Access and Exploration of Virtual Reality Interest Communities

Anton Nijholt[1]
Centre of Telematics and Information Technology (CTIT)
University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands

ABSTRACT

In this paper we present our research on developing a web-based environment, a virtual theatre, in which users can display different behaviors and have goals that emerge during the interaction with this environment. One way to support such users is to give them different interaction modalities and access to multimedia information. 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. The system presents its information through agents that use tables, chat windows, natural language, speech and a talking face. It is discussed how our virtual environment can be considered as an interest community and it is shown what further advantages can be obtained if we continue to explore this metaphor.

Keywords: Interest communities, virtual reality, multi-modality, agent technology, speech recognition and synthesis, talking faces, web technology

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 developing an environment, a virtual theatre, 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 diner, 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 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, eyebrows, 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

[1] With contributions by Joris Hulstijn, Hendri Hondorp and Boris van Schooten of the Parlevink Research Group.

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, 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. using 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.

Providing (Web-based) 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).

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 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 linguistic science 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.

Providing (Web-based) Context: Presence and Visualization

We decided to visualize the environment in which people can inform 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.

Commitment of the user to the system is another issue that helps to obtain co-operative behavior. This commitment can be enhanced by introducing agents that can be recognized by the user. All of these issues help to give the user a sense of 'presence', it adds to the social richness of the user, and it

Figure 1. Karin, the Information Agent

emphasizes the role of the user as a social actor (see Lombard et al [8].

Providing (Web-based) 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. 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 and theirs and others behavior in such environments.

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 at a cancer research institute in Seattle. This world enables people struggling with cancer to obtain information and interact with others facing similar challenges. Patients, families and friends can enter the 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. Part of the project consists of developing tools to create other applications.

3. THE TWENTE VIRTUAL THEATRE ENVIRONMENT

Our virtual theatre has been built according to the design drawings made by the architects of the building. 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 surroundings of the building, hear the carillon of a nearby church, look at neighboring pubs and a movie theatre, etc. They can enter the theatre and walk around, visit the hall, admire the paintings on the walls, enter the main performance hall, 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 blackboard that is daily updated using information from the database. Clearly, visitors may go to an information desk, see previews and start a dialogue with an agent called 'Karin'. The first version of Karin looked like other standard avatars available on World Wide Web. The second version, now available in a prototype of the system, has a more human-like appearance making visitors happy to talk with her and ask about performances and make reservations (cf. Figure 1).

One may argue the necessity of this realistic modeling of the theatre and its 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.

4. AGENTS FOR INFORMATION, TRANSACTION & NAVIGATION

The Navigation Agent

Our WWW-based virtual theatre 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 is prepared to allow 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 means 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 turns out to be 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 recognition engine that can recognize their commands and send this information to the server system. However, good speech recognition systems are very expensive and bad systems result in bad 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.

The Information & Transaction Agent

Karin, the information/transaction agent, allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karin 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 [7].

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 Hulstijn et al. [4] (see also section 5). 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.

Visualization of Agents

It has become clear from several studies 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 has been designed. It also makes a difference to interact with a talking face display or with a text display. Finally, the facial appearance and the expression of the face matters. From all these observations (see Friedman [3], for details) we conclude that introducing a talking face can help to make interactions more natural and shortcomings of the technology more acceptable to users. This is especially true in the case of speech technology.

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 an agent helps to recognize the agent’s tasks.

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

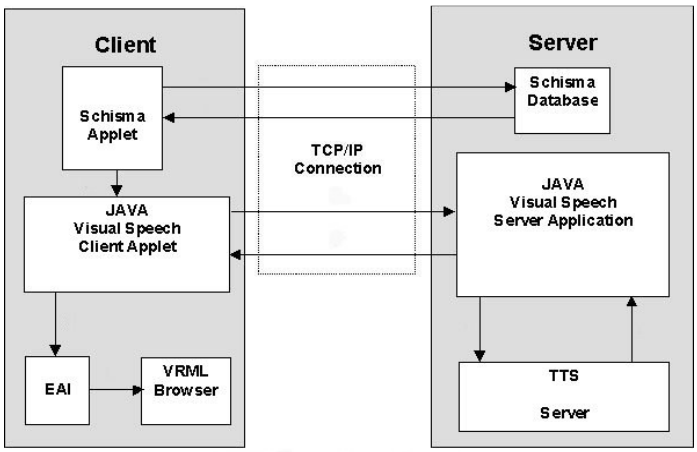


Figure 2. Client-server Architecture for Visual Speech

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 Karin 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.

5. SPEECH GENERATION AND ANIMATION

Speech Generation through Templates

The utterance generation by the information agent uses a list of utterance templates. 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 et al. [2]). The 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. 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.

Facing the Information Agent

We developed a virtual face in a 3D-design environment for the information agent. 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.

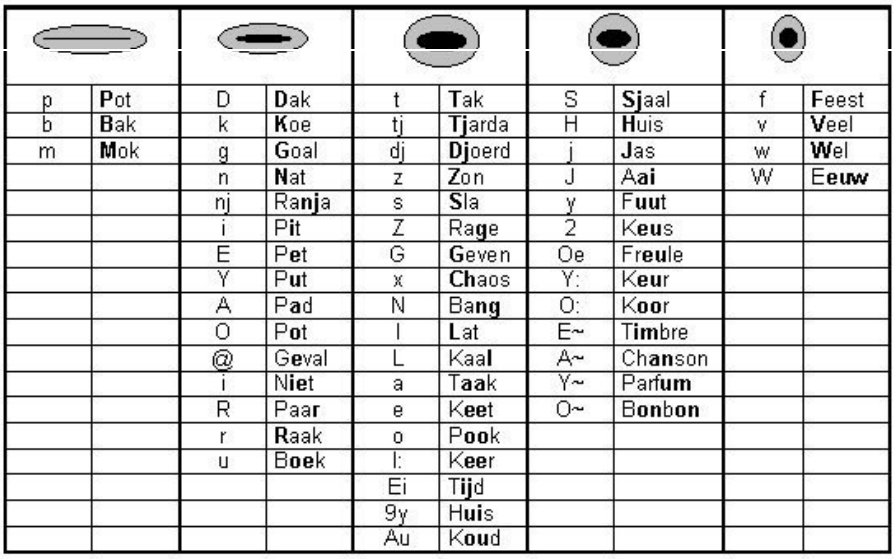
The face is capable of visualizing the speech synchronously to the speech output. This involves lip-movements according to a couple of visemes. The face has to visualize facial expressions according to user's input or the system's output. Figure 2 represents the architecture of the visual speech system. 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. We have included the cartoon face in the Karin avatar. 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. The 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 classes (cf. Figure 3).

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.

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



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	**Eeuw**
		nj	Ran**j**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~	T**im**bre		
		@	G**e**val	L	Ka**al**	A~	Ch**an**son		
		i	N**ie**t	a	T**aa**k	Y~	Par**fum**		
		R	Pa**a**r	e	K**ee**t	O~	B**on**b**on**		
		r	**R**aak	o	P**oo**k				
		u	B**oe**k	l:	K**ee**r				
				Ei	T**ij**d				
				9y	H**ui**s				
				Au	K**ou**d				

Figure 3. Phonemes and Visemes

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?

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'.

Facial Features

Apart from the lips, the virtual face has a number of dynamic control parameters (Berk [1], see 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. 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 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

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 4. Table with Facial Features

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 want to 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.

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. 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. We plan to investigate a black-board architecture where combinations of input parameters trigger rules that produce a response action, or a more permanent change of expression.

6. SPEECH, ANIMATION & WEB-BASED STREAMING AUDIO & 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.

Experiences 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 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. Roehl [9] 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) 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) 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.

7. FORMAL MODELING OF INTERACTIONS IN VIRTUAL ENVIRONMENTS

Both from an ergonomical and a software-engineering view-point, the design of interaction in virtual environments is complex. Virtual environments may feature a variety of inter-active 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 (Concurrent 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 be-tween 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, and explicitly showing which

options are available when and through which modalities. Thirdly, it enables automatic processing, such as architecture visualization and derivation of some system properties.

For more details about this approach we refer to Schooten et al [10]. 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.

8. 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.

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). In this way we have obtained three kinds of dialogues (information & transaction dialogues, command-like dialogues and conversational dialogues). Another concern which we already work on is an agent that is able to demonstrate how to play musical instruments. This requires a more detailed visualization of virtual agents, including body, arms, hands and fingers and natural movements.

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 [5]). A simplified and localized version of the virtual environment has been placed at a Dutch technology information center (Da Vinci). Here, visitors are allowed to play with the system and their (verbal) interactions with the system are logged.

We also would to use the environment for theatre-related purposes. Of course, 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 [6]). However, what we would like to investigate is how the environment can be used to stage performances with real and virtual actors and (real) audience. 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. Examples of such performances have been demonstrated on several occasions.

9. REFERENCES

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

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

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

[4] 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.

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

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

[7] 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.

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

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

[10] 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.