal Meeting Manager (M4) we are involved with research on the semantics of the interactions and the events during a meeting. Obviously, these events and interactions are of multimodal nature. Apart from the verbal and nonverbal interaction between participants, many events take place that are relevant for the interaction between

participants and that therefore have impact on their communication content and form. For example, someone enters the meeting room, someone distributes a paper, the chairman opens or closes the meeting, ends a discussion or asks for a vote, a participants asks or is invited to present ideas on the whiteboard, a PowerPoint presentation is given with the help of laser pointing and later discussed, someone has to leave early and the order of the agenda is changed, etc. Participants make references in their utterances to what is happening, to presentations that have been shown, to behavior of other participants, etc. They look at each other, to the person they address, to the others, to the chairman, to their notes and to the presentation on the screen, etc. Participants have and use facial expressions, gestures and body posture that display or emphasize their opinion, etc.

The aim of the M4 project is to design a meeting manager that is able to translate the information that is captured from microphones and cameras into annotated meeting minutes. In fact, but this is certainly too ambitious for the current project, it should be possible to generate everything that has been going on during a particular meeting from these annotated meeting minutes, for example, in a virtual meeting room, with virtual representations of the participants. The more modest goals of the M4 project include the summarization of a meeting and the retrieval of multimedia information from these annotated meeting minutes. In future projects it is expected to tackle the more ambitious goals that deal with virtual reality generation.

Clearly, we can look at the project as research on smart environments or on ambient intelligence. While in the previous sections we looked at multimodality in the context of one particular user that communicates with a computer screen, here we have a situation where there is no explicit or active communication between user and environment. The environment registers and interprets what's going on, but is not actively involved. The environment is attentive, but does not give feedback or is pro-active with respect of the users of the environment or the participants of a collaborative event. Real-time participation of the environment requires not only attention and interpretation, but also intelligent feedback and pro-active behavior of the environment. It requires also presentation by the environment of multimedia information to the occupants of the environment.

Figure 8: Three camera's capturing a M4 meeting

In order to collect multimodal meeting information scripted meetings have been organized in which participants act according to prescribed rules that define periods of monologue, discussion, note taking, or a whiteboard presentation. The corpus thus obtained allows study of meeting participants' behavior. In Figure 8 we show a three-

camera view of a meeting between four persons. In addition to the cameras there are lapel microphones and circular microphone arrays available for the meeting manager to capture audio. In the near future it is expected that white board pen capture can be added.

On a more detailed level the objectives of the project are the collection and annotation of a multimodal meetings database, the analysis and processing of the audio and video streams, robust conversational speech recognition, to produce a word-level description, recognition of gestures and actions, multimodal identification of intent and emotion, multimodal person identification and source localization and tracking. Models are needed for the integration of the multimodal streams in order to be able to interpret events and interactions. These models include statistical models to integrate asynchronous multiple streams and semantic representation formalisms that allow reasoning and cross-modal reference resolution. These models form the basis of browsing, retrieval, extraction and summarization methods. Textual "side information" (the agenda, discussion papers, slides) enables the application of useful constraints. It may be used to adapt the language model of the speech recognizer or as query expansion information for retrieval.

A straightforward meeting browser can follow the structure of an agenda. Each agenda item can be associated with different views on that topic. For example, a textual summary, a diagrammatic discussion flow indicating which participants were involved (speaker turn patterns), and audio and video key frames that give the essence of the discussion. Obviously, in order to track the discussion and find the interesting parts features need to be distinguished that can be recognized by the meeting manager.

Presently there are two approaches that are followed. The first one is the recognition of joint behavior, that is, the recognition of group actions during the meeting. Examples of group actions are presentations, discussions, consensus and note taking. Probabilistic methods based on Hidden Markov Models (HMMs) are used for this purpose [McCowan et al., 2003]. The second approach is the recognition of the actions of the individuals independently, and fuse them at a higher level for further recognition and interpretation of the interactions. When looking at the actions of the individuals during a meeting several useful pieces of information can be collected. First of all, there can be person identification using face recognition. Current speaker recognition using multimodal information (e.g., speech and gestures) and speaker tracking (e.g., while the speaker rises from his chair and walks to the whiteboard) are similar issues. Other, more detailed but nevertheless relevant meeting acts can be distinguished. In [Zobl et al., 2003] recognition of individual meeting actions by video sequence processing in the context of the M4 project is discussed. Examples of actions that are distinguished are entering, leaving, rising, sitting, shaking head, nodding, voting (raising hand) and pointing (see Figure 9). These are rather simple actions and clearly they need to be given an interpretation in the context of the meeting. Or rather, these actions need to be interpreted as part of other actions and verbal and nonverbal interactions between participants. Presently

Figure 9: Pointing, rising and voting

models, annotation tools and mark-up languages are being developed in the project that allow the description of the relevant issues during a meeting, including temporal aspects and including some low-level fusion of media streams. Higher-level fusion, where also semantic modeling of verbal and nonverbal utterances is taken into account has not been done yet. In some cases it turns out to be more convenient to make shortcuts to a pragmatic level of fusion using knowledge from the application.

The M4 meeting manager captures the events and interactions in the meeting room. After capturing the gathered information becomes available for both participants and non-participants. A next step is to allow remote participants to take part in the meeting and integrate their interactions as well. Making the meeting room intelligent and providing real-time support for efficient and effective interactions is an other objective. Clearly, as mentioned above, in this way we enter the research area of smart multi-party environments and intelligent collaborative workspaces [Mikic et al., 2000; Potter, 2003]. In all these environments we need to model the fusion of multi-sensory information in such a way that spatial and temporal aspects have to be taken into account.

3. Multimodality: Issues and Roadmaps

There is no uniformity in terminology on modes and media, on multimodality and multimedia. In [Maybury & Wahlster, 1988] media, modes and codes are distinguished. Examples of media are text, audio and video. Modes refer to the human perceptual system. For example, visual auditory or tactile modalities. Formalization is done in syntactical, semantical and pragmatic languages. These are called the code, allowing us to speak of a multimodal code. Others prefer to speak of multimodal input where every separate information stream or channel is an input modality for the system and where every information stream can have a syntactic, semantic and pragmatic representation. However, in order to obtain multimodal interpretation these input modalities have to be fused. This can be done, again, one a syntactic, semantic and pragmatic level, but it seems to be more appropriate to have partial fusion at these levels in order to allow mutual disambiguation and cross modal reference resolution.

At a low level we encounter the ‘wait’ problem. Input on multiple channels is not always strictly synchronized. How long does the system have to wait for input in other channels before it triggers an action? For example, if gestural input is available, it may be a gesture-only utterance, or part of a multimodal utterance. In the first case action should be immediately taken, in the latter the system should wait for the following natural language utterance. This is related to the distinction between early and late integration.

Early integration takes part at the data-level, late integration at the semantic level. In an early fusion architecture, the recognition process in one mode influences the course of recognition in the other. The advantage is an intensive cooperation between the various modalities, which may result in a more accurate interpretation. An example where this can be useful is with synchronized modalities such as speech and lip movements.

Late integration enables the reuse of recognition modules for individual modalities [Oviat et al., 2000]. Moreover, at the pragmatic level it may be decided how detailed the fusion at lower levels need to be for a particular multimodal domain and application.

Methods for integration include unification-based integration [Johnston et al., 1997], the melting pot approach [Nigay & Coutaz, 1995], HMM and finite state methods (e.g., [Johnston & Bangalore, 2000]), constructive type theory [Bunt & Beun, 1998] and extensions

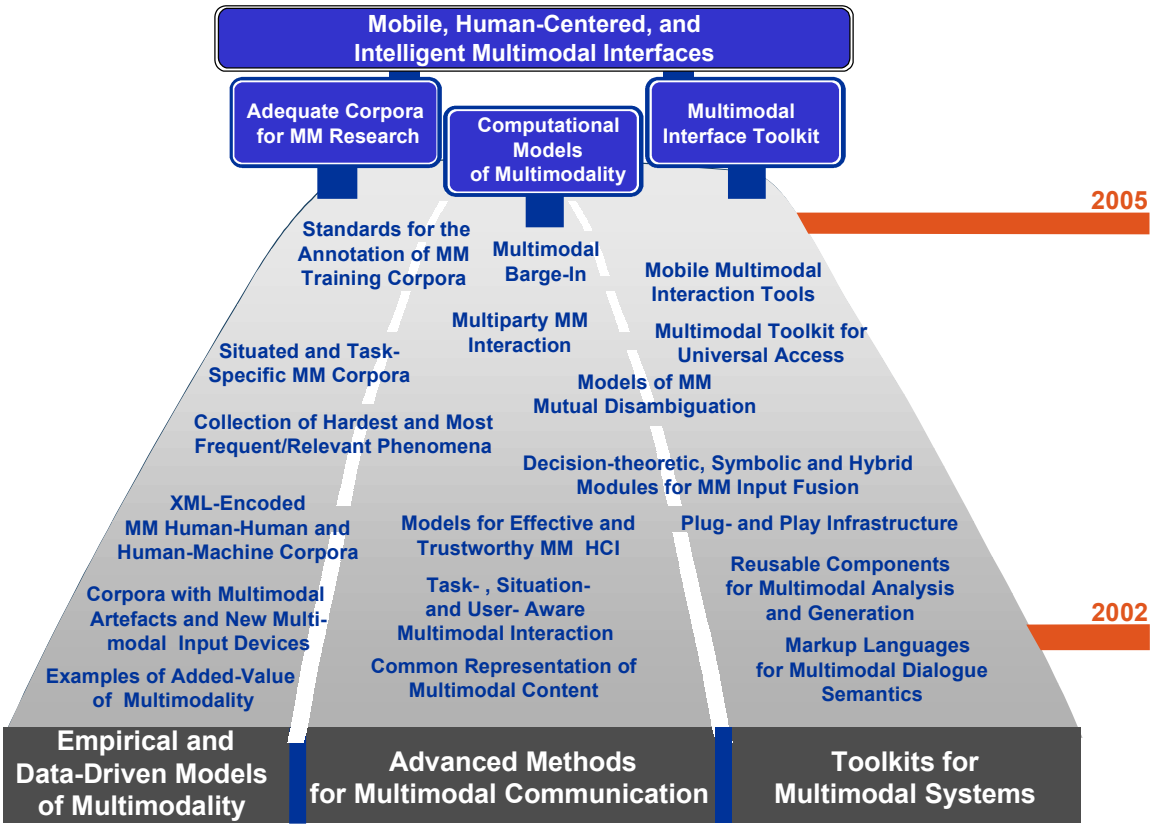

Figure 10: Near-term roadmap for multimodal communication

of discourse representation theory [Bunt et al., 2003]. Finally, multimodal output generation requires content selection, media allocation and rendering of output on presentation mechanisms and display devices.

3.1 Roadmaps

The aim of this section is to introduce the reader to some recent roadmaps that have been designed for multimodal communication {Bunt et al., 2003]. The roadmaps are about the integration and synchronization of input modes during interaction and the coordination of the multimodal output. The three lanes in Figure 10 distinguish the collecting of multimodal corpora, including coding schemes and annotation tools, the computational modeling of multimodality, including modeling of multimodal syntax, semantics and pragmatics, and, as a third lane, the development of toolkits for input fusion and output coordination. Figure 10 is the roadmap in the near-term. In Figure 11 the long-term roadmap is illustrated. Here we have more attention paid to multimodal environments, collaboration, multi-users, usability and user modeling and affective computing. As may become clear from the figures the staring point of the foreseen research is again the individual user that interacts in a multimodal way with a computer system. This is reflected in the choice of theories: from speech act to dialogue act to multimodal dialogue act, rather than starting with multimodal acts. This can be contrasted with the current M4 project (see section 2.4) where domain specific multimodal acts (meeting acts) are defined and where syntactic modeling, using for example HMMs, and semantic modeling is done to integrate modes into multimodal, domain-specific, meeting acts.

.

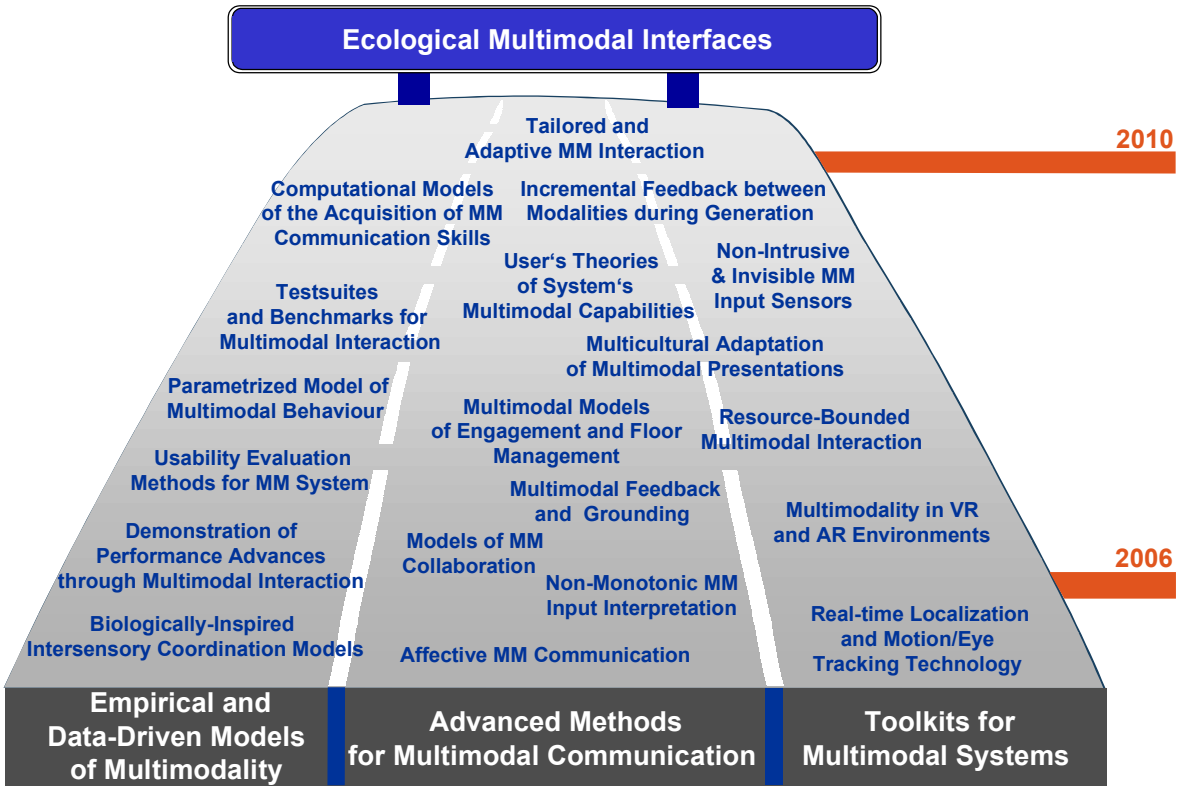

Figure 11: Long-term roadmap for multimodal communication

4. Disappearing Computers, Disappearing Social Actors, Embodied Agents

In the previous sections we discussed examples of systems that allow multimodal interactions. Many existing research and prototype systems introduced embodied agents, assuming that they allow a more natural conversation or dialogue between user and computer. In this section we will first take a look at how in general people react to computers. We will look at some of the theories, in particular the CASA ("Computers Are Social Actors") paradigm, and then discuss how new technology, for example ambient intelligence technology, needs to anticipate the need of humans to build up social relationships. One way to anticipate is to do research in the area of social psychology, to translate findings there to the human-computer situation and to investigate technological possibilities to include human-human communication characteristics in the interface. For that reason we will discuss embodied conversational agents, the role they can play in human-computer interaction (in face-to-face conversation), in ambient intelligence environments and in virtual communities.

4.1 Computers are Social Actors

In the "Media Equation" (Figure 12) Byron Reeves and Clifford Nass report about their experiments on human-computer interaction where humans assign human characteristics to computers [Reeves & Nass, 1996]. Many experiments have been done after this book has been published. They became known as the "social reactions to communication technology" (SRCT) perspective in which "computers are social actors". An example of an experiment is the following. A student is asked to sit behind a computer and to perform a particular task. When finished, the student needs to answer questions: how helpful was the computer, was it

friendly, was it polite, etc. Two computers were available for answering these questions: the computer that was used for performing the task and another computer just for presenting the questionnaire and having the student answer it. It turned out that when the questionnaire had to be answered on the computer that had been used to communicate the task with the student and to help the student when performing this task, students answered much more positive and politely than when answering similar questions posed by the second computer. Clearly, people don't like to offend a computer that has tried to be helpful to them.

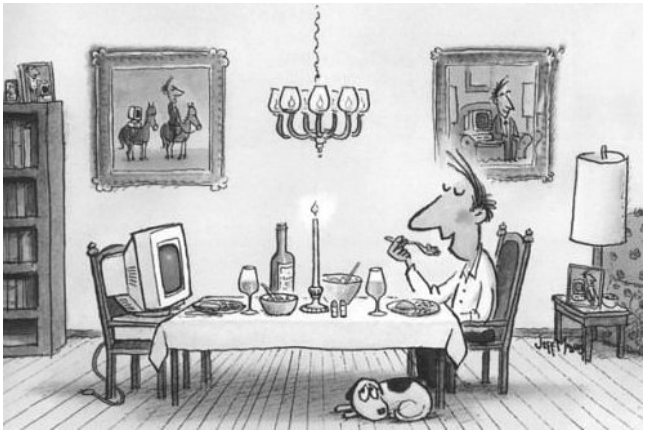

Figure 12: The media equation

Many similar experiments have been performed. Computer users turned out to be sensitive for flattery and humor; moreover, they were very much influenced, when assigning personality characteristics to a computer, by the properties of the synthesized voice in text-to-speech synthesis. And, as became clear from the experiments, it is not just a matter of contributing personality characteristics to computer, it is also a matter of being influenced by these properties while communicating with the computer. Hence, the book's conclusion was as follows:

*"Our strategy for learning about media was to go to the social science section of the library, find theories and experiments about human-**human** interaction - and then borrow. We did the same for information about how people respond to the natural environment, borrowing freely. Take out a pen, cross out "human" or "environment," and substitute **media**. When we did this, all of the predictions and experiments led to the media equation: People's responses to media are fundamental social and natural."*

For a future situation where a house, a sitting room, a working room, an office and in fact every environment and its objects allow perception of what is going on in the environment and allow interaction by its occupants and visitors to exchange information (including emotions), it is certainly useful to investigate how we can design social interfaces, emphasizing human-to-human communication properties, rather than concentrating purely on designing intelligence. One important aspect in the design is the appearance of the interface. When offering intelligence and emotion, shouldn't we offer virtual humans (or embodied conversational agents) in the interface? They offer communication properties that make us feel being appreciated and that make us feel being understood. It makes it possible for us to act in a smart, but also in a social environment.

4.2 Social Actors, Interpersonal Relationships and the Disappearing Computer

In the previous section we introduced the computer as a social actor. In human-computer interaction we recognize characteristics of human-human interaction. There is human-like behavior when interacting with the computer and human-like behavior of the computer is expected. Can we expect similar behavior when the user is interacting with an environment rather than with a desktop screen? In future environments computers will be embedded in walls, furniture, cloths, and in objects that are natural in the environment. Moreover, there is communication between these embedded computational devices allowing a much more comprehensive overview of environment and events taking place than is possible with a single

computing device. How will humans interact with such environments? Are they able to build some kind of relationship with these environments like they are able to build relationships with a computer that is perceived as a social actor? Or do we need to introduce explicit social actors, that is, embodied conversational agents, in these environments with which users can communicate and exchange information in intelligent and social ways in order to fulfill a need to establish relationships with their environments?

Some notes are in order. Firstly, it is not unusual to contribute personality characteristics to a room, a house, a mall, a street or square, to a town or even to a landscape or an other natural environment. At least one may think that thoughts and activities (i.e., interactions with the environment) are influenced by the particular environment. We won't go into that here. Secondly, it is useful to distinguish between situations. Different circumstances require different kinds of interactions. Sometimes we want to see things arranged in an efficient way. Sometimes we are more concerned with the partner's satisfaction when arranging things. Sometimes arranging itself is entertaining. Both interaction and information exchange can be goals, e.g., when we enter conversations with our children or colleagues. Efficiency has not always priority. A third note, as mentioned above, concerns the future. It is already the case that a large part of the professional population in Western countries spends the day with discussion, meetings and knowledge exchange and spends lots of time interacting with computers. The need to do this in the office will decrease and home, work and mobile situations will more and more resemble each other. Interaction forms require mixtures of efficiency, social relationship, and entertaining aspects. Our hypothesis is that people prefer to be able to interact with their 'own', personalized (but not only in the current technical sense, i.e., aimed at efficiency) and non-anonymous environment.

Although the SRCT perspective makes us aware that people react socially to computers, a more detailed view can make clear many nuances. To start with, there is no such thing as *the* computer. Its performance, as it shows in the interface, can be task oriented, it can be communication oriented and it can be oriented towards establishing and maintaining relationships. In Interpersonal Theory these types are the three tracks of conversational goals [Shechtman & Horowitz, 2003]. The task goal in human-to-human conversation is why the conversation is started, i.e., to accomplish a certain task and part of the interaction behavior is meant to reach this goal. The communication goals aim at making the interaction process run, e.g., by allowing smooth turn taking. The relationship goals of the conversational partners set the tone of the conversation. Two broad categories of relationship goals are distinguished: communion (behaviors oriented towards connecting with one another or disconnecting from another) and agency (behaviors oriented toward exerting influence or yielding to influence). Shechtman conducted experiments to study relationship behavior during keyboard human-computer interaction and (apparently) keyboard mediated human-human interaction. In the latter case participants used much more communion and agency relationship statements, used more words and spent more time in conversation.

Not all modalities that can be employed in human-computer interaction lend themselves to the same degree to the different types of performance that we distinguished above. In human-human interaction nonverbal cues play an important part in the relationship track of communication. Hence, we can ask whether we can recognize and interpret these communication aspects in human-computer interaction and whether they can play a similar role. From the SRCT perspective we know that humans react socially on social computer behavior and having the computer display more cues about its social behavior may strengthen the social reaction. Obviously, there will not necessarily be a need to consider your own computer, let alone, every computer, as a personal friend with whom you want to share your feelings. Nevertheless, there will be many situations, especially in a personal environment,

where people will prefer communicating with systems that show knowledge of the user and display reactive and pro-active behavior that shows understanding of the particular context of the user, including its mood and emotions. To do this we need other modalities in interaction and presenting information than just menu-based graphical user interfaces.

5 Embodied Conversational Agents and Multimodality

In the previous sections we saw some examples of embodied conversational agents or virtual humans. Embodied conversational agents (ECAs) have become a well-established research area. Embodied agents are agents that are visible in the interface as animated cartoon characters or animated objects resembling human beings. Sometimes they just consist of an animated talking face, displaying facial expressions and, when using speech synthesis, having lip synchronization. These agents are used to inform and explain or even to demonstrate products or sequences of activities in educational, e-commerce or entertainment settings. We saw examples, Karin knowing how to inform theatre visitors and Jacob assisting students. Experiments have shown that ECAs can increase the motivation of a student or a user interacting with the system. Lester et al. [Lester et al., 1997] showed that a display of involvement by an embodied conversational agent motivates a student in doing (and continuing) his or her learning task. Some examples of embodied conversational agents are shown in Figure 13. From left to right we see: Jennifer James, a car saleswoman who attempts to build relationships of affection, trust and loyalty with her customers, Karin, informing about theatre performances and selling tickets, Steve, educating a student about maintaining complex machinery, and Linda, a learning guide.

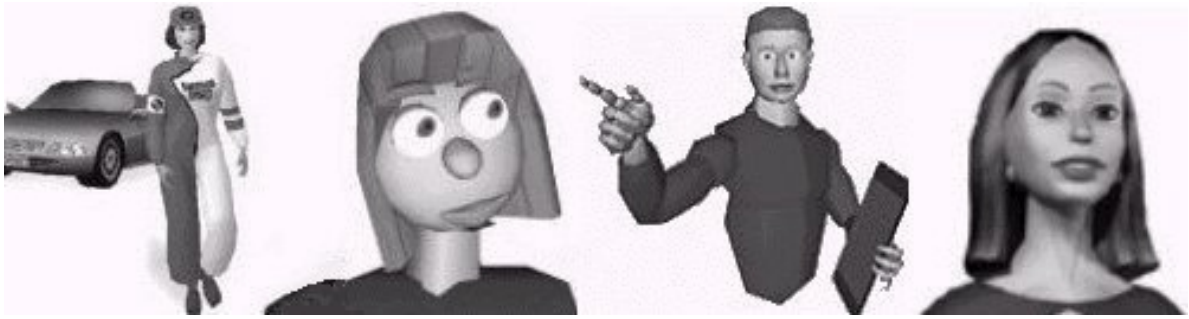

Figure 13: Examples of 2D and 3D embodied agents

In this section we will discuss the intelligence and nonverbal interaction of embodied conversational agents. We will look at the role of gestures and the role of gaze. Displaying emotions, in particular through facial expressions is another issue that will be discussed. Finally, we have a few words about ECA design that allows the development of interpersonal relationships.

5.1 Intelligence and Nonverbal Interaction

Embodiment allows more multimodality, therefore making interaction more natural and robust. Several authors have investigated nonverbal behavior among humans and the role and use of nonverbal behavior to support human-computer interaction. See e.g. [Cassell et al., 2000] for a collection of chapters on properties and impact of embodied conversational agents (with an emphasis on coherent facial expressions, gestures, intonation, posture and gaze in communication) and for the role of embodiment (and small talk) on fostering self-disclosure and trust building.

Current ECA research deals with improving intelligent behavior of these agents, but also with improving their verbal and nonverbal interaction capabilities. Improving intelligent behavior requires using techniques from artificial intelligence, in particular natural language processing. Domain knowledge and reasoning capabilities have to be modeled. Agent models have been developed that allow separation between the beliefs, desires and intentions of an agent. Together with dialogue modeling techniques rudimentary natural language interaction with such agents is becoming possible.

What role do gestures play in communication and why should we include them in an agent's interaction capability? Categories of gestures have been distinguished. Well known is a distinction in consciously produced gestures (emblematic and propositional gestures) and the spontaneous, unplanned gestures (iconic, metaphoric, deictic and beat gestures). Gestures convey meanings and are primarily found in association with spoken language. Different views exist on the role of gestures in communication. Are they for the benefit of the gesturer or for the listener? Gestures convey extra information [Kendon, 1980] about the internal mental processes of the speaker: "*. . . an alternative manifestation of the process by which ideas are encoded into patterns of behavior which can be apprehended by others as reportive of ideas.*" Observations show that natural gestures are related tot the information structure (e.g., the topic-focus distinction) and (therefore) the prosody of the spoken utterance. In addition they are related to the discourse structure and therefore also to the regulation of interaction (the turn taking process) in a dialogue**.** Apart from these viewpoints on embodiment, we can also emphasize the possibility of an embodied agent to walk around, to point at objects in a visualized domain, to manipulate objects or to change a visualized (virtual) environment. In these cases the embodiment can provide a point of the focus for interaction. When, for example, we introduce a guide in our virtual environments this is a main issue and more important than detailed facial expressions and the gestures discussed above.

5.2 The Role of Gaze

For believability and naturalness embodied conversational agents should also display life-like forms of non-verbal communication. In our environment, the face of Karen was designed to make control over its features possible. In the version of Karen that is currently accessible on the web, some changes in facial expressions are hard-coded to accompany some fixed elements in the interaction. For instance, Karin will look down at the table with performances when pointing out that more information can be found there. Currently a lot of research is going on that is specifically aimed at improving the non-verbal communication skills of Karin and other agents. This not only includes the use of facial expressions but also of gaze, gestures and posture.

Getting a system that behaves naturally in this respect involves tight co-ordination of the facial animation driver with many parameters of the dialogue manager, with the mental state of the character and its model of the user and subtle aspects of the linguistic utterance that is produced or attended to. Consider in this respect the functioning of gaze in human-human conversations [Argyle & Cook, 1976; Torres et al., 1997]. Gazing away from or towards the interlocutor can function as an important emotional signal as well as a signal to hand over the turn or avoid the turn to be taken over. As a function in the organization of turn-taking behavior, the timing of mutual gaze (eye-contact) correlates with the information-structure of the utterances (its topic-focus articulation).

In an experiment, we investigated the effects of different styles of gaze of Karin on the conversation. We had forty-eight subjects each make two reservations with different style versions. We videotaped the conversations, clocked the time they spent on the task, and had

them fill in a questionnaire after they had made the reservations. It appeared that participants that had conversed with a version in which common gaze behavior was implemented (looking away and towards users and beginnings and ends of turns, respectively) appreciated their conversation significantly better than the other participants in most respects. They not only were more satisfied overall, they found it easier to use than a version with the minimal amount of eye-movements, appreciated the personality of the agent better and thought the head movements were more natural. They were also the fastest, on average, to complete the task. For more details and more results we refer to [Heylen et al., 2003]. This short summary already shows the important effects that can be achieved by paying attention to the non-verbal language signals of embodied conversational agents.

5.3 Emotional Behavior, Personality, Friendship

Facial expressions and speech are the main modalities to express nonverbal emotion. Human beings do not express emotions using facial expressions and speech only. Generally they have their emotions displayed using a combination of modalities that interact with each other. We cannot consider one modality in isolation. Facial expressions are combined with speech. There are not only audio or visual stimuli, but also audio-visual stimuli when expressing emotions. A smile gesture will change voice quality, variations in speech intensity will change facial expression, etc. Attitude, mood and personality are other factors that make interpretation and generation of emotional expressions even less straightforward. In addition we can have different intensities of emotion and the blending of different emotions in an emotional expression. We should consider combinations and integration of speech, facial expressions, gestures, postures and bodily actions. It should be understood that these are displays and that they should follow from some emotional state that has been computed from sensory inputs of a human interactant, but also from an appraisal of the events that happen or have happened simultaneously or recently. A usual standpoint is that of appraisal theory, the evaluation of situations and categorizing arising affective states. It should be understood that what exactly is said and what exactly is done in a social and emotional setting is not part of the observations above. The importance of the meaning of words, phrases and sentences, uttered and to be interpreted in a specific context is not to be diminished. In Figure 14 we display Cyberella, an embodied agent, developed at DFKI in Saarbrücken. This agent is working as a receptionist. For example, she can provide directions to the office of a staff member. However, since she has been provided with an affective model, she also reacts emotionally to a visitor's utterances when appropriate [Gebhard, 2001].

Figure 14: Cyberella, a virtual receptionist

One of the issues we investigated was how aspects of personal attraction or friendship development [Stronks et al., 2002] can be made part of the design of an embodied agent that is meant to provide an information service to a human partner. As a 'lay psychologist', we all know that people that you like (or your friends) are able to help you better, teach you better, and generally are more fun to interact with, than people that you don't like. However, 'liking' is person dependent. Not everybody likes the same person, and one person is not liked by everyone. These observations sparked our interest in the application, effects, and design of a 'virtual friend'. An agent that observes it's user, and adapts it's personality, appearance and behavior according to the (implicit) likes and dislikes of the user, in order to 'become friends' with the user and create an affective interpersonal

relationship. This agent might have additional benefits over a 'normal' embodied conversational agent in areas such as teaching, navigation assistance and entertainment.

There is extensive knowledge about human interpersonal relationships in the field of personality and social psychology. Aspects of friendship that need to be considered in ECA design are gender (e.g., activity-based men's friendship vs. affectively-based women's friendship), age, social class and ethnic background. Effects of friendship on interaction include increase of altruistic behavior, a positive impact on task performance and an increase in self-disclosure. Interpersonal attraction is an important factor in friendship. It is governed by positive reinforcements, and similarity between subjects is a key factor. Similarity of attitudes, personality, ethnicity, social class, humor, etc., reinforces the friendship relationship. Other issues are physical attractiveness (the 'halo effect') and reciprocity of liking (whether we think that the other person likes us). In [Stronks et al., 2002] we discussed the translation of the main aspects of human-human friendship to human-ECA friendship and how we can incorporate this translation in the design process of an ECA, using a scenario-based design. One observation is that it is important to distinguish between the initial design of an ECA and the possibility to change the ECA characteristics according to an adaptation strategy based on knowledge obtained by interacting with a particular user.

5.4 Humor in Embodied Conversational Agents

In previous years researchers have discussed the potential role of humor in the interface. Humans use humor to ease communication problems and in a similar way humor can be used to solve communication problems that arise with human-computer interaction. For example, humor can help to make the imperfections of natural language interfaces more acceptable for the users and when humor is sparingly and carefully used it can make natural language interfaces much friendlier. During these years the potential role of embodied conversational agents was not at all clear, and no attention was paid to their possible role in the interface.

Humans employ a wide range of humor in conversations. Humor support, or the reaction to Humour, is an important aspect of personal interaction and the given support shows the understanding and appreciation of humor. There are many different support strategies. Which strategy can be used in a certain situation is mainly determined by the context of the humorous event. The strategy can include smiles and laughter, the contribution of more humor, echoing the Humour and offering sympathy. In order to give full humor support, humor has to be recognized, understood and appreciated. These factors determine our level of agreement on a humorous event and how we want to support the humor.

Humor plays an important role in interpersonal interactions. From the many CASA experiments we may extrapolate that Humour will play a similar role in human-computer interactions. This has been confirmed with some specially designed experiments. There is not yet much research going on into embodied agents that interpret or generate Humour in the interface. In [Nijholt, 2002] we discuss how useful it can be, both from the point of view of humor research and from the point of view of embodied conversational agent research, to pay attention to the role of humor in the interaction between humans and the possibility to translate it to the interactions between humans and embodied conversational agents. Graphics, animation and speech synthesis technology make it possible to have embodied agents that can display smiles, laughs and other signs of appreciation of the interaction or explicitly presented or generated humor. There are many applications that can profit from being able to employ such embodied agents. The designer of the interface can decide when in certain scenarios of interaction agents should display such behavior. However, much more in the line of research on autonomous (intelligent and emotional) agents we rather have an agent understand why the

events that take place generate enjoyment by its conversational partner and why it should display enjoyment because of its appreciation of an humorous situation.

6 Conclusions

In this report we discussed the role of multimodality in human-computer interaction and in environments that know how to interpret human interactions and human-related events through multi-sensory perception. Examples of multimodal interfaces were shown. We emphasized the role of visualization in the interface, allowing us to consider interactions in virtual environments and their counterparts in real environments. Near term and long term roadmaps illustrating research issues until 2010 were presented. We zoomed in on the role of virtual humans in multimodal interfaces and virtual environments. We hardly were able to discuss interaction tools based on movement sensors, data gloves, eye trackers, haptic devices or physiological measurements. Several small-sized projects using such tools and devices are in progress in our research group. Another issue we did not discuss is the role of emerging standards and toolkits for designing virtual humans. They are concerned with the generation of gestures, facial expressions and body movements. Mark-up languages are being developed that allow designers to describe these nonverbal expressions. We think that our agent-oriented approaches, our layered software architectures and our attempts to model multimodality on a sufficiently abstract level will allow us to integrate known and not yet considered modalities in the current ideas and models of fusion and coordination.

Acknowledgements. In this report we surveyed part of our research on multimodal interfaces in the last three years. Many students and researchers contributed to this research. In particular we would like to thank: Betsy van Dijk, Rieks op den Akker, Dirk Heylen, Job Zwiers, Ivo van Es, Jeroen van Luin, Bas Stronks, Wauter Bosma, Natasa Jovanovic, Dennis Hofs and Marc Evers. Hendri Hondorp took care of the final text editing and formatting of this report. Harry Bunt (Tilburg), Oliviero Stock (Trento) and Gerhard Rigoll (München) were helpful in obtaining some of the pictures included in this report.

References

Argyle, M. and M. Cook. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge.

Broersen, A. & A. Nijholt. 2002. Developing a virtual piano playing environment. In: Proc. IEEE *International Conference on Advanced Learning Technologies (ICALT 2002),* V. Petrushin, P. Kommers, Kinshuk & I. Galeev (eds.), Kazan, Russia, 278-282.

Bunt, H. and R.-J. Beun (eds.). 2001. *Cooperative Multimodal Communication*. CMC'98 Selected Papers, Springer.

Bunt, H., M. Kipp, M.T. Maybury and W. Wahlster. 2003. Fusion and coordination for multimodal interactive information presentation. Chapter in *Intelligent Information Presentation*. O. Stock & M. Zancanaro (eds.), Kluwer Academic Publishers, to appear.

Cassell, J., J. Sullivan, S. Prevost and E, Churchill (eds.). 2000. *Embodied Conversational Agents*. The MIT Press.

Darken, R.P and J.L. Silbert. 1996. Way finding strategies and behaviors in virtual worlds. Proc. CHI'96, 142-149.

Evers, M. and A. Nijholt. 2000. Jacob - an animated instruction agent for virtual reality. In: *Advances in Multimodal Interfaces - ICMI 2000*, Proc. Third International Conference on

Multimodal Interfaces, Beijing, China, Lecture Notes in Computer Science 1948, T. Tan, Y. Shi and W. Gao (Eds.), Springer-Verlag, Berlin, 526-533.

Gebhard, P. 2001. Enhanching Embodied intelligent agents with affective user modelling. *UM2001, 8th International Conferenc*e, J. Vassileva and P. Gmytrasiewicz, (eds.), Berlin, Springer.

Heylen, D., I. van Es, B. van Dijk & A. Nijholt. 2003. Experimenting with the Gaze of a Conversational Agent. Chapter in *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. J. van Kuppevelt, L. Dybkjaer & N.O. Bernsen (eds.), Kluwer Academic Publishers.

Hofs, D., R. op den Akker and A. Nijholt. 2003. A Generic Architecture and Dialogue Model for Multimodal Interaction. Submitted for publication.

Höök, K. et al. 1988. Towards a framework for design and evaluation of navigation in electronic spaces. Persona Deliverable for the EC.

Hospers, M., E. Kroezen, A. Nijholt, R. op den Akker & D. Heylen. 2003. Developing a generic agent-based intelligent tutoring system and applying it to nurse education. In: Proceedings IEEE *International Conference on Advanced Language Technologies* (ICALT '03), Athens, Greece.

Johnston, M., P. Cohen, D. McGee, S. Oviat, J. Pittman and I. Smith. 1997. Unification-based multimodal integration. In: proceedings of the *35th Annual ACL Conference*, New Jersey, 281-288.

Johnston, M. and S. Bangalore. 2000. Finite-state multimodal parsing and understanding. In: Proceedings of *COLING-2000*, Saarbrücken, Germany.

Kendon, A. 1980. Gesticulation and speech: two aspects of the process of utterance. In: *The relation of verbal and nonverbal communication*. M.R. Key (ed.), Mouton, The Hague, the Netherlands.

Lester, J.C. et al. 1997. The persona effect: Affective impact of animated pedagogical agents. *CHI '97 Human Factors in Computing Systems*, ACM, 359-356.

van Luin, J. R. op den Akker and A. Nijholt. 2001. A dialogue agent for navigation support in virtual reality. Extended Abstracts *ACM SIGCHI Conference CHI 2001: Anyone. Anywhere.* Association for Computing Machinery, J. Jacko and A. Sears (eds.), Seattle, 117-118.

Maybury, M. and W. Wahlster (eds.). 1988. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Press.

McCowan, I., S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner and H. Bourlard. 2003. Modeling Human Interaction in Meetings. Proc. *IEEE ICASSP 2003*, Hong Kong.

Mikic, I., K. Huang and M. Trivedi. 2000. Activity monitoring and summarization for an intelligent meeting room. *IEEE Workshop on Human Motion*, Austin, Texas.

Nigay, L., and J. Coutaz. 1995. A generic platform for addressing the multimodal challenge. *ACM CHI Proceedings*, 98-105.

Nijholt, A. and J. Hulstijn. 2000. Multimodal Interactions with Agents in Virtual Worlds. Chapter 8 in *Future Directions for Intelligent Information Systems and Information Science*, N. Kasabov (ed.), Physica-Verlag, Springer, Heidelberg, 148-173.

Nijholt, A. Embodied Agents: 2002. A New Impetus to Humor Research. *The April Fools Day Workshop on Computational Humour*, O. Stock, C. Strapparava & A. Nijholt (eds.), In: Proc. Twente Workshop on Language Technology 20 (TWLT 20), Trento, Italy, 101-111.

A. Nijholt, J. Zwiers and B. van Dijk. 2003. Maps, agents and dialogue for exploring a virtual world. Chapter in *Web Computing*. J. Aguilar, N. Callaos and E.L. Leiss (eds.).

Oviat, S., P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro. 2000. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions for 2000 and beyond. Report.

Potter, D. (WP5-Team). 2003. Future Workspaces: A strategic roadmap for defining distributed engineering workspaces of the future. IST-2001-38346 deliverable.

Reeves, B. and C. Nass. 1996. *The Media Equation*. Cambridge University Press, Cambridge.

Shechtman, N. and L.M. Horowitz. 2003. Media inequality in conversation: how people behabe differently when interacting with computers and people. In: *SIGCHI-ACM CHI 2003: New Horizons*, ACM, New York, 281-288.

Stronks, B., A. Nijholt, P. van der Vet and D. Heylen. 2002. Designing for friendship: Becoming friends with your ECA. In: Proc. *Embodied conversational agents - let's specify and evaluate them!* A. Marriott, C. Pelachaud, T. Rist and Zs. Ruttkay (eds.), Bologna, Italy, 91-97.

Torres, O., J. Cassell and S. Prevost. 1997.Modeling gaze behavior as a function of discourse structure. First International Workshop on *Human Computer Conversations*. Bellagio, Italy.

Zobl, M., F. Wallhoff and G. Rigoll. 2003. Action recognition in meeting scenarios using global motion features. Proc. IEEE International Workshop on *Performance Evaluation of Tracking and Surveillance*.