Multimodal Interactions with Agents in Virtual Worlds

Anton Nijholt and Joris Hulstijn

Centre for Telematics and Information Technology

University of Twente

PO Box 217, 7500 AE Enschede, the Netherlands

anijholt@cs.utwente.nl

In this chapter we discuss our research on multimodal interaction in a virtual environment. The environment we have developed can be considered as a 'laboratory' for research on multimodal interactions and multimedia presentation, where we have multiple users and various agents that help the users to obtain and communicate information. The environment represents a theatre. The theatre has been built using VRML (Virtual Reality Modeling Language) and it can be accessed through World Wide Web (WWW). This virtual theatre allows navigation input through keyboard function keys and mouse, but there is also a navigation agent which tries to understand keyboard natural language input and spoken commands. Feedback of the system is given using speech synthesis. We also have Karen, an information agent which allows a natural language dialogue with the user. In development are several talking faces for the different agents in the virtual world. We investigate how we can increase the user's commitment to the environment and its agents by providing context and increasing the user's feeling of 'presence' in the environment.

Keywords: Human-computer Interaction, Virtual Reality, Embodied Agents, Talking Faces, Text-to-Speech Synthesis, Speech Recognition, Presence.

1 Introduction

World Wide Web allows interactions and transactions through Web pages using speech and language, either by inanimate or live agents, image interpretation and generation, and, of course the more traditional ways of presenting explicitly pre-defined information of text, tables, figures, pictures, audio, animation and video. In a task- or domain-oriented way of interaction current technology allows the recognition and interpretation of rather natural speech and language in dialogues. However, rather than the current two-dimensional web-pages, many interesting

parts of the Web will become three-dimensional, allowing the building of virtual worlds inhabited by user and task agents, with which the user can interact using different types of modalities, including speech and language interpretation and generation. Agents can work on behalf of users, hence, human computer interaction will make use of 'indirect management', rather than interacting through direct manipulation of data made visible on the screen.

In this chapter we present our research on the development of an environment in which users can display different behavior and have goals that emerge during the interaction with this environment. Users who, for example, decide they want to spend an evening outside their home and, while having certain preferences, cannot say in advance where exactly they want to go, whether they first want to have a dinner, whether they want to go to a movie, theatre, or to opera, when they want to go, etc. During the interaction, goals, possibilities and the way they influence each other become clear. One way to support such users is to give them different interaction modalities and access to multimedia information. We discuss a virtual world for presenting information and allowing natural interactions about performances, associated artists and groups, availability of tickets, etc., for some existing theatres in the city of Enschede, the Netherlands, the home town of our university.

The interactions take place in a virtual theatre, a realistic model of one of our local theatres. This so-called 'Muziekcentrum' offers its potential visitors information about performances (music, cabaret, theatre, opera) by means of a brochure that is published once a year. In addition to this yearly brochure it is possible to get information at an information desk in the theatre (during office hours), to get (more recent and updated) information by phone (either by talking to a theatre employee or by using Interactive Voice Response Technology) and to get information from local daily and weekly papers and monthly announcements issued by the theatre. The central database of the theatre holds the information that is available at the beginning of the 'theatre season'. Our aim is to make this information about theatre and performances much more accessible to the general audience.

The interactions between user (the visitor) and system take place using different task-oriented agents. These agents allow mouse and keyboard input, but interactions can also take place using speech and language input. In the current system both sequential and simultaneous multi-modal input is possible. There is also multi-modal (both sequential and simultaneous) output available. The system presents its information through agents that use tables, chat windows, natural language, speech and a talking face. At this moment this talking face uses speech synthesis with associated lip movements. Other facial animations are possible (movements of head, eyes, eyebrows, eyelids and some changes in face color). These possibilities have been designed and associated with utterances of user or system, but not yet fully implemented.

In this chapter it is also discussed how our virtual environment can be considered as an interest community and it is shown what further research and development is required to obtain an environment where visitors can retrieve information about artists, authors and performances, can discuss performances with others and can be provided with information and contacts in accordance with their preferences. In addition, but this has not been realized yet, we would like to offer our virtual environment for others to organize performances, meetings and to present (video) art. We would like to offer this environment for experiments on mediated communication between visitors and for performances, done by avatars with or by avatars without user participation.

The virtual environment we consider is web-based and the interaction modalities that we consider confine to standards that are available or that are being developed for world wide web. From a more global point of view research topics that have been aimed at are:

- Modeling effective multi-modal interactions between humans and computers, with an emphasis on the use of speech and language
- Commercial transactions, (local, regional and global) information services, education and entertainment in virtual environments
- Web-based information and transaction services, in particular interactions in virtual environments

Indexing and retrieval of multi-media information (in our case, multi-media information about artists and performances) available in theatre databases and on the world wide web is an issue that will receive more attention in the near future. We assume that it will become clear how to generalize our approaches and how to tune them to domains different from our (virtual) theatre domain.

2 History and Motivation

Some years ago, the Parlevink Research Group of the University of Twente started research and development in the area of the processing of (natural language) dialogues between humans and computers. In order to do so, one can choose to take a general, domain-independent approach. This allows general research in syntactic analysis, semantic and pragmatic interpretation and the modeling of dialogues in general. Hence, research has to be embedded in current state-of-the-art research on parsing, unification, grammar formalisms, semantics and representation of dialogue utterances, discourse representation and the representation of ‘common sense’ and world knowledge. That is, knowledge that has to be represented and made accessible in order to get our system to understand user utterances and to generate intelligent system utterances.

Although this domain-independent approach has been followed, we also started research where the domain of application and the user interaction associated with the domain is a basic assumption of interaction modeling and dialogue management. In our case, the domain is that of a theatre information service and related transactions (e.g., to make a reservation). In order to be able to model 'natural' interaction, Wizard of Oz experiments have been designed and a corpus of (keyboard) natural language dialogues in this domain has been obtained.

Our dialogues involve transactions. Such dialogues display a more complex structure than mere inquiry or advisory dialogues. Two tasks are executed in parallel: obtaining information and ordering (Jönsson [8]). In our corpus similar complex behavior related to these tasks can be found. Users browse, inquire and retract previous choices, for instance when tickets are too expensive. Hence, we allow interactions where the goal is not set before; it develops during the dialogue and the user will update goals depending on the information obtained. It is also the environment in which the dialogues are embedded and the possibility to explore environment and interaction modalities that invites users to browse through the available information just like leafing through a brochure.

Our research led to the development of a (keyboard-driven) natural language accessible information system (SCHISMA), able to inform users about theatre performances and to allow users to make reservations for performances. The system made use of the database of performances in the local theatres of the city of Enschede. In the next sections we will give more information about the design of this theatre information system. The system is far from perfect. However, if a user really wants to get information and has a little patience with the system, he or she is able to get this information. A more general remark should be given: When we offer an interface to the general audience to access an information system, do we want to offer an intelligent system that knows about the domain, that knows about users, their preferences and other characteristics, etc., or do we assume that any user will adapt to the system that is being offered? This is not an exclusive or, nevertheless the point is extremely important. It has to do with group characteristics (men, women, old, young, naive, professional, experienced, etc.), but also with facilities and alternatives provided by the owner of the system. As an example, consider a transport and railway information system. Human operators are available to inform about times and schedules of busses and trains. However, the number of operators, from a user's point of view, is insufficient. Callers can wait (and pay for the minutes they have to wait) or choose for a computer-operated system to which they can talk in natural speech, but possibly have to accept that they need more interactions in order to get themselves understood. Hence, it really depends on the application, the situation and the users involved (do they want to pay for the services, do they want to adapt to the interface, does the provider offer an alternative, etc.), whether we can speak of a successful natural language accessible dialogue information system.

We do not really disagree with a view where users are expected to adapt to a system. On the other hand, wouldn't it be much more attractive (and interesting from a research point of view) to be able to offer environments, preferably on worldwide web, where different users have different assumptions about the available information and transaction possibilities, have different goals when accessing the environment and have different abilities and experiences when accessing and exploring such an environment? We like to offer a system such that we can stimulate and expect users to adapt to it and find effective, efficient, but most of all enjoyable ways to get or to get done what they want.

In the next section we discuss how we can add 'context' to our dialogue system. With 'context' we mean that we would like to add visual and auditory cues in the presentation of information and to allow users to choose the (combination of) interaction modalities that best suits his or her preferences for performing the 'task' that has to be done. With 'context' we also refer to the possibility that users come to consider the environment as an interest community, where they can exchange information with other users.

3 VR Context-Embedded Interaction

3.1 Environment Visualization

In order to add context to our natural language dialogue system we decided to visualize the environment in which people can get information about theatre performances, can make reservations and can talk to theatre employees and other visitors. VRML, agent technology, text-to-speech synthesis, talking faces, speech recognition, etc., became issues after taking this decision. They will be discussed in the next sections. Visualization allows users to refer to a visible context and it allows the system to disambiguate user's utterances by making use of this context. Moreover, the visualization can make it possible for the system to influence the interaction behavior of the user in such a way that more efficient and natural dialogues with the system become – in principle - possible.

Our virtual theatre (<http://parlevink.cs.utwente.nl/>) has been built according to the design drawings made by the architects of our local theatre. Part of the building has been realized by converting AutoCAD drawings to VRML97. Video recordings and photographs have been used to add 'textures' to walls, floors, etc. Sensor nodes in the virtual environment activate animations (opening doors) or start events (entering a dialogue mode, playing music, moving spotlights, etc.). Visitors can explore the environment of the building, hear the carillon of a nearby church, look at a neighboring pub and movie theatre, etc. and they can enter the theatre (cf. Figure 1) and walk around, visit the concert hall, admire the paintings

Figure 1: Entrance of the Theatre

on the walls, go to the balconies and, take a seat in order to get a view of the stage from that particular location. When the performance hall is entered, the lights dim, spot lights are moving over the stage and some music starts playing. Information about today's performances is available on an information board that is automatically updated using information from the database with performances. In addition, as may be expected, visitors may go to the information desk in the theatre, see previews of performances and start a dialogue with an information and transaction agent called 'Karen'. This agent has a 3D talking face (see section 4).

Apart from navigating, clicking on interesting objects (resulting in access to web pages with information about performances, access to web magazines, etc.) and interacting with person-like agents we allow a few other interactions between visitors and virtual objects. For example, using the mouse, the visitor can play with the spotlights and play notes on a keyboard that is standing in some far away part of the building. There is a floor map near the information desk where people can click on positions in order to be 'transported' to their seat in the performance hall so they can see the view they have. On the desk is also a monitor on which they can see pictures or video previews of performances. Unfortunately, most performances do not have a video preview available yet, so we can not display them for every performance that is in the database.

3.2 Visualizing Agents

We assign natural tasks in our environment to agents. It can be useful to visualize them using talking faces and animated 3D avatars. From several studies (cf. Friedman [6]) it has become clear that people engage in social behavior toward machines. It is also well known that users respond differently to different

‘computer personalities’. It is possible to influence the user’s willingness to continue working even if the system’s performance is not perfect. Users can be made to enjoy the interaction and they can be made to perform better, all depending on the way the interface and the interaction strategy have been designed. People behave differently in the presence of other people than they do when they are alone. Similarly, in experiments it has been shown that people display different behavior when interacting with a talking face than they do with a text-display interface. This behavior is also influenced by the facial appearance and the facial expressions that are shown. People tend to present themselves in a more positive light to a talking face display and they are more attentive when a task is presented by a talking face (cf. Sproull et al. [12]). From these observations we conclude that introducing talking faces can help make interaction more natural and shortcomings of technology more acceptable to users.

There is another reason to introduce visualized task-oriented agents. The use of speech technology in information systems will continue to increase. Most currently installed information systems that work with speech, are telephone-based systems where callers can get information by speaking aloud some short commands. There are exceptions. For example, there are applications where customers can call a specific telephone number which connects them to a website. Data at the site can be accessed and retrieved (using speech recognition and speech synthesis). This is especially useful when a caller needs to order, to book, to fill in forms, etc., or to access information that is available in tables. Another example is the demonstrator for the SNCF (French railways) train travel information kiosk, where train passengers have access to information services using multimodal input and multimedia output. Speech and touch are among the input modalities. Sound, video, text and graphics are the output modalities. Speech is becoming more and more important in these multimodal interfaces. One of the main problems in speech-only dialogue systems is the limitation of the context. As long as the context is narrow they perform well, but wide contexts are causing problems. One reason to introduce task-oriented agents is to restrict user expectations and utterances to the different tasks for which agents are responsible. Obviously, this can be enhanced if the visualization of the agents helps to recognize the agents tasks.¹

¹ It may be the case that different specialist agents are taken more seriously than one generalist agent. See Nass et al. [11] who report about different appreciation of television programs depending on whether they were presented in a ‘specialist’ or in a ‘generalist’ setting.

3.3 Multi-modal and Multi-agent Approach

When a user has the possibility to change easily from one modality to an other, or can use combinations of modalities when interacting with an information system, then it is also more easy to deal with shortcomings of some particular modality. Multi-modality has two directions. That is, the system should be able to present multi-media information and it should allow the user to use different input modalities in order to communicate with the system. Not all communication devices that are currently available for information access, exploration of information and for transaction allow more than one modality for input or output. This is especially true if we look at world wide web interfaces.

More serious, however, hardly any research has been done to distinguish discourse and dialogue phenomena, let alone to model them, for multi-modal tasks. The same holds for approaches to funnel information conveyed via multiple modalities into and out of a single underlying representation of meaning to be communicated (the cross-media information fusion problem). Similarly, on the output side, there is the information-to-media allocation problem.

In addition to the issue of multi-modality, there is a need for a multi-agent approach. Our system deals with actors that present information, reason about information, communicate with each other and that realize transactions (e.g. through negotiation). In this case agents can take roles ranging from presenting windows on a screen, reasoning about information that might be interesting for a particular user, and being recognizable (and probably visible) as being able to perform certain tasks.

3.4 Virtual Communities

Today there are examples of virtual spaces that are visited and inhabited by people sharing common interests. These spaces can for example, represent offices, shops, class rooms, companies, etc. However, it is also possible to design virtual spaces that are devoted to certain themes and are tuned to users (visitors) interested in that theme or to users (visitors) that not necessarily share common (professional, recreational or educational) interests, but share some common conditions (driving a car, being in hospital for some period, having the same therapy, belonging to the same political party, etc.).

In the previous subsections we have looked at possibilities for theatre visitors to access information, to communicate with agents designed by the provider of the information system and to explore an environment with the goal to find information or to find possibilities to enter into some transaction. Hence, we have a community of people interested in theatre, in music, in performers and their environment has been modeled along the lines of an existing theatre. We need to investigate how we can allow communication between users or visitors of this

web-based information and transaction system. Users can help each other to find certain information, they can inform each other (especially when they know about the other's interests), they can have conversations about common interests and they can have domain-related collaboration (e.g., in our case, they can decide to perform a certain play where the actors are distributed among different web sites but sharing the same virtual stage).

As a (not too complicated) example we mention a virtual world developed by the virtual worlds group of Microsoft in co-operation with The Fred Hutchinson Cancer Research Center in Seattle. This so-called "Hutch World" enables people struggling with cancer to obtain information and interact with others facing similar challenges. Patients, families and friends can enter the password protected three-dimensional world (a rendering of the actual outpatient lobby), get information at a reception desk, visit a virtual gift shop, etc. Each participant obtains an avatar representation. Participants can engage in public chat discussions or invitation-only meetings. A library can be visited, its resources can be used and participants can enter an auditorium to view presentations.

4 Agents in the Virtual Theatre

4.1 An Agent Platform in the Virtual Environment

In the current prototype version of the virtual theatre we distinguish between different agents: We have an information and transaction agent, we have a navigation agent and there are some agents under development. An agent platform has been developed in JAVA to allow the definition and creation of intelligent agents. Users can communicate with agents using speech and natural language keyboard input. Any agent can start up other agents and receive and carry out orders of other agents. Questions of users can be communicated to other agents and agents can be informed about each other's internal state. Both the information & transaction agent and the navigation agent are in the platform. But also the information board, presenting today's performances, has been introduced as an agent. And so can other objects in the environment.

4.2 The Information & Transaction Agent

Karen, the information & transaction agent, allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karen (Figure 2) wants to give information and to sell tickets. Karen is fed from a database that contains all the information about performances in the (existing) theatre.

Figure 2: Karen at the Information Desk

Our current version of the dialogue system of which Karen is the face uses a ‘rewrite and understand’ approach. User utterances are simplified using a great number of rewrite rules. The resulting simple sentences are parsed. The output can be interpreted as a request of a certain type. System response actions are coded as procedures that need certain arguments. Missing arguments are subsequently asked for. The system is modular, where each ‘module’ corresponds to a topic in the task domain. For example, a module has to take care of a date a user is referring to (next Wednesday, over two weeks, tomorrow).

There are also modules for each step in the understanding process: the rewriter, the recognizer and the dialogue manager. The rewrite step can be broken down into a number of consecutive steps that each deal with particular types of information, such as names, dates and titles. The dialogue manager initiates the first system utterance and goes on to call the rewriter and recognizer process on the user's response. Also, it provides an interface with the database management system. Results of queries are represented as bindings to variables, which are stored in the global data-structure, called context. Based on the user utterance, the context and the database, the system has to decide on a response action, consisting of database manipulation and dialogue acts. The arguments for the action are dug out by the dedicated parser, associated with the topic. All arguments that are not to be found in the utterance are asked for explicitly. More information about this approach can be found in Lie et al. [9].

Presently the input to Karen is keyboard-driven natural language and the output in our for the general audience WWW accessible virtual world allows for a mix of speech synthesis and information presentation on the screen. As mentioned earlier, Karen's spoken dialogue contribution is presented by visual speech, that is, a 'talking face' on the screen, embedded in the virtual world, mouths the questions and part of the responses. If necessary, information is given in a window on the screen, e.g., a list of performances or a review of a particular performance. The user can click on items to get more information or can type in further questions concerning the items that are shown.

4.3 The Navigation Agent

Navigation in virtual worlds is a well known problem. Usually, navigation input is done with keyboard and mouse. This input allows the user to move and to rotate, to jump from one location to an other, to interact with objects and to trigger them. There is more to say about navigation:

- *Input tools*: Which tools do we use to navigate? Keyboard, mouse, 3D stick, bookmarks, menus, speech, gestures, gaze control?
- *Physical*: How do we move in the virtual world? How can we avoid collisions with physical objects and with other agents? How can we employ knowledge of the environment and knowledge of tasks in order to control the movements of our agents?
- *Orientation*: How does a user know where s/he is and how s/he should move to get to another location in the virtual environment. How do we represent spatial knowledge and how do we assist users to think, reason and argue about objects, locations and routes in the virtual world?
- *Social*: How can we improve social navigation? There are so many users who know how to get at certain information that suits their interests. So why not exploit their knowledge to identify routes, to filter and to select information?

Related issues to navigation in multi-user environments are personal space, group space and privacy. Including these issues involves also the study of social norms, both in virtual and physical environments.

We developed a navigation agent that helps the user to explore the environment by means of speech commands. It is left to the user to choose between interaction modes for navigation. A smooth integration of the pointing devices and speech in a virtual environment requires that the system has to resolve deictic references that occur in the interaction. Moreover, the navigation agent should be able to reason (in a modest way) about the geometry of the world in which it moves. It knows about the user's coordinates in the virtual world and it has knowledge of the coordinates of a number of objects and locations. This knowledge is necessary

when a visitor refers to an object close to the navigation agent in order to have a starting point for a walk in the theatre and when the visitor specifies an object or location as the goal of a route in the environment. The navigation agent is able to determine its position with respect to nearby objects and locations and can compute a walk from this position to a position with coordinates close to the goal of the walk.

Verbal navigation requires that names have to be associated with different parts of the building, objects and agents. Users may use different words to designate them, including references that have to be resolved in a reasoning process. The current agent is able to understand command-like speech or keyboard input. However, it hardly knows how to communicate with a visitor. The phrases to be recognized must contain an action (go to, tell me) and a target (information desk, synthesizer). It tries to recognize the name of a location in the visitor's utterance. When the recognition is successful, the agent guides the visitor to this location. When the visitor's utterance is about performances the navigation agent makes an attempt to contact Karen, the information and transaction agent. In progress is an implementation of the navigation agent in which the navigation agent knows more about (or is able to compute) current position and focus of gaze of the user, geometric relations between objects and locations, knowledge of previously visited locations or routes and knowledge of the previous communication with the visitor.

4.4 Language Skills of the Agents

At this moment our agents have different language skills. On the one hand we have Karen and a grammar specification of the input for Karen based on a corpus of WoZ obtained keyboard-based dialogue utterances. On the other hand we have a navigation agent with language skills that are based on the current limitations of speech input. We would like to automate the process of assigning language skills to agents in our environment as much as possible. Therefore we hope to see speech recognition technology move forward from keyword spotting, to finite state utterance specification to (word-graph) based context-free language specification. At this moment we follow the developments of Philips speech recognition software when looking at the recognition of spoken dialogue utterances for different agents. More fundamental, however, is our approach to induce grammars (context-free, probabilistic, unification constraints) from corpora of utterances (see Ter Doest [5]) collected in Wizard of Oz experiments. In this approach we tag a corpus with syntactic categories and superficial structure using Standard Generalized Markup Language (SGML). From this tagged data grammar rules, unification constraints and probabilities are derived. We have tested grammars on 'seen' and 'unseen' data from Karen's same domain using a probabilistic left-corner parser for PATR II unification grammars. In a similar way we have induced for our navigation agent a probabilistic grammar from a corpus

of user utterances that have been obtained from several scenarios presented to (potential) visitors from the theatre. This grammar is a start. It allows the design of a primitive system and it allows bootstrapping this system from the original corpus and from corpora obtained from logging the interactions between visitors and the navigation agent. Clearly, this approach still requires the integration of speech recognition technology with natural language specification and understanding. For that reason it may be useful to investigate the generation of finite state probabilistic (unification) grammars from corpora of utterances.

In our current web-based system we have speech recognition on the server side. This requires the recording of commands on the client side and a robust transporting of the audio files. It does not require users to install speech recognition software or to download a speech recognition module as part of the virtual world from the server. Users do need however audio-software, which is usually available anyway. For speech recognition we use Speech Pearl, commercial speech understanding software from Philips. Recognition is based on keyword spotting. A next version of the software will allow a finite state specification of the user's input for speech recognition. For text-to-speech synthesis we use a Dutch text-to-speech system [4]. It runs on top of the MBROLA diphone synthesizer. Lexical resources of *Van Dale Dictionaries* have been used to obtain phonetic descriptions of the words.

5 Facial Displays and Speech Generation

5.1 Introduction

We developed a virtual face in a 3D-design environment. The 3D data has been converted to VRML-data that can be used for real-time viewing of the virtual face on WWW. We are researching various kinds of faces to determine which can be best used for our applications. Some are rather realistic and some are more in a cartoon-style (cf. Figure 3). The face is the interface between the users of the virtual theatre and the theatre information system. A dialogue window is shown when users approach the information-desk while they are navigating in the virtual theatre. Users can type their questions in a user's window. Karen answers the user in spoken language and by

Figure 3: Cartoon Face for Karen

showing her response in a window. In addition, a list of performances can be presented on the screen. While listening and answering Karen's face may change, reflecting her attitude towards the user and the contents of the dialogue.

How do we control the responses of the system, the prosody and the artificial face? The central module of our dialogue system is the *dialogue manager*. It maintains two data-structures: a representation of the *context* and a representation of the *plan*, the current domain-related action that the system is trying to accomplish. Based on the context, the plan and a representation of the latest user utterance, the dialogue manager selects a certain response action. Planning and action selection are based on a set of principles, the *dialogue rules*. A response action is a combination of basic domain related actions, such as database queries, and dialogue acts to convey the results of the query. Dialogue acts describe the intended meaning of an utterance or gesture. The *response generation* module selects a way to express it. It determines the utterance-structure, wording, and prosody of each system response. In addition, it also controls the orientation and expression of the face, the eyes, and the coordination of sounds and lip movement.

5.2 Prosodic and Facial Features

5.2.1 Prosodic Features and Templates

In the design of utterance generation by the information agent a list of annotated utterance templates is used. The response generation module uses a set of parameters to control the templates. Templates contain gaps to be filled with *information items*: attribute-value pairs labeled with syntactic, lexical and phonetic features. An appropriate template for a given dialogue act is selected by the following parameters: *utterance type*, *body* of the template, *given* information, *wanted* and *new* information. The utterance type and body determine the word-order and main intonation contour. The given, wanted and new slots, as well as special features, affect the actual wording and prosody. Templates respect rules of accenting and de-accenting. As a rule, information that is assumed to be given in the dialogue is de-accented, expressed as a pronoun, or even left out. Given information is repeated whenever the system is not confident it was recognized correctly, e.g., by a speech recognition module. Such verification prompts are distinguished by a rising intonation. Information that is to be presented as new, is accented. Quoted expressions, like artist names or titles of performances, are set apart from the rest of the utterance. For reading the longer texts and reviews that describe the content of performances, the system assumes a 'reading voice'.

Our text-to-speech system operates at three levels: the grapheme level, the phoneme level and a low-level representation of phones where the length and pitch of sounds is represented. It has phonetic descriptions of the usual Dutch words as they appear in lexical resources. Other prosodic information is derived

by heuristic rules. It is possible to manipulate prosody by adding punctuation at the grapheme level, by adding prosodic annotations at the phoneme level or by directly manipulating the phone level.

5.2.2 Facial Features and Templates

Apart from the lips, the virtual face has a number of dynamic control parameters (Figure 4). The *eyes* can gaze at a certain direction. This can be used to direct attention towards an area. The *eyelids* may be opened and closed, for blinking. The *eyebrows* can be lifted to indicate surprise or lowered for distress. The shape of the *mouth* can be manipulated into a smile or and angry expression. The *color* of the face can be deepened, to suggest a blush that indicates shyness or embarrassment. The *orientation* of the head can be manipulated, leaning forward and backward or tilting left and right. This may produce important facial gestures like nodding and shaking one's head. It can also be used to indicate attention; leaning forward means being interested, leaning backward means loosing interest. In general the character is not still. The head will wiggle a bit and its eyes will wonder. This is called *idle behavior*. Many existing 'talking heads' look artificial because of their stillness. Moreover, not moving can also be taken as a sign. For instance, a fixed stare can indicate a misunderstanding in the dialogue. The *frequency* of idle movements is an indicator of the liveness of the character; it serves as a type of volume, to the existing emotion. So, many random movements of the head, combined with smiles and attentive eyes, indicate a very happy personality; stillness, a neutral mouth shape and looking away, indicate a withdrawn and unhappy personality. But an angry face, combined with a blush and a lot of movement, indicate increased anger. Jerky movements with wondering eyes indicate nervousness. Since our agent is supposed to be

Feature	Manipulation	Meaning
Eyes	Gaze direction	Idle behavior, attention, indexing
Eyebrows	Lift, lower	Surprise, distress, angry
Lips	Form visemes	Talk
	Stretch, round	Smile, laugh, neutral, angry, kiss
Mouth shape	Stretch, round	Smile, neutral, angry
Color	Blush	Shyness, embarrassment
Head	Orientation	Nodding, shaking head, attention
	Idle behavior	Neutral
	Movement frequency	Emotional 'volume'
Shoulders	Shrug	Indifference

Figure 4: Facial Features

professionally friendly, she will be generally smiling and will have a moderate movement frequency.

Each of these basic features can be combined into facial *gestures* that can be used to signal something. Gestures like nodding, shaking and shrugging can be used separately, but often utterances are combined with gestures or utterance related facial expressions. The timing of the gesture or the expression must be aligned with the utterance. We use the following general heuristic for alignment of gestures.

Like any event, an utterance and a gesture have an *entry* and an *exit* point. Moreover, an utterance can be broken down into phrases; each phrase has a so called *intonation center*, the moment where the pitch contour is highest. Since pitch accents are related to informativeness, we can assume that the accent lands on the most prominent expression. Usually the accent lands towards the end of an utterance. Similarly, each gesture has a *culmination point*. For instance for pointing, the moment that the index finger is fully extended. The visual animator extrapolates a nice curve from the entry point to the culmination and again to the exit point. Our current working hypothesis is that gestures synchronize with utterances, or precede them. So we link the gesture's entry and exit points to the entry and exit points of the utterance and make sure that the culmination point occurs before or on the intonation center.

5.3 Facial Behavior

Facial behavior is linked to personality and to the conversation that is being held. Facial behavior may also be related to the task an agent has to perform. For a chosen face there are different types of behavior that need to be modeled.

Firstly, permanent features like the facial expression, gazing direction and general movement characteristics, both when speaking and when idle. These can be controlled by two parameters: *mood* and *attention*. The *mood* parameter indicates the general attitude of the personality in the conversation. It is a state, that extends over a longer period. Is the agent happy, sad, angry or uncertain? The *attention* parameter controls the eyes and gazing direction. We believe that one of the benefits of a talking face is that turn taking and attention management in dialogues will be made easier. The gazing direction of the eyes and the head position are crucial for this. In section 6 we will discuss the possibility to include appropriate gaze behavior in our human-like agents for regulating a conversational process. Usually mood and attention are fixed for a given personality. Temporary changes in emotion and attention, may result from previous utterances or from the general conversation. For instance, anger at an insult, or increased interest after a misunderstanding.

Figure 5: Blackboard Architecture

Secondly, utterance related attitudes. Since we cannot monitor the user's utterances in real-time, at the moment this is limited to system utterances only. Think of smiling at a joke, raising eyebrows at a question or a pointing gesture at an indexical. Conventional gestures can be modeled as a special instance of response actions. Nodding or shrugging are coded like any other utterance synchronized with a gesture, except that they can be silent. Utterance related features are controlled by the existing utterance parameters, extended with a new parameter, *gesture*, that labels one or more facial movements to be synchronized with the utterance template. Because we know all utterance templates in advance, the synchronization can be manually adjusted if needed. The extend of the gesture and its final shape also depend on the general emotional state and attention level.

To control the many features of facial behavior we need a blackboard architecture as depicted in Figure 5. The reason for this architecture is that the parameters influence each other. Combinations of input parameters trigger a rule that produces a response action, or a more permanent change of expression.

We hope to introduce some variation in the exact choice of movement. Variation is important. For instance, it is natural to combine "yes" with a nod, but when every affirmative is combined with the same nod it looks mechanical. Another example is the raising of eyebrows. In an early version of the system the eyebrows were directly controlled by pitch level. Thus, the agent would nicely express uncertainty on a question, which has a rising intonation. But of course, pitch is also used for accenting. So the agent wrongly expressed surprise at expressions that were stressed. Synchronizing the apparently random movements with fixed speech from templates is difficult. We have chosen to align the culmination points of the movement with the intonational centers of the phrases uttered. But the exact frequency and extent of the movements will be randomly distributed, partly based on mood and attention.

5.4 Importance of Nonverbal Behavior

Nonverbal behavior as described in the previous sections is important. It provides feedback for the conversation partner of the embodied agent. This feedback helps to obtain a more smooth conversation and exchange of information. Cassell et al [2] have compared different kinds of nonverbal feedback in some experiments. They distinguished:

- content-only feedback: answering questions and executing commands related to the topic of the conversation
- envelope feedback: nonverbal behaviors related to the process of conversation, e.g., glances towards and away from the partner, manual gestures, head movements, etc.
- emotional feedback: facial displays that reference a particular emotion (smiles, look of puzzlement, etc.)

In the experiments they compared three feedback behaviors: content-only feedback (speech and actions), envelope feedback added to content feedback, and emotional feedback added to content feedback. Some questionnaires were presented to users in which they could give their opinion on life-likeness and ease of interaction. Moreover, there were measurements on the relative number of utterances, hesitations and expressions of frustrations. It could be concluded that the envelope feedback is much more important than the emotional feedback. In fact, it is argued that the possibility to visualize such behavior is the strongest argument for using embodied agents.

6 Future and More Sophisticated Embodied Agents

6.1 Introduction

There is no need to visualize all agents in a virtual environment, especially if the environment is designed in an agent-oriented way. In that case a window appearing on the screen may be an agent, a door knob in a virtual environment may be an agent, but also a non-visible dialogue manager or database management system can be designed as an agent. We think it is useful to have a visualization of the agents that have explicit dialogues with the visitors of our environment. Maybe not necessarily during the dialogue, but, for example, to initiate a dialogue or to refer to a previous dialogue. It is clear that Karen needs to be represented as a person in the virtual world. Moreover, she needs an animated talking face and text-to-speech synthesis. It is not really clear whether the navigation agent requires

a human face and body. However, if this agent is supposed to walk in front of us, it needs some visualization. Especially if we want agents that are visible for the visitor, agents that walk, agents that show how to do certain things and agents that interact with users in a natural way, then we need rather natural visualization of movements of agents (including movements of body, legs, arms, fingers, etc.) and animation of facial expressions, all in accordance with the tasks that the agent has to perform and the interaction with the visitor that is required. Generation of natural language and the corresponding synthesis of speech and facial expressions has been discussed in the previous sections. What remains is the interaction between visitors and visualized agent such that the agent's movements in the virtual environment are natural, given the task the visitor wants it to perform.

6.2 Embodied Agents and Gaze Behavior

In our environment we have different human-like agents. Some of them are represented as communicative humanoids, more or less naturally visualized avatars standing or moving around in the virtual world and allowing interaction with visitors of the environment. In a browser which allows the visualization of multiple users, other visitors become visible as avatars. We want any visitor to be able to communicate with agents and other visitors, whether visualized or not, in his or her view. That means we can have conversations between theatre agents, between visitors, and between visitors and agents.

If a visitor, whether visualized or not, has a visual conversation partner on the screen, gaze behavior of this partner can help in making the conversation more natural. This behavior can be used to better convey the message (just as face expressions, head movements, gestures and intonation) and to have more natural turn-taking in dialogue or actions. Clearly, gaze behavior of a conversational partner on the screen (whether human or artificial) should be oriented towards the human partner looking at the screen or towards its representation (maybe just as a hand or a 3D face) on the screen.

In general we want our embodied agent to look at the user when it speaks to the user or when the user is looking or speaking to the agent. However, the agent can temporarily look away from the user when starting a turn, it can look away to draw attention to something else on the same screen (another agent or user representation, object, menu or window) on the same screen or a situation not on its screen, but somewhere in the user's environment. It may also follow the user's gaze in order to show interest and to anticipate a next topic that may arise in the conversation.

Similarly, in a neutral situation we assume that a user looks at an embodied conversational partner. The user may look away from an agent to start the turn, to follow the agent's gaze or to take an independent action to look to another agent, object or window on the same screen, to look at another screen or object in the

environment, or to use the keyboard or perform another task related to the conversation. The user may also decide to start looking or to look back at a particular agent. This can be taken as a sign that the user wants to address this particular agent or, in the process of a conversation, wants to give the turn, so the agent can start speaking.

We conclude that there are many reasons to model gaze behavior of our agents, whether they represent humans or theatre employees and whether they talk to each other or to a human looking at the screen. However, as may already have become clear from above, there are so many factors that determine gaze behavior (including the relation between the conversational partners, the number of partners, noise in the environment, tasks that have to be performed at the same time, mood and personality of the partners, etc.) that we can not expect to be able to model all of them. Best investigated is the relation between gaze behavior and turn-taking in a conversation. In summary: often a speaker looks away from a hearer when she starts an utterance and she looks towards the hearer when ending an utterance.² Cassell et al. [1,3] have investigated the relation between propositional content of utterances and gaze behavior. This is interesting from the point of view that we can make further steps towards generating the behavior of agents (syntactic, semantic and pragmatic content of utterances, intonation, face expressions, gaze behavior, head and body movements) from a representation of (the history of) previous interactions, the representation of beliefs, desires and intentions and the representation of some personality characteristics. In the tradition of Halliday, Cassell et al distinguish between the 'theme' and the 'rheme' of an utterance. The 'theme' can be considered as the topic of the conversation. Hence, the part of the utterance that links the conversation to previous parts is the 'theme', it does not really add something new to the conversation. The 'rheme' of an utterance adds something new to the conversation. That is, something which is new information for the hearer. From an empirical analysis of experimental data they conclude that:

- starting the thematic part of an utterance is frequently accompanied by gaze behavior that looks away from the hearer, while beginning of the rheme is usually accompanied by gaze behavior that looks towards the hearer; and
- if the start of the thematic part coincides with the start of the turn, the speaker always looks away from the hearer; in cases where the rheme coincides with the end of the turn, the speaker always looks towards the listener.

Obviously, the behavior of computer-generated agents, their movements, face expressions, intonation of speech and the gaze behavior described above can be

² This gaze behavior is only one of many cues that speakers employ. There may be gestures that indicate the end of a turn, the pitch of the utterance may fall towards the end and some specific phrases ('don't you think so?') may give the floor to an other conversational participant.

controlled by our algorithms and programs in the virtual environment. In particular when the agent's contributions to the dialogue are generated from templates the 'theme' and 'rheme' of the generated utterance are known and, assuming the utterance is not too short, gaze behavior can be adjusted to the roles the different phrases in the utterance have.³

In an attempt to profit as much as possible from the user's input, whether given deliberately or not, to the (agents on the) screen, it is useful to see whether the gaze behavior of the user or visitor can not only give information about the way the conversation should proceed, but also how certain input from other modalities (speech, keyboard, mouse) should be interpreted. Obviously, this is part of our philosophy how to deal with imperfect interaction technology, i.e., the integration of different interaction modalities makes it possible to reduce the number of possible interpretations of whatever modality and to better return a system or system agent response that helps the user to continue his task and exploration. In particular we want an agent to know that the visitor is addressing him or her. That is, the agent should detect that the visitor is looking at him or her rather than looking at an other agent (window, menu, etc.) visible at an other position on the screen.

Presently, we are doing experiments with an eyetracker system. Such a system makes it possible to determine where a person is looking at. In particular, it is possible to determine to which avatar a person is looking. This allows management of multi-user conversations in a virtual environment, where each user knows when and which other users are looking at him or her. This leaves to a certain degree open how the user is represented in the environment, but at least user gaze directional information can be conveyed. Hence, visitors of our environment can address different task-oriented agents in such a way that speech recognition and language understanding are tuned to the particular task of the agent; therefore quality of recognition and understanding can increase, since the agent may assume that words come from a particular domain and that language use is more or less restricted to this domain. That is, we can restrict lexicon and language model to the utterances that are reasonable considering the agent. Obviously, we should try to visualize agents in such a way that it is clear from their appearance what they're responsible for and what a visitor can ask them. An attempt should be made to ensure that any agent is able to determine that she isn't the right agent to answer a visitor's questions and therefore should direct the visitor to an other task-oriented agent or to an agent having global knowledge of the task-oriented knowledge of the other agents in the virtual environment.

³ Notice that in our database with theatre information we also have reviews of performances which can be rather long and which can be read to the visitor by Karen.

6.3 Embodied Agents that Sell, Advise, Buy, . . .

In the case of Karen, it will be clear that her boss would like her to sell as many tickets as possible. A theatre director will have certain preferences in choosing performances for the theatre, but once they have been chosen the aim is to be sold out every night. How can we program Karen (and maybe also the navigation agent) such that the behavior towards a visitor increases the chance to reach this 'private' goal?

In the future, webpages and virtual environments will become inhabited by sales agents, hosts, guides, etcetera that offer information and answer questions (and try to sell products) in a way similar to Karen or our navigation agent, but also in a way that allows building up relationships with users through social "chit-chat" about family, work or sports. That is, these agents should have specific knowledge about a certain domain, a domain which may range from detailed knowledge about Shakespeare, performances in a next season and cars in a showroom to drinks that are available in a bar. But, in addition to that, they have some superficial global knowledge that allows them to give socially appropriate answers to keep up a conversation about topics they hardly know of and they have some special hobbyhorses and some specific knowledge with which they can steer a conversation and which makes them 'believable' to the user.

One of our concerns in the near future will be the introduction of such properties in our present agents. Maybe Karen shouldn't be allowed to give too personal opinions about performances and artists, but some more human-like conversational behavior should be considered. Moreover, if a user really wants to know or to exchange opinions about performances and artists she should be able to communicate with other visitors of the environment or be able to address agents employed by the theatre who can share their specific knowledge (embedded in some kind of social conversation) with the visitor.

Agents that allow human-like conversation have been designed, both in research and in commercial environments. An example of a virtual web agent is Jennifer James, designed by *Extempo Systems Inc.* She sells cars in a virtual auto show room, 24 hours a day. Jennifer has a history in racing, which is useful in talking with customers. She wants to know about a customer's background and preferences. Questions are asked in a friendly and sometimes ironic way. Jennifer knows about cars and racing, but questions or comments on family or country music can be dealt with in a believable way. Jennifer has been visualized as a smiling saleswomen dressed in a red and white jump suite. Jennifer has personality, customers feel confident to talk with her, they give information about themselves and a company that employs Jennifer will build relationships of affection, trust and loyalty with its customers. The information that is elicited from a customer is stored in marketing and customer databases.

Jennifer makes a charming attempt to understand natural language and to interact with objects in the scene. The customer uses the keyboard to communicate with Jennifer. Her body and face are real-time animated, where animations attempt to be consistent with role and personality, and with the events of the dialogue. Jennifer talks back using speech synthesis. More examples of agents that display human and social qualities have been introduced. Of course, we can go back to 'conversational agents' such as Weizenbaum's Eliza and Colby's Parry (both from the sixties) or Julia, a chatterbox character of the early nineties. However, these characters have not been given the task to inform visitors, to give information, to guide and help or to purposely try to seduce potential buyers to buy commercial products, to visit sites and events and to take part in activities.

7 MM-Retrieval in the VR Environment

We investigate how to store, index and retrieve theatre-related multi-media objects (text, pictures, audio, video) in such a way that a visitor of our virtual theatre environment can address Karen in such a way that she is not only able to inform the user about artists, performances, dates and available seats, but also knows how to provide the user with pictures, audio and video fragments (previews) of performances. In general, if a visitor asks about a particular performance, it is not difficult to let Karen present the (available) associated multi-media information. However, if the visitor's questions are about or address associated information rather than explicitly making references to a particular performance, then our theatre agents have to start reasoning with the information that is available in the database and information that is available in a knowledge representation scheme in which general knowledge about the theatre domain has been stored.

At this moment no explicit, comprehensive and suitable ontology of the theatre information and booking domain has been designed. Karen knows about this domain, but her knowledge is hidden in the linguistic (syntactic, semantic) and the dialogue management (pragmatic) knowledge of the present system. This knowledge nevertheless allows a better mapping on the information that is available in the theatre database. However, other agents have no access to this implicit knowledge. Similarly, our navigation agent knows about the geography of the theatre and how users generally ask questions about the theatre. However, knowledge about the geography has been represented as a list of coordinates of particular locations and possible ways to refer to them. Hence, the user has some freedom how to address these locations. However, whatever synonyms are allowed, the list is predefined and only the navigation agent knows how to access this information. Our speech recognition software, that is, the number of phrases it can recognize reasonably correctly in normal circumstances, does very much determine the navigation's agent intelligence.

How to handle a user's question about artists and performances, where the user doesn't use names that are available in the theatre database? That is, can we provide Karen with general knowledge about artists and performances which goes beyond the information that is available in the database? , and, if it turns out that the user doesn't know about Karen's knowledge and such knowledge is not complete, can we nevertheless map the user's question to a database query?

As an example, suppose someone asks: "Well, what's her name, Michelle, eh, I know she received an Oscar last week. Are there any movies with her showing this week?" At this moment we think that it is reasonable to expect from our system that it starts searching WWW with a question that has as keywords: 'Oscar' and 'Michelle'. The intelligent theatre search engine should deliver the name of the actress and this should be sufficient to search the database for movies showing this week (or maybe later this month) with this particular actress in one of the roles. A knowledge representation network with the main concepts and their relationships in the world of theatre (opera, musicals, plays, music, cabaret, etc.) and theatre actors (directors, performers, technicians, reviewers, etc.) has to be designed in order to connect information available on WWW and information available in local databases through a process of reasoning.

Textual information is important, since it is widely available. Newspapers, magazines and WWW pages contain reviews about performances, movies and performers. However, there is no need to confine ourselves to textual information. We are involved in national and European projects on indexing and retrieval of multi-media information, including pictures, captions, audio, video, subtitles and (transcripts of) spoken movie dialogues. Integrating this research on multimedia indexing and retrieval and natural language (spoken) dialogue systems embedded in environments where users can 'say what they see' (ranging from handheld devices to immersive virtual reality environments) is a topic high on our research agenda. We think it is fruitful to consider WWW as an extension of our theatre databases and therefore we need an intelligent agent that knows how to filter and present the results of a WWW search to a naive user looking for some information that may help in having a nice evening.

8 Presence and Believable Agents

'Presence', as defined by Lombard and Ditton [10], is the perceptual illusion of nonmediation, that is, 'a person fails to perceive or acknowledge the existence of a medium in his/her communication environment and responds as he/she would if the medium were not there.' They mention that this illusion can occur in two distinct ways:

- The medium can appear to be invisible, with the medium user and the medium content sharing the same physical environment; and
- The medium can appear to be transformed into a social entity.

From the previous sections it may have become clear that rather naturally emerging topics of our interest are closely related to the issue of presence. The environment that is offered and the locations that can be visited look familiar, the functions of several objects and what to do with them is clear from their appearance and the multimodality approach allows a variety of user input and the production of different sensory outputs. The agents in the environment are assumed to be friendly and cooperative and the embedding of talking faces and moving avatars in the environment will increase the tendency to treat agents as social actors, rather than as software modules.

We have looked at possibilities to increase a user's commitment to the system (like we would in similar systems, e.g., for electronic commerce) with the aim to obtain co-operative behavior. One obvious reason which makes us loose a user is when use of imperfect technology is not sufficiently backed up by context (including other modalities) which seduces the user to a certain interaction behavior and helps to disambiguate the users utterances.

In our world mediated entities (our agents in the virtual world) play the role of social actors, they increase the feeling of presence and they help to increase the user's commitment to the system. A more technical reason to have agents as social actors is that they can influence the interaction behavior of the users in such a way that it remains restricted to the task and/or domain knowledge, therefore hiding shortcomings of imperfect interaction technology, in particular speech and language technology. With other words, in order to increase the quality of the web-based information and transaction services we are offering, it seems to be useful to exploit the possibility to increase the role of social actors in our environment.

It should be clear from previous sections that in our environment we decided to explore the possibility to increase the user's feeling of presence in order to increase his or hers commitment to the system. We certainly do not want to advocate such an approach in general. In our application we think this is a useful approach. We don't think this should be the general approach to the design of computer interfaces. Is the interface serving a leisure activity, is it task-oriented or is it assumed to change the user?

9 Conclusions

In this chapter we reported about on-going and future research on interactions in virtual worlds. We intend to continue with the interaction between experimenting with the virtual environment (adding agents and interaction modalities) and theoretic research on multi-modality, formal modeling, natural language and dialogue management. In particular the integration of visitors in an agent platform that models a uniform verbal and non-verbal behavior is required in order to be able to maintain and extend the virtual environment.

Our approach to designing a virtual environment for an interest community is bottom-up. At this moment the system has two embodied agents with different tasks and limited interaction between them. Moreover, the agents do not employ a model of a user or of user groups. In general, when we talk about interface agents we mean software agents with a user model, that is, a user model programmed in the agent by the user, provided as a knowledge base by a knowledge engineer or obtained and maintained by a learning procedure from the user and customized according to his preferences and habits and to the history of interaction with the system. In this way we have agents that make personalized suggestions (e.g. about articles, performances, etc.) based on social filtering (look at others who seem to have similar preferences) or content filtering (detect patterns, e.g. keywords) of the items that turn out to be of interest to the visitor. In this way, visitors can be provided with personal assistants ('butlers') that know about the visitors' preferences, that can exchange information with other personal assistants and that can search for and filter information that is of interest for the visitor. In particular we hope to integrate research that comes available from the European projects *Magic Lounge* (which aims at developing tools that allow intelligent communication services in virtual meeting places; these tools include shared white boards and chat environments, but also tools that record and store interaction events such that it becomes possible to browse earlier interactions and inspect individual contributions) and *PERSONA* (a project about navigation in two- and three-dimensional interfaces; in this project the concept of social navigation is explored, that is, navigation that exploits the possibility to talk to other users and to agents for obtaining information, including the following of their trails in the information space).

Contrary to many other virtual environments the public version of the environment is in use by the general audience. The environment contains actual information about the theatres, it is accessible on WWW and in addition it is part of an exposition in a technology activity center where visitors can get explanation about the environment.

References

- 1 Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S. & M. Stone. Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Proceedings of *SIGGRAPH '94*, 1994.
- 2 Cassell, J. & K.R. Thórisson. The power of a nod and a glance: envelope vs. Emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, to appear.
- 3 Cassell, J., O.E. Torres, S. Prevost. Turn taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In: Wilks (ed.), *Machine Conversations*. The Hague: Kluwer, to appear.
- 4 Dirksen, A. and Menert, L. Fluent Dutch text-to-speech. Technical manual, Fluency Speech Technology/OTS Utrecht, 1997.
- 5 Doest, H. ter. *Towards Probabilistic Unification-Based Parsing*. Ph.D. Thesis, University Twente, February 1999.
- 6 Friedman, B. (ed.). *Human Values and the Design of Computer Technology*. CSLI Publications, Cambridge University Press, 1997.
- 7 Hulstijn, J. & A. van Hessen. Utterance Generation for Transaction Dialogues. Proceedings *5th International Conf. Spoken Language Processing (ICSLP)*, Vol. 4, Sydney, Australia, 1998, 1143-1146.
- 8 Jönsson, A. *Dialogue Management for Natural Language Interfaces*. PhD thesis, Linköping University, 1993.
- 9 Lie, D., J. Hulstijn, A. Nijholt, R. op den Akker. A Transformational Approach to NL Understanding in Dialogue Systems. Proceedings *NLP and Industrial Applications*, Moncton, New Brunswick, August 1998, 163-168.
- 10 Lombard, M. & T. Ditton. At the heart of it all: The concept of presence. *Journal of Mediated Communication* 3, Nr.2, September 1997.
- 11 Nass, C., B. Reeves & G. Leshner. Technology and roles: A tale of two TVs. *Journal of Communication* 46 (2), 121-128.
- 12 Sproull, L., M. Subramani, S. Kiesler, J. Walker & K. Waters. When the interface is a face. In [6], 163-190.