Speech and Language Interactions in a Web Theatre Environment

Anton Nijholt, Joris Hulstijn, Arjan van Hessen

Centre for Telematics and Information Technology (CTIT)
University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

ABSTRACT

We discuss research on interaction in a virtual theatre that can be accessed through Web pages. In the environment we employ several agents. The virtual theatre allows navigation through keyboard and mouse, but there is also a navigation agent which listens to typed input and spoken commands. We also have an information agent which allows a NL dialogue, where input is keyboard-driven and output is by tables and template driven NL generation. In development are talking faces for the agents. A user's commitment to this environment is increased by increasing 'presence'.

1 INTRODUCTION

We present the current state of 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, for example, may 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: they first want to have a dinner, or they want to go to a movie, theatre, or to opera, 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 theatre, and of course, in particular, the performances in this theatre. The interactions take place with different task-oriented agents. These agents allow mouse and keyboard input, but interactions can also take place using speech and NL input. In the 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 synchronized lip movements.

2 HISTORY AND MOTIVATION

Some years ago, our research group started research and development in the processing of NL dialogues between humans and computers. This research led to the development of a (keyboard-driven) NL accessible information system, able to inform users about theatre performances and to allow them to make reservations. This rather primitive system used a database of performances. However, if a user really wants to get information and has little patience, he or she is able to get this information. A more general remark is in place: 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, about users, their preferences and other characteristics, etc., or do we assume that any user will adapt to the system that is being offered? The latter point is important. It has to do with group characteristics, but also with facilities and alternatives provided by the designer.

We do not disagree with a view where users are expected to adapt to a system. On the other hand, it is more attractive (and interesting) to offer environments, where users can 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 the environment? We like to offer a system such that users are stimulated to adapt to it in a natural way.

3 PROVIDING CONTEXT

3.1 Multimodality

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, it is 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 on WWW allow more than one modality for input or output.

Looking at multi-modal human-computer interaction it is clear that hardly any research has been done to distinguish discourse and dialogue phenomena and to model them, for multi-modal tasks. The same holds for approaches to funnel information conveyed via multiple mo-

dalities into and out of a single underlying representation of meaning to be communicated (the cross-media information fusion problem). Similarly, for output, there is the information-to-media allocation problem.

Our second observation, certainly not independent from the observation above, 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 the 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.

3.2 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 using this context. Moreover, it allows the system to influence the interaction behavior of the user such that more efficient and natural dialogues with the system become possible.

Our theatre has been built according to design drawings of a local theatre. Sensor nodes in the virtual environment activate animations (opening doors) or start events (entering a dialogue mode, playing music, moving spotlights, etc.). Information about today's performances is available on a notice board that is automatically updated using information from the database with performances. In addition, visitors may go to the information desk in the theatre to see previews and to start a dialogue with an information & transaction agent called 'Karin'.

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. Users can be made to enjoy the interaction and to perform better all depending on the way the interface and the interaction strategy have been designed. It makes a difference to interact with a talking face instead of a text display. People tend to present themselves in a more positive light to a talking face and they are more attentive when a task is presented by such a face.

From these observations we conclude that introducing a talking face can help to make the interaction more natural and the shortcomings of the technology more acceptable to users. One problem in spoken dialogue systems is the limitation of the context. As long as the context is (very) narrow they perform well. Task-oriented agents can help to restrict user expectations and utterances to the different tasks for which agents are responsible. This can be enhanced if the visualization of the agents helps to recognize the agents tasks.

3.3 Interest Communities

It is interesting to investigate how we can allow communication between users 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. Such communities have been around for some years. They have evolved from text environments to 2D graphical and 3D virtual environments with sounds, animation and video. Visitors enter 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 take responsibility for the environment. It is our aim to extend the current facilities of our environment in such a way that multiple users can meet each other, talk to each other and inform each other, not only by chat windows but also by lectures and presentations in the theatre itself (see[4]).

4 AGENTS IN THE VIRTUAL THEATRE

4.1 An Agent Platform in the Virtual Theatre

In the current prototype version of the virtual theatre we distinguish between different agents: We have an information & transaction agent, 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 keyboard NL. Any agent can start up other agents and receive and carry out orders of others. 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 become an agent. And so can be done with other objects in the environment.

4.2 The Information & Transaction Agent

Karin, the information/transaction agent, allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karin wants to give information and to sell tickets. She is fed from a database that contains the information about performances in the theatre.

The approach used can be summarized as ‘rewrite and understand’. User utterances are simplified using a 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. 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. More information about this approach can be found in [2].

Presently the input to Karin is keyboard-driven natural language and the output is both screen and speech based. Based on the most recent user utterance, on the context and on the database, the system decides on a response, consisting of database manipulation and dialogue acts. Textual output of the system is filtered in parts that are to be shown in a table or a dialogue window and parts that have to be converted to speech output for Karin.

4.3 The Navigation Agent

Navigation is done with keyboard and mouse. This 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 has to resolve deictic references in the interaction. The current version of the navigational agent is not conversational. Straightforward typed commands or similar speech commands make it possible for the user to explore the virtual environment. The phrases to be recognized must contain an action (go to, tell me) and a target (information desk, keyboard). Speech recognition can be improved by using 'word graphs', grammars and context depending word lists, something that will be implemented in our next version. For the speech recognition we currently use the SpeechPearl engine from Philips. Users may use different words to designate parts of the building, including references that have to be resolved during reasoning.

5 SPEECH GENERATION AND ANIMATION

When users approach the information-desk while they are navigating in the virtual theatre they can see an avatar (Karin) standing there and also a dialogue window and a window for presenting information about several performances is shown. We developed also a virtual face for Karin in a 3D-design environment and imported it in VRML. The face is capable of visualizing the speech synchronously to the speech output. For pronouncing we use the Fluent Dutch TTS system which runs on top of the MBROLA diphone synthesizer. It operates at three levels: a grapheme level, a phoneme level and a low-level representation of phones where the length and pitch of sounds is represented. Visualization involves lip-movements according to a couple of visemes and generation of facial expressions according to user's input or the system’s output.

How do we plan to control the responses of the system, the prosody and the artificial face? 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 the latest user utterance or signal (such as a pointing gesture) the dialogue manager selects a response action. A response action is a combination of 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* module selects a way to express it. The module determines the structure, wording, and prosody of each response. It also controls the orientation and expression of the face, the eyes, and the coordination of sounds and lip movement.

In the design of utterance generation a list of annotated templates is used. They 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 possibly empty lists of information items that are marked *given*, *wanted* and *new*. Utterance type and body determine the word-order and the main intonation contour. The presence of information items in the *given*, *wanted* and *new* slots, as well as special features affect the actual wording and intonation of the utterance. Templates respect rules of accenting and de-accenting. Information that is given in the dialogue is de-accented, expressed as a pronoun, or even left out. It is repeated whenever the system is not confident it was recognized correctly by the speech recognition module. Verification prompts are distinguished by a rising intonation. Information that is to be presented as new, is accented. Quoted expressions (artist names, titles of performances) are set apart. For reading texts that describe the content of performances, the system assumes a ‘reading voice’.

Apart from the lips that are controlled by the phoneme sequences, the virtual face has a number of dynamic control parameters. They deal with the gazing of eyes, movement eyelids and eyebrows, and head orientation. 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. Our current working hypothesis is that gestures synchronize with utterances, or precede them. So we link the ges-

ture'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. To control the many features we propose a blackboard architecture where combinations of input parameters trigger rules that produce utterance related facial movements and more permanent changes of expression.

6 THE ROLE OF PRESENCE

'Presence' [3] 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.' This illusion can occur in distinct ways:

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

Many conceptualizations of presence available in the literature contain more detailed viewpoints, such as:

- Is the medium perceived as sociable, warm, personal, etc., in the interaction? Do users overlook the artificial nature of entities within a medium with which they interact?
- Does the user have the feeling that he/she has been transported to a different place, does the user have the feeling to share it with others?
- How much are the user's senses immersed in the virtual world and involved in the interaction?

It may be clear that our topics of interest are closely related to these issues. The environment that is offered looks familiar, the functions of several objects 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. We have looked at possibilities to increase a user's commitment to the system with the aim to obtain co-operative behavior. One obvious reason which makes us loose a user is when clumsy technology (like speech and language technology) is not sufficiently backed up by context (including different modalities) which seduces the user to a certain interaction behavior and which helps to disambiguate the users utterances.

7 FORMAL MODELING OF INTERACTIONS

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 VR techniques 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 these issues by means of a formal modeling technique that is based on the process algebra CSP. 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 (video versus audio, text versus graphics) may also be modeled as separate channels.

This technique has some strong points. A simplified and formal model enables a clear and unambiguous specification of architecture and dynamics. Secondly, it may be useful as a conceptual model, modeling the fact that a user experiences interaction with 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 system properties. In [5] we show 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 REFERENCES

1. Hulstijn, J. & A. van Hessen. Utterance Generation for Transaction Dialogues. 5th *ICSLP*, Vol. 4, Sydney, 1998, 1143-1146.
2. Lie, D., J. Hulstijn, A. Nijholt, R. op den Akker. A Transformational Approach to NL Understanding in Dialogue Systems. *NLP and Industrial Applications*, Moncton, 1998, 163-168.
3. Lombard, M. & T. Ditton. At the heart of it all: The concept of presence. *J. of Mediated Communication* 3, Nr.2, September 1997.
4. Nijholt, A. The Twente Virtual Theatre Environment. *Interactions in Virtual Worlds (IVW'99)*. TWLT 15, University of Twente, May 1999.
5. Schooten, B. van, O. Donk & J. Zwiers. Modeling interaction in virtual environments using process algebra. *Interactions in Virtual Worlds (IVW'99)*. TWLT 15, University of Twente, May 1999.