ANTON NIJHOLT*
Parlevink Research Group
University of Twente, Enschede, The Netherlands
E-mail: anijholt@cs.utwente.nl

Speech, Language and Gaze in Multi-modal Navigation in Virtual Environments

Abstract

In this paper we survey our research on navigation in virtual environments. The research is done in a laboratory environment we designed several years ago. The environment is a virtual theatre that has been made available to the general audience on WWW. A navigation agent has been introduced that knows about the environment and that can be addressed using speech. However, the interaction between navigation agent and user is rather primitive. We survey our research to give more intelligence to the navigation agent and to employ it in a framework of interacting, intelligent agents. In this framework we look at speech and language as interaction modalities in a context of verbal and nonverbal interactions. Gaze and its role in regulating a conversation is one of the modalities we try to model in combination with speech and language interaction.

1. Introduction

In this paper we introduce our research on navigation and advisor agents in 3D virtual reality (VR) environments. This research does not aim at replacing human decision-making by agent decision-making. Rather it emphasizes the cooperation between an agent and a human in search of particular information. This cooperation should take place in a natural way.

What does 'natural' mean? First of all, interaction with a navigation agent requires multi-modality. Especially in a VR environment, the visualization of information may lead visitors to make references to information in many different ways. Information may be looked at and when we talk about it, ask questions about it, we make references to what we see. Our language use and our interaction behavior will be influenced by this information. When we have an embodied agent our interaction behavior is also influenced by the assumed intelligence and the human-like similarity. It should know where we point at with the mouse or even where we look at when asking a question.

Especially in VR these are interesting and sometimes confusing issues. The user has been navigating, using a mouse or a joystick, in order to get to a certain position with a certain field of vision in the virtual environment. The environment is presented from that viewpoint on the screen and the system knows what is in this field of vision (as if the user is an avatar at that position). However, there is also the user, not being in the virtual environment but looking at the screen and looking at what the (invisible avatar) representing himself is looking at. In a non-immersive VR environment the user can focus on certain positions on the screen without changing his avatar's sight, which leads to interesting mixtures of first and second person perspectives.

Gaze and how to exploit it in a situation like this, but also in a situation where we talk to one or more embodied agents on a screen, is one of the topics we look at in our research on navigation or advisor agents. A navigation agent should be able to understand us, whether we interact through speech, language and gaze, or, eventually, gestures, body movements, and facial gestures. Similarly, on the output side, a navigation agent should be able to generate and present multi-media information, explain and demonstrate, guide the visitor through the environment and, when it has been embodied, do it using similar nonverbal signals humans make in face-to-face communication.

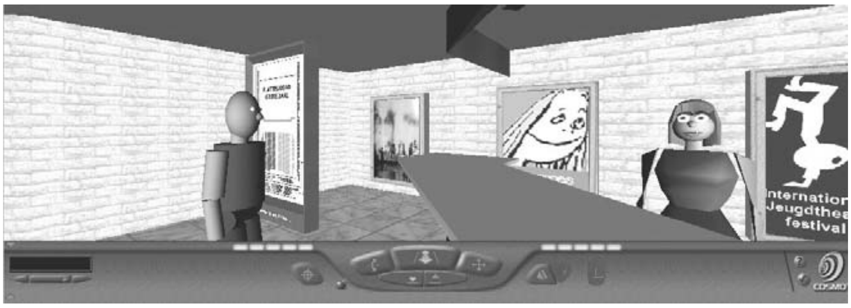


Figure 1: A user's avatar visits domain agent Karin

In this paper we survey our approaches to the design of an intelligent navigation agent. These approaches are followed in separate subprojects because of being funded by different companies or foundations, having different objectives with the projects. Nevertheless we hope to be

*Special Conference in this **7th International Symposium on Social Communication**.

able to integrate the different research lines in the near future. The paper is organized as follows. In section 2 we give a short explanation of our virtual environment. A more comprehensive overview can be found in [Nijholt & Hulstijn, 2000]. Section 3 introduces the navigation task and how we dealt with it in a first version of the navigation agent that was built using commercial speech recognition software. In section 4 we introduce the approach we took in one of our subprojects where we built a navigation agent that has much better knowledge of language and dialogue, but where the natural language utterances of the user can only be given using the keyboard. Section 5 is devoted to a subproject where we make a first attempt to have personalized navigational aid. That is, we experiment with an agent that knows about the users preferences. Here the emphasis is more on an agent-based design than on verbal intelligence of the navigation agent. In section 6 of this paper is on our experiments on gaze and the way we expect to integrate our research about gaze modelling with the other capabilities of the navigation agent.

2. The Virtual Environment

The Twente virtual theatre environment is a VRML built 3D environment that resembles an existing local theatre building in our hometown. The environment has been made accessible on WWW. Visitors can enter the building, walk around, visit the performance halls and start a conversation with an embodied agent that knows about performances and can make reservations for them. This agent has access to a database containing information about all the performances that take place in the theatre. It can be accessed using natural language. The theatre has many additional features. An animated piano player on stage takes care of some background music. A baroque dancer, imported from another research group, can be animated to do dances. The basic version of this environment is available on the Web. Moreover, it has been installed in a permanent science and technology exhibition where especially children use it to explore the environment and ask questions to the information agent. Several versions have emerged from this basic environment. In one of them we introduced an agent framework, which gives the freedom to introduce agents that have different tasks and that can communicate with each other and with the visitor. One of the agents that has been introduced in this framework is a speech-accessible navigation agent. We will return to this agent in the next section. Yet another version is a multi-user theatre environment. Using DeepMatrix (Reitmayr, 1999) it has become possible that different users enter the environment, can see each other and can chat with each other. Unfortunately, current web technology does not allow us to integrate these approaches in one system that can be offered to the general audience. Moreover, in order to have an environment where we have both multiple users and multiple agents and we want to allow them to communicate in any arbitrary configuration there are a lot of problems that require more research in verbal and nonverbal interaction and more standards to share environments and virtual reality objects that have been developed by different creators (Nijholt & Hondorp, 2000; Nijholt, 2000). In Figure 1 we illustrate a situation we now can have in our virtual environment when users use it and domain agents represented by avatars.

3. Speech for Navigation in Virtual Reality

Since it turns out that non-professional users have tremendous problems navigating in virtual environments we introduced a navigation agent in our environment. It 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. It is not that difficult to introduce speech recognition software in our environment. Commercial software is available that can be embedded in a windows environment and which allows a user to give speech commands to the web environment and its plug-ins.

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. This has only been realized in a primitive way. 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 through the agent framework to contact Karin, the environment's information and transaction agent.

4. Language and Navigation in Virtual Reality

In (continuous) progress is an implementation in which the navigation agent knows about (or should be able to compute):

- Current position of the user, what is in the eyesight of the user and where does he or she focus at;
- Objects and the properties they have; geometric relations between objects and locations;
- Possible walks towards objects and locations; some knowledge of previously visited locations;
- The action it is performing (or has performed); some knowledge of the previous communication.

And, of course, apart from knowing about objects, geography, history, etc., we expect that it can use this knowledge in order to give relevant advice and suggestions to users and we expect that it can do this in a natural language dialogue with the user.

Only part of this has been realized yet. In the present situation (cf. van Luin, 2000) we have a system (Demosthenes), which accepts the user's utterance and presents it to its parser based on probabilistic unification grammar. This grammar has been generated automatically from an annotated corpus of user utterances (see ter Doest, 1999) collected in Wizard of Oz experiments. In this approach we tag a corpus with syntactic categories and superficial structure using Standard Generalised Markup Language (SGML). From this tagged data grammar rules, unification constraints and probabilities are derived. We have tested grammars on 'seen' and 'unseen' data from different domains using a probabilistic left-corner parser for PATR II unification grammars. In a similar way we have induced for our navigation agent a probabilistic grammar from a corpus of user utterances that have been obtained from several scenarios presented to (potential) visitors from the theatre. This grammar is a start. It allows the design of a primitive system and it allows bootstrapping this system from the original corpus and from corpora obtained from logging the interactions between visitors and the navigation agent. Clearly, this approach still requires the integration of speech recognition technology with natural language specification and understanding. For that reason it may be useful to investigate the generation of finite state probabilistic (unification) grammars from corpora of utterances.

The current parser uses the grammar and a lexicon to find possible parses of the user's utterance. Parses are assigned probabilities that are obtained from the corpus. If no parse can be obtained, the parser presents a subparse for the longest part of the user's utterance. The most possible parse is used to construct a Sentence Object (a feature structure) and an Action Object. In the latter it should be clear which actions have to be performed by the navigation agent. Therefore it is necessary to solve multimedia references ('this', 'that', there', etc.) by looking at mouse positions and recent objects that have been introduced before in the history of the dialogue and that could have been used in the utterance at that particular position. Obviously, the user is inquired about the decisions that have been made concerning reference resolutions. When the Action Object requires more information than has been given or obtained from reference resolution, the user is prompted to give additional information.

When the navigation agent recognizes an object, a concept or a reference in an utterance, it performs a look-up in its own representation of the virtual theatre. In this particular approach we have lists of objects and concepts, a map of the world (objects, routes, distances) and a so-called focus list (objects and concepts which have been referred to in previous parts of the dialogue). Basic properties of objects and concepts (size, color, position, price) have been stored, including aliases that can be used by the user. Some general relations between objects have been stored too; indeed, a table is a piece of furniture and if we say 'not the big one' we refer to the smaller one in the scene.

When the Action Object contains the information the navigation agent needs, the corresponding action is performed. The action can consist of the computation of a route in the virtual music centre which then will be followed, or the retrieval of some information that can be presented to the user.

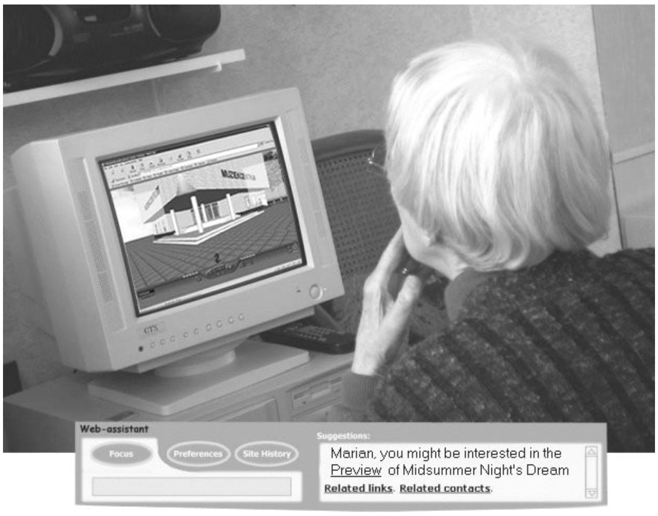
5. Advisor Agents and User Preferences

Presently a more general approach to agents that assist the user in web-based VR environments is followed in the U-WISH (Usability of Web-based Information Services for Hypermedia) project in which we participate as members of the Dutch Telematics Institute. 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 is the virtual music center. In the context of this project a new agent-based advisor has been built. Rather than exploring the problems associated with addressing such an agent using speech and language, here the emphasis is on agent architecture design issues, the cooperation between agents and the possibility to obtain an evaluation framework in which different kinds of user interfaces employing (embodied) agents can be compared. Especially this latter issue required many simplifications on our side, but also some useful extensions, e.g. the introduction of user profiles.

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), that helps the system to anticipate behavior and to select or suggest information or actions that help to satisfy these preferences and, depending on the kind of virtual environment and its application, even help to guide a user’s avatar to act in this 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.

Obviously, this is an extremely simplified view on what can be done with user models. Generally, a user model should contain the knowledge an advisor agent has about what the user knows, wants and has done so far (the history of the user in this particular environment). Apart from what the user has been willing or been able to make known about himself during the interactions the system is able to observe the user’s behavior, both in form: he spent most of his time in that part of the environment, he spent a lot of time looking at a particular restoration of a painting, rather than questions he is using commands, etc., as in content: he seems to be interested in sculptures, he is interested in the history of a particular painting (or is he interested in restorations in general?), he is very straightforward, etc. Knowledge obtained from such observations has to be integrated in an existing user model.

However, what kind of user profile is used is not that important as long as the advisor system has been designed in such a way that more complicated user profiles can be ‘plugged in’. In our design, 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. The window may contain text, hyperlinks and internal links to other parts of the environment. The current agents are rule-based, but as long as they comply with the input/output conventions in the communication with the administrator agent more sophisticated agents can be introduced.



Marian, you might be interested in the Preview of Midsummer Night's Dream

Figure 2: User is getting advice from the Advisor Agent

During the U-WISH navigation experiments that now take place 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 advisor assistant, the other half not. Results of experiments will become available soon.

After these experiments we will further explore the approach we take for advisor agents that can make use of user profiles and the history of events (places that have been visited, advice that has been given). Zwiers and Van Dijk (2000) is a first attempt to give the design of an agent-based structure. One possible subagent that can be introduced in this architecture is the speech or language accessible navigation agent that can be asked about the theatre and that in addition to give advice, can also guide the user to a particular position in the theatre.

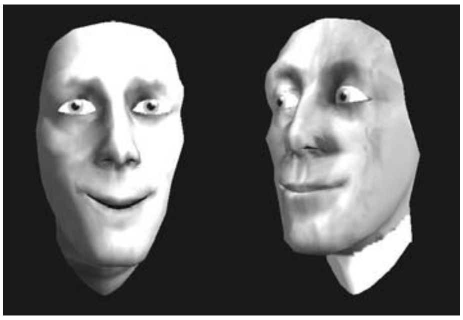


Figure 3: Conversation with two agents

6. Gaze, Attentive Agents and Turn taking

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 modeling of conversations and measure 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. The main project here is done in cooperation with Queens University in Kingston, Canada.

Seeking or avoiding looking at the face of conversational partners serves a number of functions, one of which involves the regulation of the flow of conversation. Certain patterns in gaze behavior of speaker and hearers 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.

Current experiments are meant to see how we can implement research results on gaze and communication with embodied agents and to find the problems when others than ourselves use these prototypes. One of these prototype systems is described in Vertegaal et al. (2000). The system establishes where the user looks by means of a desk-mounted LC Technologies eye tracking system (http://www.eyegaze.com). In this 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. This makes it possible to model turn-taking behavior during conversations and to help the user regulate conversations with multiple agents. Figure 3 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.

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 Figure 4 we have a situation where the dialogue is 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

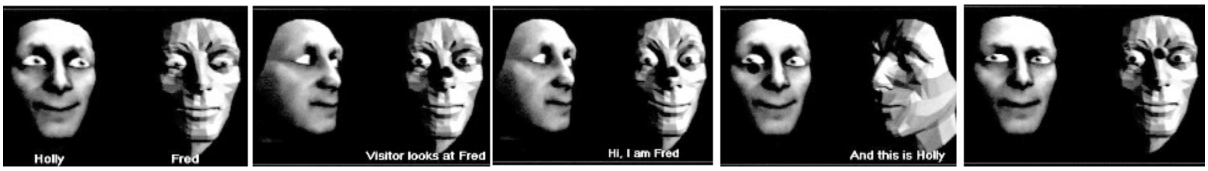


Figure 4: Agents interacting with each other while talking to a human user

speakers and silent bystanders). The black circle which is visible in the second figure (on Fred's nose) and in the following figures denotes where the user is looking. Such an experiment is also 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.

7. Conclusions

We surveyed our research on navigation and advising in our virtual theatre environment. Speech recognition, natural language and dialogue processing, agent communication and multi-modality (including nonverbal interactions) were among the issues that we have to deal with. We hope to be able to combine the results of the different approaches presented in this paper in a future design of an advisor agent in our virtual environments. To do this it is useful to make more clear distinctions between different ways of navigation, way finding and spatial orientation (cf. Volbracht & Domik, 2000).

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 may 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 5; Martijn Polak, with guidance from Roel Vertegaal, realized the prototype gaze experiment environment mentioned in section 6. Other help in writing this paper was obtained from Dirk Heylen and Job Zwiers.

References

Doest, H. ter. (1999). Towards Probabilistic Unification-based Parsing. Ph.D. Thesis, February 1999.

Höök, K., D. Benyon & A. Munro. Proc. Workshop on Personalized and Social Navigation in Information Space. http://www.sics.se/humle/projects/persona/webold/workshop/

Jacob, R. (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, 258-288.

Luin, J, van (2000). Navigatie in het Virtuele Muziekcentrum. M.Sc. Thesis, August 2000.

Matlock, T., 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.

Neerincx, M.A., 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.

Nijholt, A. 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, 148-173.

Nijholt, A. and H. Hondorp (2000). Towards communicating agents and avatars in virtual worlds. In: Proceedings *EUROGRAPHICS 2000*, A. de Sousa & J.C. Torres (eds.), Interlaken, August 2000, 91-95.

Nijholt, A. (2000). Towards virtual communities on the Web: Actors and audience. In: Proceedings ICSC Symposium on Interactive and Collaborative Computing (ICC'2000). University of Wollongong, Australia, December 2000, to appear.

Reitmayr, G. et al. Deep Matrix: An open technology based virtual environment system. *The Visual Computer Journal* 15: 395-412, 1999.

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

Volbracht, S. and G. Domik (2000). Developing effective navigation techniques in virtual 3D environments. In: *Virtual Environments 2000*, J.D. Mulder & R. van Liere (eds.), Springer Computer Science, Wien/New York, 2000, 55-64.

Zwiers, J. and B. van Dijk (2000). Specification and design of advisor agents for multi-modal virtual environments. Draft Deliverable Work-Package 2.3, Centre of Telematics and Information Technology (CTIT), University of Twente, Enschede, the Netherlands.