Inhabited interfaces: attentive conversational agents that help

A Nijholt[1], D Heylen[1] and R Vertegaal[2]

[1]Department of Computer Science, University of Twente,
PO Box 217, 7500 AE Enschede, THE NETHERLANDS

[2]Computing & Information Science, Queen's University, Kingston, Ontario, CANADA

[1]*{anijholt,heylen}@cs.utwente.nl,* [2]*roel@acm.org*

ABSTRACT

We discuss the role of attentive agents in virtual reality interfaces. This discussion is guided by experiences and experiments with a virtual reality environment we designed and implemented. In this environment we have introduced agents, sometimes embodied, with which the users can communicate using different input modalities. These agents provide information or are able to perform certain transactions or they help the user in finding her way in the virtual environment, allowing a mix of user exploration and guidance. Among the input modalities that are considered are speech, natural language, mouse and keyboard and gaze. Output includes natural language, visual speech, changes in the virtual environment, animations and menus. Gradually this environment evolves to an environment where multiple users and agents live and communicate with each other. Apart from offering different input modalities and attentive agents, in the near future we also hope to be able, based on current experiments, to offer suggestions to the users based on preferences obtained from their user profile and their visit history.

1. INTRODUCTION

This paper is a progress report on our research, design and implementation of a virtual reality environment where users/visitors/customers can interact with agents that help them to obtain information, to perform certain transactions and to collaborate with agents in order to get some tasks done. We consider this environment as a laboratory for doing research and experiments on users interacting with agents in multimodal ways, referring to visualized information and making use of knowledge possessed by domain agents, but also by agents that represent other visitors of this environment. As such, we think that our environment can be seen as a laboratory for research on users and user interaction in (electronic) commerce and entertainment environments. Moreover, we expect that despite whatever is said about ubiquitous computation, disappearing computers, etc., especially in the home environment, there will be a growing need for social interfaces that can as well be considered as interest communities, inhabited by domain agents, user agents, friends and relatives, etc., that help, advise, discuss and 'negotiate' on matters that range from how and what to prepare for dinner until how to end a relationship with a boyfriend. Our current experiments include the detection of the user's gaze by agents that inhabit the screen and speech commands for a navigation and guidance agent that helps the user to explore a virtual environment by voice rather than using keyboard and mouse. All these experiments are part of the main goal to construct habitable environments that can be approached using natural interaction techniques.

Whether or not the use of embodied conversational agents is appreciated by users is to a large extent an empirical matter (see Rickenberg and Reeves, 2000, for instance). Much will depend on the quality of the agents and the appropriate use that is made of them. The same applies to the use of such agents in interactions with disabled users. The benefits and drawbacks of high quality natural interactions between virtual agents and disabled users will depend on the kind of disability, the kind of task that the agents are put to and many other factors. If the disabled are hindered in their human-human communication situations than they might encounter the same difficulties in the virtual case. If multiple modes of communication are offered to the user than this may provide alternatives to the common keyboard and mouse input modes or the screen and audio channels that may not be available to the disabled user. The major goal is to introduce these other modes of communication like they are used in face-to-face human interactions as Jacob (1995)

advocates. In our experiments we are interested in theoretical issues that concern the modeling of natural interaction as well as in the practical use that can be made of this.

2. BACKGROUND

In Lie et al. (1998) we discussed a natural language dialogue system that offered information about performances in some (existing) theatres and that allowed visitors to make reservations for these performances. The intelligence of this system showed in the pragmatic handling of user utterances in a dialogue. Although the 'linguistic intelligence' was rather poor, the outcome of a linguistic analysis could be given to pragmatic modules which in the majority of cases (assuming 'reasonable' user behavior) could produce system responses that generated acceptable utterances for the user. The general idea behind this system was that users learn how to phrase their questions so that the system produces informative answers. The system prompts can be designed in such a way that users adapt their behavior to the system, the prosody of system utterances (in a spoken dialogue) can invite user's to provide information that they already assumed to be known by the system and, more generally, the system may allow the user to assume and address information available to the system either because that information has been visualized in the dialogue context or because the user may assume that the system employs agents that can start searching for certain information (on WWW). Both aspects – visualization and agents – have been the main topics of further research.

In Nijholt et al. (1998) we reported about embedding our theatre system in a virtual reality environment that allowed visitors to walk around in the theatre, to approach an information desk with an agent (Karin, see Figure 1) with a talking face that is able to address the user in a natural language dialogue about available performances. The theatre has been built according to construction drawings provided by the architects of the building. Visitors can explore this environment, walk from one location to another, ask questions to available agents and objects, click on objects, etc. Karin, the receptionist of the theatre, has a 3-D face that allows simple facial expressions and lip movements that synchronize with a (Dutch) text-to-speech system that mouths the system's utterances to the user. Presently, in our implementation of the system, there is no sophisticated synchronization between the (contents of the) utterances produced by the dialogue manager and corresponding lip movements and facial expressions of the Karin agent. Design considerations that allow an agent to display believable behavior can be found in Nijholt and Hulstijn (2000).

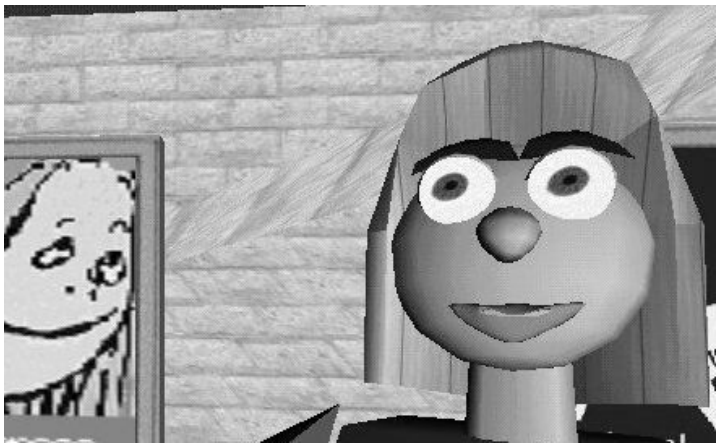

Figure 1. *A close-up of Karin, the virtual receptionist in our virtual environment*

Gradually we have moved from text-based dialogue systems to virtual environments in which the dialogues are embedded in a visual context with embodied agents as conversational partners. Several input and output modes have been added to make the conversations look more like naturally occurring interactions between humans. Below we discuss some of our latest experiments.

3. NATURAL INTERACTIONS WITH AGENTS

The embodied agents in virtual worlds can serve as representations of the user or they can act as autonomous virtual humans. In the first case, there are various means by which users can be represented. Avatars can take the form of simple visual forms or may look like animated characters, possibly resembling the human user

(by photographic means). Also the degree in which the user can manipulate the representation can differ a lot. Simple manipulations might be restricted to moving the representation around in the virtual world. With more expressive means of representation, like manipulable faces, the user might be able to change the facial expressions or these may be copied from the natural expressions by the user. Thus also the degree of correspondence in terms of truthfulness between the representation and the represented is an issue in these matters. The avatar functions as a mediator between the user, the virtual world, and possibly other avatars/users in the same world. This means that there are two stages of interactions: between the user seeing/manipulating the avatar and the virtual world and between the avatar and the other users or the virtual environment.

The embodied virtual agents in the virtual world need not be representations of the user but may be autonomous synthetic characters. One aspect of such characters that we are particularly interested in is their capacity to interact with users and other agents; their conversational skills. Because the interaction between them and the user is modeled on the interaction between humans, their success can be judged on the basis of how well they mimic human-human interaction. Another qualitative aspect has to do with the appropriateness of using such agents for specific interface tasks. For certain applications other forms of interaction may be more efficient.

In our projects we are not only interested in the theoretical and implementation issues that are involved in the creation of naturally conversing embodied agents but also in human factors that determine the appropriateness of using such agents in real applications. In order to investigate such aspects we look for ways to make interactions with virtual agents mimic face-to-face (body-to-body) conversations as well as possible, building experimental set-ups. Natural interactions must be modeled as joint-actions in the sense of Clark (1996) using multiple synchronized input/output channels of different modality (sound, vision).

There are a number of reasons why conversations have to be modeled as joint-actions. For one, the production of an utterance by a speaker is only part of a conversation if it is accompanied by perception on the part of the hearer. Secondly, a conversation is made up of a sequence of turns. Participants alternate between speaker and hearer roles. Modeling a conversation simply as a sequence of turns ignores important aspects of a natural conversation. While a listener is attending to the speech by the speaker, he is also continuously providing (back channel) feedback that the speaker perceives and reacts to during his turn, adjusting his production in response to this feedback. This fact is often not taken into account in models and implementations of conversational systems.

Natural, face-to-face conversation between human agents also involves multiple modes of expression. Speech output is accompanied by different types of gestures, facial expressions, body posture and gaze. Current research in multimodal interaction is concerned with modeling these different ways of expressions and their synchronization using animated agents. In the ideal situation these agents should be able to perceive and decode multimodal input from the human user or other agents on the one hand and to produce this type of output convincingly on the other. The common multimedia desktop computer is fluent in providing images and sounds to the user but has only limited and unnatural devices for input (keyboard and mouse). Currently, the conversational agents that people may be familiar with are animated characters that react to typed input only though speech input is becoming more widely available.

In our current environment we have several agents implemented in a Java-based agent framework allowing (primitive) communication between the agents. Moreover, visitors can address these agents. Presently we are experimenting with the DeepMatrix multi-user browser (Reitmayr, 1999), allowing a visitor to see others and to start chat sessions with other visitors. Although this is a nice addition to our present environment, the chat extension does not fit the agent framework, there are separate channels for communicating with system agents and for communicating with other visitors. In the following sections we report about some steps that are taken in order to solve this problem. In general, these can be characterized as follows.

- Redesigning and extending our agent framework such that individual agents can represent (human) visitors (e.g., movements, posture, nonverbal behavior) and can stand for artificial, embodied domain agents that help visitors in the virtual environment (using natural language).
- Designing VRML agents that are controlled through the protocol of the agent framework, that can walk around in the virtual environment (either acting as a domain agent, hence displaying intelligent and autonomous behavior, or representing a visitor and its moving around in the environment).
- Investigations into (and partly discovering) linguistic and dialogue modeling problems that are specific for multiple dialogue partners present in a virtual environment.

In particular we look at two types of studies. The first, described in section 4, addresses our implementations of navigation agents that assist the visitor of the virtual environment. The second type of experiments deals with modeling of gaze behavior in multi-agent conversations (section 5).

4. NAVIGATION

4.1 Navigation Using Speech and Language

Since it turned out that non-professional users have tremendous problems navigating in virtual environments we introduced a navigation agent in our environment, which can be addressed in limited natural language using the keyboard or spoken utterances. Apart from the well-known shortcomings of state of the art speech technology it turned out to be a useful addition. Because of ownership problems of the commercial software that is used (Speech Pearl, Philips) the navigation agent has not yet been included in the publicly accessible websites that have been made available for our system. It is left to the user to choose between interaction modes (speech and keyboard) or to use both, sequentially or simultaneously. In general, a smooth integration of the pointing devices and speech in a virtual environment requires that the system has to resolve deictic references that occur in the interaction. Moreover, the navigation agent should be able to reason (in a modest way) about the geometry of the world in which it moves. The navigation agent knows about the user's coordinates in the virtual world and it has knowledge of the coordinates of a number of objects and locations. This knowledge is necessary when a visitor refers to an object close to the navigation agent in order to have a starting point for a walk in the theatre and when the visitor specifies an object or location as the goal of a route in the environment. The navigation agent is able to determine its position with respect to nearby objects and locations and can compute a walk from this position to a position with coordinates close to the goal of the walk.

In our case, verbal navigation requires that names have to be associated with different parts of the building, objects and agents. Users may use different words to designate them, including references that have to be resolved in a reasoning process. The current agent is able to understand command-like speech or keyboard input. Otherwise it hardly knows how to communicate with a visitor. The phrases to be recognized must contain an action (go to, tell me) and a target (information desk, synthesizer). It tries to recognize the name of a location in the visitor's utterance. When the recognition is successful, the agent guides the visitor to this location. When the visitor's utterance is about performances the navigation agent makes an attempt to contact Karin, the information and transaction agent. In progress is an implementation of the navigation agent (cf. Van Luinen, 2000) in which the navigation agent knows about (or should be able to compute):

- Current position and focus of gaze of the user;
- What is in the eyesight of the visitor;
- Objects and the properties they have;
- Geometric relations between objects and locations;
- Possible walks towards objects and locations;
- Some knowledge of previously visited locations or routes;
- The action it is performing (or has performed)
- Some knowledge of the previous communication with the visitor.

Presently two other approaches are followed in our research on navigation aids in virtual environments. These approaches, unfortunately, have to be followed in different projects. One is the U-WISH (Usability of Web-based Information Services for Hypermedia) project in which we participate as members of the Dutch Telematics Institute and the other is the Jacob project which we do as members of the VR-Valley Foundation, an initiative which aims at establishing a regional knowledge center on virtual reality in the Netherlands. We hope to be able to combine the results of the three approaches in a future design of navigation agents in our virtual environments.

4.2 Navigation in the U-WISH Project

In the U-WISH project (Neerinck et al., 1999) cognitive engineering techniques are used to develop and test support concepts for networked user interfaces and to derive HCI guidelines based on the test results. One of the test services being used in the U-WISH project is the virtual music center. In the context of this project a new agent-based navigation assistant has been built. Rather than exploring the problems associated with addressing such an agent using speech and language, here the emphasis is on the possibility to obtain an

evaluation framework in which different kinds of user interfaces can be compared. This required some simplifications on our side, but also some useful extensions, e.g. user profiles.

It is clear that in many situations we can expect different user interaction behavior and different user preferences with respect to the 'content' that is offered. These differences follow from different interests, background, culture, intelligence and interaction capabilities of users. These issues can become part of a user profile (obtained by learning, by assuming or by asking), help the system to anticipate the users preferences and even help to guide a user's avatar acting in the virtual environment. For experimental purposes the user profiles in the U-WISH project are fixed. They just contain a few fields containing, among others, name, profession and interests of a user.

In this project, in the user's browser we have an 'eavesdropper' that listens to the interactions of the user with the virtual environment (our virtual music center) and sends them to the server. For each user the server has an administrator agent that creates (or loads) a user profile, an event history and an advice history. Moreover, it creates a number of sub-agents. Events coming from the client are received by the administrator agent, entered into the event history and then send to an appropriate sub-agent. Responses from a sub-agent are logged in the advice history and send to the client's virtual music center. For instance, there is a sub-agent called the PositionAgent, which generates responses based on the position (triggered when the user passes a sensor in the virtual environment), the event history and the profile of a user. Similarly, there is a sub-agent called the DialogAgent, which monitors the dialogue with Karin for certain keywords. The responses by these and other possible sub-agents take the form of suggestions to the user, which, at this moment, are displayed, in an advice window. This window may contain text, hyperlinks and internal links to other parts of the virtual environment. The current agents are rule-based, but as long as they comply with the input/output conventions in their communication with the administrator agent more sophisticated agents can be introduced.

During the U-WISH navigation experiments that are now in preparation tasks have to be performed. They are embedded in scenarios about fictive users. Some of the tasks are open (find some general information within a certain limit of time), others are closed (find a specific piece of information). Half of the test participants will be supported by the navigation assistant, the other half not. Results will be presented in a forthcoming paper.

4.3 Navigation in the Jacob project

In the Jacob project (cf. Evers and Nijholt, 2000) we have the task to design an animated agent, which is called Jacob, in virtual reality, which gives instruction to the user. In this project software engineering plays a prominent role. We apply object-oriented techniques, design patterns and software architecture knowledge. In the architecture we have separated the concerns of the 3D visualization from the basic functionality, which follows from a task model, an instruction model and a user model. Presently, the task and instruction model form Jacob's mind, a control system that observes the world and tries to reach specific objectives by having Jacob perform a certain task (e.g., show the user what to do next) or to produce an utterance to direct the user. Presently Jacob's task is to teach a user to solve the Towers of Hanoi problem. This is chosen as an example task since we think that the design solutions found there can be generalized very well and when Jacob will be integrated in the virtual music center it can help to navigate through the environment (to teach the user what to do and to find where).

5. IMPLEMENTING GAZE BEHAVIOR AND GAZE DETECTION

Our research into embodied conversational agents is concerned with improving the naturalness and fluency of conversations, addressing a number of issues mentioned above such as the continuity of receiving and producing information in joint interaction and the coordination of different modalities. By defining and implementing different set-ups, using the virtual theatre environment as a basis, we want to achieve insight into the proper modeling of conversations and also measure the effects in terms of user satisfaction.

Several of the projects we are currently engaged in concern the use of a gaze detector in conversations with multiple agents, focusing on the interaction between gaze and turn taking. Seeking or avoiding looking at the face of conversational partners serves a number of functions (Kendon, 1967), one of which involves the regulation of the flow of conversation. Certain patterns in gaze behavior of speaker and hearers that are correlated with turn-taking patterns. For instance, a person tends to look away when beginning to speak and returns to look at the hearer at about the end of utterances or turns. In Vertegaal (1998) such patterns were examined in the context of conversations between a number of human dialogue participants and the

implications for representing conversational participants in groupware systems were worked out and implemented in an experimental setting.

In addition to the agent-oriented, the computational linguistic and the dialogue management approaches mentioned above, we are currently working to implement our findings on gaze behavior (Vertegaal et al, 2000) in our environment. That is, the system establishes where the user looks by means of a desk-mounted LC Technologies eye tracking system (http://www.eyegaze.com). In our system multiple conversational agents can be embodied by means of cartoon faces or by using 3D texture-mapped models of humanoid faces. Based on work by Waters and Frisbee (1995), muscle models are used for generating accurate 3D facial expressions. Each agent is capable of detecting whether the user is looking at it, and combines this information with speech data to determine when to speak or listen to the user.

Certain aspects of the structure of the linguistic signal, for instance the topic-focus organization, play another part in gaze and turn-taking behavior during conversations (Torres et al, 1997). At the moment we are investigating how to build agents that know how to behave according to these patterns of natural conversation. To help the user regulate conversations, agents should generate display appropriate gaze behavior. Figure 2 exemplifies this. Here, the agent speaking on the left is the focal point of the user's eye fixations. The right agent observes that the user is looking at the speaker, and signals it does not wish to interrupt by looking at the left agent, rather than the user.

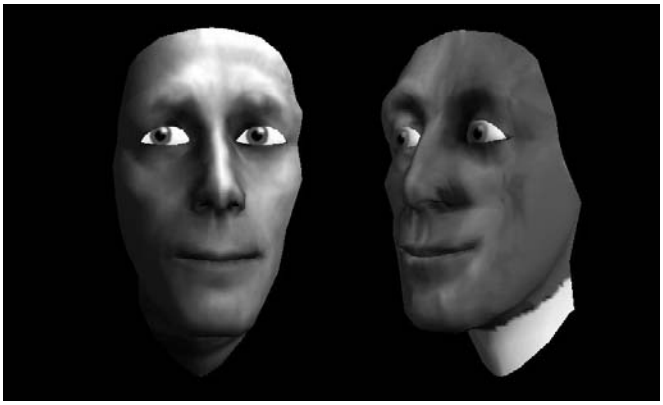

Figure 2. *Gaze behavior experiments with two synthetic agents and a user.*

Our experimental set-up thus consists of two animated talking faces, possibly displayed on two separate screens. The agents can turn their heads and eyes in a number of relevant directions: looking at the user, each other and several other positions. The user's eye-movements are being tracked and the agents are informed about this when their field of vision includes the eyes of the user. Simple conversations will be conducted in which various parameter settings are tested that concern the gaze behavior of the agents, the way in which they and the user take turns in correlation with the informational organization of the utterances they and the user produce. In one of the experiments that will be done in the next months we will have a set-up where we the two agents have related tasks.

A number of variant of this experiment will be carried out. For one version we expect to make an explicit distinction between the information task and the reservation task of our information and transaction agent Karin. Hence, we have a Karin-1 and a Karin-2 who have to communicate with each other (information about user and chosen performance) and with the visitor. Clearly, when during the reservation phase with Karin-2 it turns out that the desired number of tickets is not available or that they are too expensive, it may be necessary to go back to Karin-1 in order to determine an other performance. Although the separation of tasks may look a little artificial, it gives us the opportunity to experiment in the existing environment and with a (modified) existing dialogue system. In other versions the dialogue will be restricted to some canned phrases with variations in timing, turn-taking and gaze behavior of the agents and with variation in the active participation of the individual faces (introducing speakers and silent bystanders). This will provide us with more data on the gaze behavior of users in these virtual settings. Such an experiment will also be used to find out how different emotional factors or personality traits of these synthetic characters can be defined by tweaking the parameters that determine their gaze and turn-taking behavior.

6. CONCLUSIONS

Although originally we didn't intend to build an environment to assist people with disability, we now slowly approach a virtual environment that can be compared with a social setting where different people are ready to help in a conversational way. Obviously, a lot of work has to be done, but recognizing this line of research has been a stimulating attainment. Working towards total communication is not just of theoretical interest may useful to enhance human machine interactions and to bypass the restrictions of the common input-output modes of the stereotypical desktop computer. This is even more true when we move beyond the personal computer. In intelligent environments there is not necessarily a central screen and keyboard. Instead we may expect to have attentive environments where joint voice and gaze information will (unambiguously) activate one or more devices and agents (out of many) in the environment (see Matlock et al., 2000). And from the opposite point of view, agents and devices that try to get our attention by using speech and gaze when necessary.

Acknowledgements: Support for this research was obtained from VR-Valley Twente and the Telematics Institute. Casper Eijckelhof, with help from Job Zwiers and Betsy van Dijk, implemented the navigation agent discussed in section 4; Marc Evers realized the Jacob agent discussed in the same section. Martijn Polak, with guidance from Roel Vertegaal, realized the prototype gaze experiment environment mentioned in section 5.

7. REFERENCES

H. Clark (1996). *Using Language*, Cambridge University Press, Cambridge.

M. Evers and A. Nijholt (2000). Jacob – An animated instruction agent in virtual reality. In: Proceedings 3rd *International Conference on Multimodal Interfaces (ICMI 2000)*, Beijing, Lecture Notes in Computer Science, Springer, to appear.

R. Jacob (1995). Eye Tracking in Advanced Interface Design. In *Virtual Environments and Advanced Interface Design*, ed. by W. Barfield and T Furness, Oxford University Press, New York, pp. 258-288.

A. Kendon (1967). Some functions of Gaze Direction in Social Interaction. *Acta Psychologica* 32, pp. 1-25.

D. Lie, J. Hulstijn, R. op den Akker and A. Nijholt (1998). A Transformational Approach to NL Understanding in Dialogue Systems. Proceedings *NLP and Industrial Applications*, Moncton, New Brunswick, pp. 163-168.

J. van Luinen (2000). M.Sc. thesis, in preparation.

T. Matlock, C.S. Campbell, P.P. Maglio, S. Zhai and B.A. Smith (2000). On gaze and speech in attentive environments. In: Proceedings 3rd *International Conference on Multimodal Interfaces (ICMI 2000)*, Beijing, Lecture Notes in Computer Science, Springer, to appear.

M.A. Neerincx, S. Pemberton and J. Lindenberg (1999). U-WISH; Web usability: methods, guidelines and support interfaces. TNO-report TM-99-D005, TNO Human Factors Research Institute.

A. Nijholt, A. van Hessen and J. Hulstijn (1998). Speech and language interaction in a (virtual) cultural theatre. Proceedings *NLP and Industrial Applications*, Moncton, New Brunswick, pp. 176-182.

A. Nijholt and J. Hulstijn (2000). Multimodal Interactions with Agents in Virtual Worlds (to appear). Chapter in *Future Directions for Intelligent Information Systems and Information Science*, N. Kasabov (ed.), Physica-Verlag: Studies in Fuzziness and Soft Computing.

G. Reitmayr, S. Carroll, A. Reitemeyer and M.G. Wagner (1999). Deep Matrix: An open technology based virtual environment system. *The Visual Computer Journal* 15, pp. 395-412.

R. Rickenberg and B. Reeves (2000). The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. Proceedings *CHI 2000*, April 2000, pp. 49-56.

O. Torres, Justine Cassell and Scott Prevost (1997). Modeling Gaze Behavior as a Function of Discourse Structure. *First International Workshop on Human Computer Conversation* Bellagio.

R. Vertegaal (1998) *Look Who's Talking to Whom*. PhD Thesis, Twente University, Enschede.

R. Vertegaal, R. Slagter, G. van der Veer and A. Nijholt (2000). Why conversational agents should catch the eye. Proceedings *CHI 2000*, pp. 257-258.

K. Waters and J. Frisbee (1995). A coordinated muscle model for speech animation. In Proceedings of *Graphics Interface'95*. Canada.