The Twente Virtual Theatre Environment: Agents and Interactions

Anton Nijholt (Parlevink Research Group)[1]
Centre for Telematics and Information Technology (CTIT)
University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

ABSTRACT

In this paper we discuss our research on interaction in a virtual theatre. It has been built using VRML and therefore can be accessed through Web pages. In the virtual environment we employ several agents. The virtual theatre allows navigation input through keyboard and mouse, but there is also a navigation agent which listens to typed input and spoken commands. Feedback of the system is given using speech synthesis. We also have an information agent which allows a natural language dialogue with the system where the input is keyboard-driven and the output is both by tables and template driven natural language generation. In development are several talking faces for the different agents in the virtual world. At this moment an avatar with a cartoon-like talking face driven by a text-to-speech synthesizer provides users with information about performances in the theatre. 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. Societal and ethical implications of VR environments are discussed. Moreover, we spend some notes on real-time and video performances in our virtual theatre.

Keywords: Virtual Reality, Talking Faces, Text-to-Speech Synthesis, Agent Technology, Speech Recognition, Presence, Societal Aspects, Electronic Commerce, Web Theatre

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 by allowing users access to text, tables, figures, pictures, audio, animation and video. In a task- or domain-restricted 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, the interesting parts of the Web will become three-dimensional, allowing the building of virtual worlds inhabited by interacting user and task agents, and 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 by users.

In this paper we present our research on the development of an environment in which users can display different behaviors 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, both goals, possibilities and the way they influence each other become clear. One way to support

[1] This paper describes research carried out in the Parlevink research group of the CTIT (Centre of Telematics and Information Technology) of the University of Twente. The following members of the group have contributed to this paper: Joris Hulstijn, Arjan van Hessen, Hendri Hondorp, Danny Lie, Mathieu van den Berk and Boris van Schooten.

such users is to give them different interaction modalities and access to multimedia information. We discuss a virtual world for representing information and allowing natural interactions that deal with an existing local theatre, and of course, in particular, the performances in this theatre. The interactions between user (visitor) and the 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, eye-brows, eyelids and some changes in face color), but at this moment these possibilities have not yet been associated with utterances of user or system.

It is 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, but also 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.

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. This 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. The system is rather primitive. 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? Clearly, the latter 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 the Dutch public transport and railway information system. Human operators are available to inform about times and schedules of busses and trains. However, the number of operators is insufficient. Callers can wait (and pay for the minutes they 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 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, e.g. 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 and efficient ways to get or get done what they want.

3 PROVIDING CONTEXT: ACCESS AND EXPLORATION

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. Research done on information access and transaction in the context of modalities (and especially the sequential and simultaneous combination of modalities), that is, a multi-modality approach for WWW, can be embedded in the attempts to develop standards for access to the web and presentation on the web. For example, standards are being developed for speech access (voice browsing), 3D visualization (virtual reality modeling languages) and the combination of access and visualization (MPEG standards).

Figure 1: Entrance of the Virtual MUZIEKCENTRUM

When we look at multi-modal human-computer interaction it is clear that 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.

Our second observation, certainly not independent from the observation above on modalities for access, exploration and presentation, deals with the actors in a system that has to deal with presenting information, reasoning about information, communicating between actors in the system and realizing transactions (e.g. through negotiation) between actors in the system. In addition to a multi-modality approach, there is a need for a multi-agent approach, where 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.

Both multi-modality and multi-agent technology can be considered from a cognitive science point of view, an artificial intelligence point of view or a computer science (i.e., design, algorithmic & data structures) point of view.

At this moment the cognitive science point of view, at least at our side, is rather undeveloped. The ideas that are available on the cognitive science point of view deal with syntax, semantics and pragmatics of natural language communication. That is, although we would like to see it differently, they are more closely related to linguistics than to cognition science in general. On the other hand, some modest approaches to include concepts of cognitive science in the definition and the behavior of agents are available and cognitive ergonomics helps to design user interfaces and interaction modalities for given tasks and users.

From the artificial intelligence point of view we know we can use results on domain-independent and domain-dependent representation and reasoning. Frame- and script-based methods in AI are available and compromises have been established between cognitive science, artificial intelligence and computer science, in order to design and develop useful applications. From the computer science point of view we can discuss methods for design, specification and implementation of multi-modal and multi-agents systems. In the next sections we will return to these topics. The roles of speech, language and visualization will be emphasized.

4 PROVIDING CONTEXT: VISUALIZATION

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, it allows the system to influence the interaction behavior

of the user in such a way that more efficient and natural dialogues with the system become possible.

Our virtual theatre has been built according to the design drawings made by the architects of a 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 hall, admire the paintings on the walls, go to the balconies and, take a seat in order to get a view of the stage from that particular location. Information about today's performances is available on a notice 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 and start a dialogue with an information and transaction agent called 'Karin'. The first version of Karin looked like other standard avatars available on World Wide Web. The second version, available in a recent prototype of the system, makes use of a 3D talking face.

One may dispute the necessity of this realistic modeling of theatre and agents, the environment and the information and transaction services. We have taken the point of view that (potential) visitors are interested in or are already familiar with the physical appearance of this theatre. Inside the virtual building there should be a mix of reality (entrance, walls, paintings, desks, stages, rooms, etc.) and new, non-traditional, possibilities for virtual visitors to make use of interaction, information, transaction and navigation services that extend the present services of the theatre.

It has become clear from several studies (cf. Friedman [9]) 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. They can be made to enjoy the interaction, they can be made to perform better, etc., all depending on the way the interface and the interaction strategy have been designed. It makes a difference to interact with a talking face display or with a text display. People behave differently in the presence of other people than they do when they are alone. 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. [28]).

From these observations we conclude that introducing a talking face can help to make interaction more natural and to make shortcomings of the technology more acceptable to users.

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. Also real dialogue systems wherein people can say normal phrases become more and more common, but one of the problems in this kind of 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.[2]

5 PROVIDING CONTEXT: COMMUNICATION

In the previous subsections we have looked at possibilities for users 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. It is also interesting to investigate how we can allow communication between users or visitors of a web-based information and transaction system. For that purpose it is useful to look at experiences with web-based digital cities, chat environments and interest communities.

Web-based digital cities have been around for some years. Like computer games they have evolved from text environments to 2-dimensional graphical and 3D virtual environments with sounds, animation and video.

[2] It may be the case that different specialist agents are taken more seriously than one generalist agent. See Nass et al. [18] who report about different appreciation of television programs depending on whether they were presented in a 'specialist'or in a 'generalist'setting.

Visitors, or maybe we should call them residents, of these cities visit libraries, museums, pubs, squares, etc., where they can get information, chat with others, etc. In these environments people get the feeling of being together. They are listening to each other and, in general, they take responsibility for the environment

Figure 2: Lobby of the Hutch World

Today there are examples of virtual spaces that are visited and inhabited by people sharing common interests. With virtual spaces or environments we want to refer to computer accessible environments where users (visitors, passers-by) can enter 3D environments, browse (visual representations of) information and can communicate with objects or agents (maybe other visitors in the same environment). These spaces can for example, represent offices, shared workspaces, 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 or educational) interests, but share some common conditions (driving a car, being in hospital for some period, have the same therapy, belonging to the same political party, etc.).

As an 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 pass-word protected three-dimensional world (a rendering of the actual outpatient lobby), get information at a reception desk, visit a virtual gift shop, etc. (Figure 2). 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. Part of the project consists of developing tools to create other applications.

6 Agents in the Twente Virtual Theatre

6.1 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 typed dialogue. 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.

Figure 3: Karin at the Information Desk

6.2 THE INFORMATION & TRANSACTION AGENT

Karin, the information/transaction agent, allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karin (Figure 3) wants to give information and to sell tickets. Karin is fed from a database that contains all the information about performances in our local theatre. Developing skills for Karin, in this particular environment, is one of the aims of our research project. This research fits in a context of much more general 'intelligent' (web-based) information and transaction services.

Our current version of the dialogue system of which Karin is the face is called THIS v1.0 (Theatre Information System). The approach used can be summarized as rewrite and understand. 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. 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 (DBMS). Queries to the database are represented using a standard query language like SQL. Results of queries are represented as bindings to variables, which are stored in the global data-structure, called context. The arguments for the action are dug out by the dedicated parser, associated with the category. 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 [15].

Presently the input to Karin is keyboard-driven natural language and the output is both screen and speech based. In development is an utterance generation module. Based on the most recent user utterance, on the context and on the database, the system has to decide on a response action, consisting of database manipulation and dialogue acts.

6.3 THE NAVIGATION AGENT

The WWW-based virtual theatre we are developing allows navigation input through keyboard and mouse. Such 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. In addition, a navigation agent has been developed that helps the user to explore the environment and to interact with objects in this environment by means of speech commands. A smooth integration of the pointing devices and speech in a virtual environment requires has to resolve deictic references that occur in the interaction. The navigation agent should be able to reason about the geometry of the virtual world in which it moves. The current version of the navigational agent is not really conversational. Straightforward typed commands or similar speech commands make it possible for the user to explore the virtual environment. Navigation also requires that names have to be associated with the different parts of the building, the objects and the agents, which can be found inside of it. Clearly, users may use different words to designate them, including implicit references that have to be resolved in a reasoning process.

Speech Recognition on local machines is pretty good, but speech recognition on the World Wide Web results in various problems. Many of these problems are caused by the lack of standards and the lack of interest of big companies (providing operating systems, WWW browsers and Virtual Reality languages and environments) to cooperate in order to establish standards. When we confine ourselves to speech recognition, we distinguish between two approaches.

- First Solution: Every user should have a speech

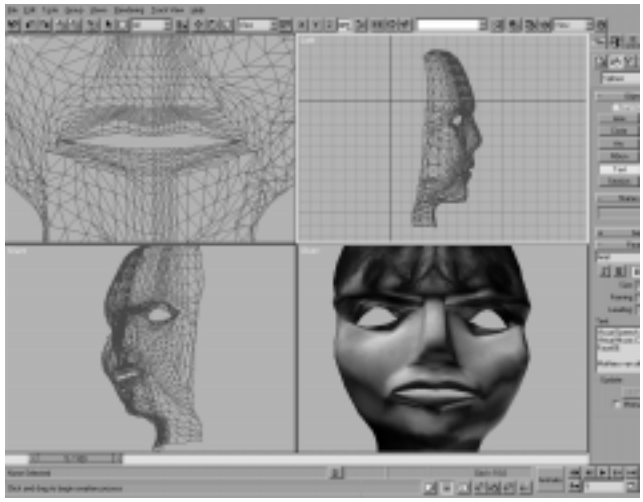

Figure 4: 3-D Face Development

Figure 5: Cartoon Face

recognition engine that can recognize their commands and send this information to the server system. However, good speech recognition systems are very expensive, require large resources and bad systems result in badly recognized commands.

- Second Solution: Another solution would be to have the speech recognition on the server side. This requires the recording of commands on the client side and a robust transporting of the audio files.

In our system we have chosen for the second solution. 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. They do need audio-software which is usually available anyway.

7 SPEECH GENERATION AND ANIMATION

7.1 SPEECH GENERATION THROUGH TEMPLATES

In the design of utterance generation by the information agent a list of utterance templates is used. Templates contain gaps to be filled with information items: attribute-value pairs labeled with syntactic and lexical features. Templates are selected on the basis of five parameters: utterance type, the body of the template and possible empty lists of information items that are to be marked as given, wanted and new. The utterance type and body determine the word-order and the main intonation contour. The presence and number of information items in the given, wanted and new slots, as well as special features affect the actual wording and intonation of the utterance.

For pronouncing the utterance templates we use the Fluent Dutch Text-to-Speech system (Dirksen [7]). Fluent Dutch runs on top of the MBROLA diphone synthesizer (Dutoit [8]). It uses a Dutch voice, developed at the Utrecht institute of linguistics (OTS). Fluent Dutch 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. For many words, the phonetic description is taken from lexical resources of Van Dale dictionaries. 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.

7.2 FACING THE INFORMATION AGENT

We developed a virtual face in a 3D-design environment (cf. Figure 4). The face consists of various three-dimensional coordinates and is connected through faces. These faces are shaded to visualize a three-dimensional virtual face. The 3D data is converted to VRML-data that can be used for real-time viewing of the virtual face. A picture of a real human face can be mapped onto the virtual face. We are researching various kinds of faces to determine which can be best used for this application. Some are rather realistic and some are more in a cartoon-style (cf. Figure 5). 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.

The face is capable of visualizing the speech synchronously to the speech output. This involves lip-movements according to a couple of visemes. The face has to visualize facial expressions according to user's input or the system's output (see further sections on facial features). Figure 6 represents the architecture of the visual speech system.

The last element in the chain of this figure (the VRML-browser) is also the first element. We use Cosmo Player, which is a plug-in for an HTML-Browser, for viewing VRML-files. These files are specifications of a three-dimensional virtual environment. The whole virtual theatre is a collection of VRML files, which can be viewed by the browser. As mentioned earlier, the user will see a virtual face when the information desk is

approached. A dialogue window, the JAVA Schisma applet, is available for the user to formulate questions or to give answers to the system's questions. The user types the questions on a keyboard in Dutch sentences. The answers to the questions are to be determined on the server side: the Schisma server. Answers or responding questions are passed to the JAVA Visual Speech Server Application on the server side.

This application filters the textual output of the dialogue system in parts that are to be shown in a table or a dialogue window and parts that have to be converted to speech. The parts that are to be shown in the dialogue window or a table, like lengthy descriptions of particular shows or lists of plays are send to the Schisma Client Applet where they are showed on the screen. The parts of the Schisma output that are to be spoken by the virtual face are converted to speech with the Text-to-Speech Server. The input is the raw text and the output is the audio file of this spoken text and information about the phonemes in the text and their duration.

For example, the Dutch word for "speech generation" is "spraakgeneratie". This word contains the following phonemes: S p r *a k x e n @ r a t s I. When the resulting audio file is played, each phoneme has it's own duration. This information is gathered from the

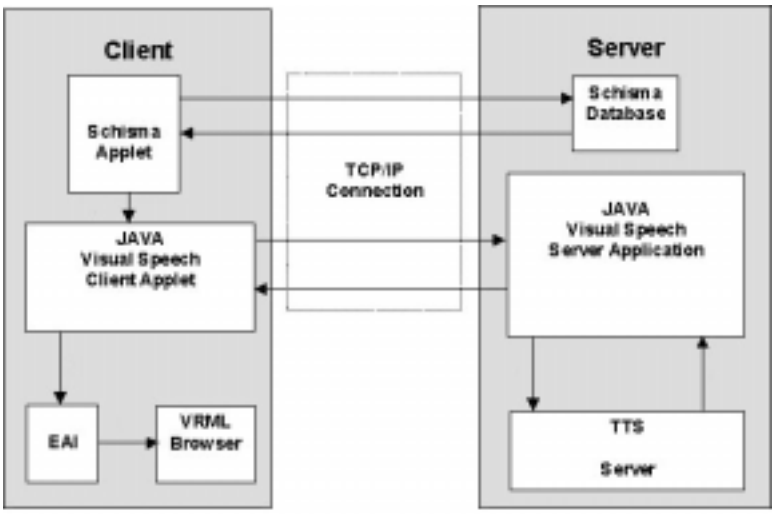

Figure 6: Architecture Visual Speech System

TTS-server:

s 79 p 71 r 38 a 106 50 127 k 53 x 90 e 113 20 102 n 60 @ 38 r 53 a 101 t 23 s 113 I 119 20 75

The characters are the phonemes and the first number after the characters are durations of the corresponding phonemes in milliseconds. If more numbers follow then the first number is a percentage of the whole duration in which the pitch of the voice changes to the following number. So the first 'a' is spoken for 106 milliseconds and on 50% of this 106 milliseconds the pitch changes to 127 Hz. The previously described information from the TTS-server will be sent to the JAVA Visual Speech Client Applet together with the audio file. The Visual Speech Client Applet uses the phoneme information to map the phonemes onto different mouth states or visemes. All the phonemes are categorized in five visemes (cf. Figure 7).

p	**P**ot	D	**D**ak	t	**T**ak	S	**Sj**aal	f	**F**eest
b	**B**ak	k	**K**oe	tj	**Tj**arda	H	**H**uis	v	**V**eel
m	**M**ok	g	**G**oal	dj	**Dj**oerd	j	**J**as	w	**W**el
		n	**N**at	z	**Z**on	J	A**ai**	W	E**euw**
		nj	Ra**nj**a	s	**S**la	y	F**uu**t		
		i	P**i**t	Z	Ra**g**e	2	K**eu**s		
		E	P**e**t	G	**G**even	Oe	Fr**eu**le		
		Y	P**u**t	x	**Ch**aos	Y:	K**eu**r		
		A	P**a**d	N	Ba**ng**	O:	K**oo**r		
		O	P**o**t	l	**L**at	E~	**Tim**bre		
		@	G**e**val	L	Ka**al**	A~	Ch**an**son		
		i	N**ie**t	a	**Taa**k	Y~	Par**fum**		
		R	Paa**r**	e	K**ee**t	O~	**Bonbon**		
		r	**R**aak	o	P**oo**k				
		u	B**oe**k	l:	K**ee**r				
				Ei	**Ti**jd				
				9y	H**ui**s				
				Au	K**ou**d				

Figure 7: Phoneme Classes and Visemes

When the audio file is loaded on the client side, the mouth states and their durations are

passed to the External Authoring Interface (EAI). This is an interface between JAVA and the VRML browser. This interface triggers animations in the virtual environment. It starts the sound playback and all the corresponding animations. Only the mouth states are specified in the VRML-file. The animation is done by interpolating between mouth states in the given amount of time. This results in reasonable smooth lip-movements.

7.3 PROSODY, FACIAL EXPRESSIONS AND EMOTIONS

How do we control the responses of the system, the prosody and the artificial face? The central module of a dialogue system is called the *dialogue manager*. The dialogue manager 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 or signal, such as a pointing gesture, the dialogue manager selects a certain response action. Planning and action selection are based on a set of principles, called *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. Now, it should also control the orientation and expression of the face, the eyes, and the coordination of sounds and lip movement. What parameters are needed to control response generation?

7.4 PROSODIC FEATURES

The spoken utterance generation module uses a set of parameters to control prosodically annotated utterance 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 by the 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 texts and reviews that describe the content of performances, the system assumes a 'reading voice'.

7.5 FACIAL FEATURES

Apart from the lips, the virtual face has a number of dynamic control parameters (Figure 8).

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

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

Figure 8: Facial Features

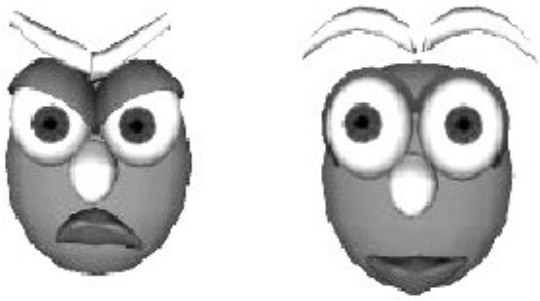

Figure 9: Angry and Uncertain

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, Clermont et al. [5] found that a fixed stare indicates 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 (see Figure 9). Jerky movements with wondering eyes indicate nervousness. Since our agent is supposed to be 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.

So how do we control this wealth of features? We propose a blackboard architecture, as depicted in Figure 10. Combinations of input parameters trigger a rule that produces a response action, or a more permanent change of expression. The reason for a blackboard architecture is that the parameters influence each other. Roughly, there are two types of facial 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 (Vertegaal [30])[3]. Usually mood and attention are fixed for a given personality. Temporary changes in emotion and attention, may result from previous utterances or to the general conversation. For instance, anger at an insult, or increased interest after a misunderstanding.

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

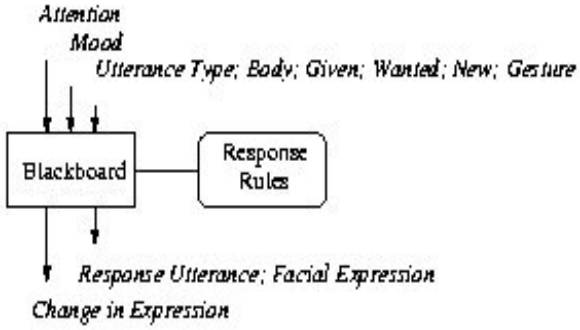

Figure 10: Blackboard Architecture

[3] Experiments with different video conferencing environments have shown that gaze information is more important to a smooth conversation, than a simple television type image of the talking person, when it is unclear what he or she is looking at.

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.

We also 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.

8 SPEECH, ANIMATION & WEB-BASED STREAMING AUDIO AND VIDEO

At this moment VRML97 is the standard specification for VRML. This specification allows the definition of AudioClip Nodes in a VRML world. AudioClip Nodes have stereometric properties, that is, the volume of sound increases when approaching a sound object and when a user moves in the world the sound will adapt. Hence, when the user moves to the right the volume in the left speaker will increase and the volume in the right speaker will decrease. The present standard AudioClip Nodes are uncompressed WAV and MIDI, respectively.

Our experiences until now show that the use of uncompressed WAV slows down animation considerably (often 30 seconds or more) because the WAV file has to be written to the hard disc by the TTS Server and both the EA (External Authoring) Interface and the VRML browser have to read this file completely before the animation can be started. Short sentences hardly cause problems, but long texts often take 300 kB or more.

It is investigated whether it is possible to process the audio output of the TTS Server in such a way that a compressed audio stream can be created that can be synchronized with the VRML animation. When we look at RealAudio compression/streaming there are the following disadvantages:

- For the VRML browser this is not a standard format. However, maybe we can deal with this since it is possible to add unknown Nodes to a VRML world using properties of existing Nodes. In his way it is possible to define a RealAudio Node which can be driven with the help of RJMF (RealAudio Java Media Framework).
- The VRML browser claims the full Audio System of the PC if it has to use WAV files; therefore it is not possible for applications outside the VRML world to play audio fragments. This would mean that all WAV files to be used in the VRML world have to be played as mentioned above.
- If outside the VRML world an Audio file is played then the stereometric properties of the sound can not be modeled in the VRML world. A partial solution can be obtained since the EAI can observe changes of position and the RJMF can change the volume of the sound. Whether it allows implementation of stereo sound is not clear yet.

Roehl [25] discusses audio streaming in VRML. He argues for a standard way for a content creator to indicate to the browser that the data in an AudioClip should be streamed, rather than being completely downloaded prior to being presented to the user. Moreover, whenever possible, we should use existing open standards. Examples are RTSP, SMIL and RTP.

RTSP (Real Time Streaming Protocol (RFC 2326)) is an existing draft Internet standard for accessing streaming media. The content creator is able to identify which data should be streamed by specifying "rtsp:" as the scheme in the URLs instead of "http:" or "ftp:", so the browser should use RTSP to obtain data for that node. RTSP does not specify the use of any particular transport mechanism for the actual streaming data itself. RTP (the Real Time Protocol (RFC 1889)) does. It is an application level protocol for the transport of streaming multimedia. Synchronization of the audio and video data can be achieved using the timestamp information provided in the RTP headers. Part of the RTP standard is a separate protocol RTCP (Real Time Control Protocol) which, among other things, provides

NTP based timestamps for the purpose of synchronizing multiple media streams.

An important recent development is SMIL (Synchronized Multimedia Integration Language), a proposed World Wide Web Consortium Standard. SMIL is HTML-like and describes multiple sequential or concurrent media streams along with timing information. Hence, it allows the synchronization of Audio/Video files with other events. We have to investigate how VRML fits in this development and we plan to investigate whether it is possible to generate SMIL information from the phoneme output of the Text-to-Speech Server. Together with the RTP and RTSP transport mechanisms it should be possible to obtain exact synchronization with the help of the timestamps in NTP.

In Osterman et al. [20] it is explained how MPEG4 can specify a set of face animation parameters and a sound output interface such that phonemes and their timing (obtained from the TTS) can be translated into sequences of facial animation parameters and there is synchronization between speech and animation. The TTS system is explained in Beutnagel et al [4]. We have not yet investigated their approaches in detail.

9 Formal Modeling of Interactions in Virtual Environments

Both from an ergonomical and a software-engineering viewpoint, the design of interaction in virtual environments is complex. Virtual environments may feature a variety of interactive objects, agents which may use natural language to communicate, and multiple simultaneous users. All may operate in parallel, and may interact with each other concurrently. Next to this, the possibility of using Virtual Reality techniques to enhance the experience of virtual worlds offers new ways of interaction, such as 3D navigation and visualization, sound effects, and speech input and output, possibly used so as to complement each other.

One new line of research we have taken is an attempt to address both of these issues by means of a formal modeling technique that is based on the process algebra CSP (Communicating Sequential Processes). For that reason, in our virtual theatre a simplified flow of interaction has been specified, showing all relevant interaction options for any given point in time. The system architecture has been modeled in an agent-oriented way, representing all system- and user-controlled objects, and even the users themselves, as parallel processes. The interaction between processes is modeled by signals passing through specific channels. Interaction modalities (such as video versus audio and text versus graphics) may also be modeled as separate channels.

This modeling technique has some strong points. Firstly, and most generally, such a simplified and formal model enables a clear and unambiguous specification of system architecture and dynamics. Secondly, it may be useful as a conceptual model, modeling the fact that a user experiences interaction with other users and agents in a similar way than in a completed system, and explicitly showing which options are available when and through which modalities. Thirdly, it enables automatic prototyping, such as architecture visualization and verification of some system properties.

For more details about this approach we refer to Schooten et al. [26]. There it is also shown how a CSP description can be coupled to a simplified user interface and executed, so that the specified system can be tried out immediately. Specifications map closely to software architecture, reducing the cost of building a full prototype.

10 The Role of Presence

'Presence', as defined by Lombard and Ditton [16], 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:

(a) The medium can appear to be invisible, with the medium user and the medium content sharing the same physical environment; and

(b) The medium can appear to be transformed into a social entity.

The authors show that many conceptualizations of presence that appear in the literature are more detailed viewpoints on this distinction. These viewpoints are:

- Social richness. Is the medium perceived as sociable, warm, personal, etc., in the interaction? How does the medium transmit cues of human communication?
- Realism. How real is the experience, how accurate is the representation of objects?
- Transportation. Does the user have the feeling that he/she has been transported to a different place (or

the place to the user), does the user have the feeling to share the place with others?

- Immersion. How much are the user's senses immersed in the virtual world, how much is a user involved in the interaction with the medium?
- Social actors within a medium. Do mediated entities play the role of social actors? That is, do users overlook the mediated and artificial nature of entities within a medium with which they interact?
- Medium as social actor. Do users respond to the medium as a social entity? Do they follow social rules in the interaction?

Other definitions of presence exist, often closely related as in Witmer et al. [31] ("presence is defined as the subjective experience of being in one place or environment, even when one is physically situated in another"), sometimes using new names ('telepresence'), sometimes introducing new viewpoints, e.g., in Zahorik et al. [33], where in the tradition of Heidegger presence is strongly related to one's successfully supported action in an environment.

Causes and effects of presence have been studied in experiments, but the number of parameters involved, e.g., task and user characteristics, makes systematic research difficult.[4] We mention Thie et al. [29], who attempt to measure the effects of social virtual presence in a decision making task. Social virtual presence is the feeling that other people are present in the virtual environment. Their hypothesis was that presence will be higher if social presence cues are maximized and that decision task performance improves if presence increases. While the first part of this hypothesis was supported by their experiments, due to lack of good measuring questionnaires, technical malfunctioning, etc., no useful results could be reported about an increase of performance.

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.

[4] Papers on presence can be found in the journal: *Presence: Teleoperators and Virtual Environments*, MIT Press.

However, increasing the user's feeling of presence has not been an explicit goal. Rather 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 clumsy technology (like speech and language technology) is not sufficiently backed up by context (including other modalities) which seduces the user to a certain interaction behavior and which helps to disambiguate the users utterances.

Needless to say that for many tasks and activities one may not expect that presence will necessarily enhance performance. On the contrary, presence can be noise which distracts a user from performing a task and which throws away the advantages of individual access from a home PC to, e.g., an information service. There is no need for a virtual waiting queue or the sound of a leaving plane when all we want to know is when the next flight to Amsterdam leaves.

11 SOCIETAL AND ETHICAL ASPECTS

Long time ago, in the early eighties, Terry Winograd mentioned that in presenting machines as 'intelligent' we produce an illusion which may be beneficial, may lead to breakdown in the interaction, or may be used by parties to deceive and exploit others. His concern was illustrated with the following quote (from a researcher at a major computer firm):

> *"From my point of view, natural language processing is unethical for one main reason: It plays on the central position which language holds in human behavior. I suggest that the deep involvement Weizenbaum found some people to have with ELIZA, is due to the intensity with which most people react to language in any form. When a person sees a linguistic utterance in any form, the person reacts, much as a dog reacts to an odor. We are creatures of language ….. Since this is so, it is my feeling that baiting people with strings of characters (tokens) clearly intended by someone to be interpreted as symbols, is as much misrepresentation as would be your attempt to sell me property for which you had a false deed. In both cases, an attempt is made to encourage someone to believe that something is a thing other than what it is, and only one party in the interaction is aware of the deception."*

We would like to add a more recent quote (late nineties) by Ben Shneiderman (head of the HCI-Lab of the University of Maryland):

"Designers should restrain themselves from exploiting feelings that can be evoked by machines. Computers should not be laden (!) with emotion."

Interesting it is to observe that in the early eighties the emphasis was on natural language, and in the nineties the emphasis is on feelings and emotions. More interesting for our purposes is to note that in both quotations the authors assume that computers can be used to deceive and exploit users and that that is wrong.

One reaction to that may be: What's wrong about that? Newspapers do, PR departments do, political parties do and advertisements do. What makes computers more special than newspapers that they shouldn't be used for those purposes? However, even if we accept such a view – or understand that we have to live with it - it nevertheless is useful to consider the social and ethical consequences of research that aims at making or presenting computers as intelligent. Clearly, in our research we are developing systems - or agents in systems – that exhibit human behavior. For that reason we would like to give some views on societal and ethical aspects.

In a previous section, when we discussed the role of 'Presence', it became clear that 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 stays in a certain domain of task and/or domain knowledge, 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.

As already mentioned in Bates [2] (in the context of agents performing in virtual worlds), rather than build smart but narrow agents we can build agents with a more broad though perhaps more shallow capabilities. That is, for several applications we may be able to use agents that are not intelligent, as long as they are not clearly stupid or unreal. That is,

"An agent that keeps quiet may appear wise, while an agent that oversteps its abilities will probably destroy the suspension of disbelief. Thus, we propose to take advantage of the "Eliza effect", in which people see subtlety, understanding, and emotion in an agent as long as the agent does not actively destroy the illusion."

Now we would like to introduce a second view, presenting support to the quotations given above. For that purpose we want to refer to Friedman and Kahn [10], who give a clear overview of the relation between social actors in computer systems (or computer systems as social actors) and the moral perspective, that is, aren't we introducing relationships between users and computational systems that delegate responsibility from the human user to the computational system? Can computers or entities which become visible in the interface between user and computer held responsible for actions initiated (directly or indirectly) by a user?

In order to answer the question whether a computer can be a moral agent (and thus be held morally responsible for a decision) they argue that this can only be the case if we attribute (human) intentionality to a computer (or an agent in the interface). They take the position that this can not be the case. If so, either the human is reduced to the status of computers, or the computer is raised to the status of humans. Since, in their opinion, nobody has been able to undermine Searle's 'Chinese room' argument, computational systems cannot have intentionality and cannot be held responsible for their decisions and the consequences. Unfortunately, systems can be designed such that the user's sense of his or her moral agency is diminished. This can be done by placing human users into mechanical roles without understanding of their individual actions, or by masquerading the system as an agent with beliefs, desires and intentions. Their conclusion is, humans should not inappropriately attribute agency to computational systems and design practice should discourage this perception.

In order to discourage such a perception, designers should refrain from anthropomorphizing computational systems. Non-anthropomorphic design increases responsible computing. Hence, do not try to model human-human communication and do not (necessarily) try to model human intelligence when designing interaction systems. Rather than blurring the boundaries between people and computers they should be sharpened.

These viewpoints return in the well known debate on 'direct manipulation' vs. 'interface agents' (see e.g. Shneiderman and Maes [27]). While Shneiderman argues that anthropomorphic representations are misleading (destroying the users' sense of accomplishment), Maes has advocated agents that appear on the screen with animated facial expressions and body language.

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.[5]

12 THEATRE-RELATED PURPOSES OF THE VIRTUAL ENVIRONMENT

In the previous sections we concentrated on:

- the public, that is, men, women, children who want attend a certain performance or who want to know about performances in general, in a certain city or region, and at a certain date or in a certain period; the public has expectations about the information that is provided, it knows, for example, that different newspapers have different opinions about performances, hence, it is necessary to be careful in pushing the visitor to attend a certain performance
- the theatres, that is, the organizations that want to sell tickets, want positive reviews for the performances they hire, want to give correct and relevant information to the public, and have to offer contracts to the managers of artists and theatre companies in such a way that they are not loss-making

However, now that we have a virtual theatre where people can look around and get information on performances, wouldn't it be nice to apply this virtual reality environment to other theater-related purposes? Why not look more closely at possibilities that can be offered to:

- the professionals such as stage directors, choreographers, stage crew, sound and light people, etc.
- the performers, hence, the actors, the musicians, the dancers, the artists, authors and poets who present their work and prefer more or other interaction with each other or/and the audience
- the public in its role of audience attending a performance; not necessarily a passive audience just enjoying a performance, but also a web-audience that can (real-time) influence the running of things during a performance or can even more explicitly take part in a performance by taking the role of an actor

In this paper we will not elaborate the possibility to use our environment for scenographic simulations. There are projects aiming at providing professionals tools and environments to help in pre-producing performances. In these projects users can build a scenography of a performance, they can move through virtual models of stage sets in real time, they can experiment with lights or camera effects, change points of view, etc. An example of such an project is CATS (see Gil et al. [11]). See also Diamond et al. [6]).

Rather we would like to add the possibility to look at our environment as a stage on which we can have on-line performances or pre-recorded performances on request. It has been mentioned before, that the computer screen can be looked at as a stage and it has been argued that the theater metaphor can help in understanding human-computer interaction (Laurel [14]). The metaphor needs to be explored further, especially with regard to interface agents and the 'artistic' agents that have been introduced in CMU's Oz project (cf. Mateas [17]) and the Virtual Theatre project of Stanford University (see http://www-ksl.stanford.edu/projects/cait/publicity.html). This requires further investigations in agent theory, but also in the possibilities to use WWW and the computer (not only its screen, but also data gloves, head-mounted devices, special clothing, etc.) to stage performances with real and virtual actors and (real) audience.

It is not unusual today to have meetings in virtual environments. Lectures have been given (and attended) in chat environments and meetings have been held in visualized meeting places. So, why not have live theatre performances over the web?

In the traditional theatre, performers and audience are physically together. There is a focus of attention of the audience in things happening on stage and performers are aware of the audience's attention. Rather than to have one special physical space where performers and audience gather, performers can be geographically dispersed and so can the audience. Moreover, there is no need to maintain the distinction between audience and performers. The environment should allow an (web) audience that can (real-time) influence the running of things during a performance or can even take part in a performance by taking the role of an actor. This requires special attention for the presence issue, both for actors and audience (see Reeve [22]).

[5] There are successful approaches using virtual reality environments in the treatment of certain phobia. This success in what can be called 'desensitization' should be a warning against careless design of human-computer interfaces. How does the interface influence the user, not only in the interaction with the system, but also in his or her interactions with reality?

Figure 11: Hermia and Lysander

Early online performances include a *Hamlet* parody on IRC (Internet Relay Chat) and *The Odyssey* by Homer. Well known is a VRML production of Shakespeare's *A Midsummer Night's Dream* performed live on April 26, 1998. The various avatars playing the roles were controlled by actors and the performance could be seen from any avatar's point of view of the stage (see Figure 11 and http://www.vrmldream.com/ for more details).

13 FUTURE RESEARCH & CONCLUSIONS

In this paper we reported about on-going research and it is clear that all issues that have been discussed here need further research. 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.

As may have become clear from the previous sections, our approach to designing a virtual environment for an interest community is bottom-up. At this moment the system has two agents with different tasks and with no interactions 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 user. These agents can be passive that wait until they are put into action or they sense changes, take initiative and perform actions, e.g. to inform the user without being asked about new information.

One of our concerns in the near future will be the introduction of a conversational agent (which has some general knowledge about well known artists and some well known performances). It may be the case that this agent will resemble Erin ("the coolest virtual bartender in cyberspace") who serves drinks in Spence's Bar, one of the virtual characters built by Extempo Systems (http://www.extempo.com/).

With this conversational agent we have obtained three kinds of dialogues (information & transaction dialogues, command-like dialogues and conversational dialogues). Another concern is an agent that is able to demonstrate how to play musical instruments. This agent requires a much more detailed visualization, including body, arms, hands and fingers and natural movements. In the Spring of 1999 Ph.D. research in this area will start.

In 1999 some versions of our virtual environment have become available for other research groups to work on. For example, in a joint project with the TNO Human Factors Research Institute user evaluation studies will be done and we hope this will help in future decisions about the direction of our work on the theatre information and transaction service interactions and the environment where they take place. Together with KPN Research we hope to investigate the possible role of MPEG4 for visualization and interactions (see Koenen [13]). A simplified and localized version of the virtual environment has been placed at a Dutch technology activity center (Da Vinci). Here, visitors are allowed to play with the system and their (verbal) interactions with the system are logged.

When developing our system there are similar systems which inspire us and from which we hope to learn. We mention:

Trilogy (cf. Norman & Jennings [19]) is a project from the University of London for the development of a virtual research laboratory with intelligent agents. In the laboratory, students are trained in the area of 'traffic engineering' for telecommunication. Agents are used to present information and to give access to tools. Agents can make suggestions when a user is not familiar with the possibilities of the system. In addition they can take care of the efficient use of the available resources.

Steve (cf. Rickel & Johnson [23,24]) stands for Soar Training Expert for Virtual Environments, a project performed at the University of Southern California, Marina del Rey, CA. Steve is a pedagogical agent with the task to help students learn to perform procedural tasks, such as maintaining equipment. It demonstrates

how to perform actions and it uses locomotion, gaze, facial expressions and deictic gestures in its communication with the student. Speech recognition and speech generation have been added to allow task-oriented dialogs between Steve and student.

MUeSLI (cf. Wyard & Churcher [32]) is a project of British Telecom Labs on MUltimodal Spoken Language Interfaces. The multimodal interface that is designed allows language and touch access to a 3D retail system via a kiosk or over the internet. The retail system shows a 3D virtual living room, a 2D fabric palette and a virtual assistant (a 3D talking head). The user may select fabrics and have them displayed on furniture, curtains and walls in the living room.

Figure 12: Jennifer James

It is also worthwhile to look at already existing attempts to commercialize (aspects of) systems similar to the one we are building. A particular nice 'In-Person Service Character' is **Jennifer James** (Figure 12, see again (http://www.extempo.com/), a spokeswoman at a virtual auto show who greets visitors, engages them in a dialogue and a presentation of available cars. She listens to comments and questions typed by a visitor and talks using speech synthesis technology. Her face and body are 3-D animated, actions and reactions are consistent with role and personality and are coordinated with the events of the dialogue. The information she obtains from visitors is stored for follow-up and market research.

In addition to the projects which concentrate on VR environments and interaction modalities there are projects on virtual interactive studio television (e.g., the European VISTA project with applications on Interactive Drama and Interactive Presenters; in this environment viewers can actively participate and direct the programme being transmitted), on digital cinema (multi-treaded movies, interactive series with audience participation, etc.) and on agent-based TV broadcasting.

REFERENCES

[1] Agah, A. & K. Tanie. Multimedia Human-Computer Interaction for Presence and Exploration in a Telemuseum. *Presence: Teleoperators and Virtual Environments* 8 (1), February 1999, 104-111.

[2] Bates, J. Virtual Reality, Art, and Entertainment. *Presence: Teleoperators and Virtual Environments* 1 (1), Winter 1992, 133-138.

[3] Berk, M. van den. Visuele spraaksynthese. Master's thesis, University of Twente, 1998.

[4] Beutnagel, M., A. Conkie, J. Schroeter, Y. Stylianou & A. Syrdal. The AT&T Next-Gen TTS System. In: *ICSLP–98*, Sydney, Australia, November 1998.

[5] Clermont, Th., M. Pomplun, E. Prestin and H. Rieser. Eye-movement research and the investigation of dialogue structure, Proceedings of TWLT13: *Formal Semantics and Pragmatics of Dialogue (Twendial'98)*, J. Hulstijn and A. Nijholt (eds.), 1998, 61-75.

[6] Diamond, D. & T. Berliner (eds.). The Stage Directors Handbook: Opportunities for Directors and Choreographers. Theatre Communications Group, New York, ISBN 1-55936-150-6, 1999.

[7] Dirksen, A. and Menert, L. Fluent Dutch text-to-speech. Technical manual, Fluency Speech Technology/OTS Utrecht, 1997.

[8] Dutoit, T. High-quality text-to-speech synthesis: An overview. *Electrical and Electronics Engineering* 17 (1997), 25-36.

[9] Friedman, B. (ed.). *Human Values and the Design of Computer Technology*. CSLI Publications, Cambridge University Press, 1997.

[10] Friedman, B. & P.H. Kahn. Human agency and responsible computing: Implications for computer system design. In: Friedman [9], 221- 235.

[11] Gil, F.M. et al. 3D Real-time graphic environments for theatrical and TV scenographic simulations. In: Proceedings *Nimes '98*, LLIA Nos 134-135-136, 1998, 163-167.

[12] 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.

[13] Koenen, R. MPEG-4: Multimedia for our time. http://drogo.cselt.stet.it/mpeg/koenen/mpeg-4.html, 1999. See also: MPEG-4: Overview of the MPEG-4 Standard. ISO/IEC JTC1/SC29/WG11, N2459, October 1998, Atlantic City.

[14] Laurel, B. *Computers as Theatre*. Addison-Wesley 1991; 2nd edition 1993.

[15] 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.

[16] Lombard, M. & T. Ditton. At the heart of it all: The concept of presence. *Journal of Mediated Communication* 3, Nr.2, September 1997.

[17] Mateas, M. An Oz-centric review of interactive drama nad believable agents. CMU-CS-97-156, Carnegie Mellon University, Pittsburgh, June 1997.

[18] Nass, C., B. Reeves & G. Leshner. Technology and roles: A tale of two TVs. *Journal of Communication* 46 (2), 121-128.

[19] Norman, T.J. & N.R. Jennings. Constructing a virtual training laboratory using intelligent agents. Manuscript, University of London, 1999.

[20] Ostermann, J., M. Beutnagel, A. Fischer & Y. Wang. Integration of talking heads and text-to-speech synthesizers for visual TTS. In: *ICSLP-98*, Sydney, Australia, November 1998.

[21] Reany, M. The Theatre of Virtual Reality: Designing Scenery in an Imaginary World. *Theatre Design and Technology*, Vol. XXIX, No.2, 1992, pp. 29-32.

[22] Reeve, C. Presence in Virtual Theatre. *BT Presence Workshop*, BT Labs, 10-11 June 1998.

[23] Rickel, J. & W.L. Johnson. Task-oriented dialogs with animated agents in virtual reality. In: Proceedings of the First Workshop on *Embodied Conversational Characters*, Tahoe City, CA, October 1998.

[24] Rickel, J. & W.L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. To appear in *Applied Artificial Intelligence*.

[25] Roehl, B. Draft Proposal for the VRML Streaming Working Group. http://www.vrml.org/WorkingGroups/vrml-streams/proposal.html, 1998.

[26] Schooten, B. van, O. Donk & J. Zwiers. Modeling interaction in virtual environments using process algebra. In: Proceedings *Interactions in Virtual Worlds (IVW'99).* Twente Workshop on Language Technology 15, University of Twente, May 1999.

[27] Shneiderman, B. & P. Maes. Direct manipulation vs interface agents. Exerpts from debates at IUI 97 and CHI 97. *Interaction*, November-December 1007, 42-61.

[28] Sproull, L., M. Subramani, S. Kiesler, J. Walker & K. Waters. When the interface is a face. In [9], 163-190.

[29] Thie, S. & J. van Wijk. Experimental evaluation of social virtual presence in a decision making task. In: Proceedings *BT Workshop on Presence*, May, 1998.

[30] Vertegaal, R. *Look who's talking to whom: mediating joint attention in multiparty communication and collaboration.* Ph.D. Thesis, University of Twente, Enschede, 1998.

[31] Witmer, B.G. & M.J. Singer. Measuring presence in virtual environment: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7 (3), June 1998, 225-240.

[32] Wyard, P.J. & G. E. Churcher. Spoken Language Interaction with a Virtual World in the MUeSLI Multimodal 3D Retail System. In: Proceedings *Interactions in Virtual Worlds (IVW'99)*. Twente Workshop on Language Technology 15, University of Twente, May 1999.

[33] Zahorik, P. & R.L. Jenison. Presence as being-in-the-world. *Presence: Teleoperators and Virtual Environments* 7 (1), February 1998, 78-89.