Towards the Automatic Generation of Virtual Presenter Agents

Anton Nijholt
HMI Research Group, University of Twente,
Enschede, the Netherlands

anijholt@cs.utwente.nl

Abstract

There are many ways to present information to visitors and users of 2D and 3D interface environments. In these virtual environments we can provide visitors with simulations of real environments, including simulations of presenters in such environments (a lecturer, a sales agent, a receptionist, a museum guide) and including audience participation in these environments. Our research aims at generating presentations from available multimedia information. In particular, we would like to see the generation of presentations by embodied conversational agents that employ verbal and nonverbal capabilities. In the past we have seen the introduction of embodied agents and robots that take the role of a museum guide, a news presenter, a teacher, a receptionist, or someone who is trying to sell insurance, houses or tickets. In all these cases the embodied agent needs to explain and to describe. The automatic generation of presentations and presentation agents from information sources is still too ambitious a task. Therefore we look at research from the perspective of the design of tools that can support presenters or can help to provide natural access to presentations and lectures. Can we use a given collection of sheets and maybe other accessible media sources to design, create and generate an embodied presenter? Among others we discuss manual annotation of available information and the way in which presenter agents can use it. Clearly, the development of tools for these purposes is a first step towards automating the generation of presentations and presentation agents.

Keywords: information presentation, virtual presenters, virtual agents, embodied agents

Introduction

In this paper we are concerned with the research and development of embodied presenter agents. These agents have 2D or 3D appearances and they perform on 2D web pages or in 3D (web) environments. They can appear as a talking head showing facial expressions, text-to-speech synthesis and lip synchronization or as more fully embodied agents with postures, gestures and the ability to walk or fly around, demonstrate products or guide visitors to certain locations. Among the latter are agents to explain paintings in a virtual museum, agents to explain routes on a roadmap, agents to present the weather or agents to give a PowerPoint presentation.

This paper is a survey of research on 3D embodied conversational agents (ECAs) explaining a 3D visualized environment (e.g. routes between locations) or explaining 2D visualized information (a

Editor: Eli Cohen

painting, a PowerPoint sheet) presented in a 3D visualized environment. The emphasis will be on the latter.

The research question we want to survey is how can we put knowledge about presentations into use in these embodied virtual presenters? That is, how can we model this knowledge in software and hardware and then give computer support in situations where presentations need to be delivered? Ultimately, our aim is to make virtual presenters available on websites, in 3D and virtual reality environments, that are human-like (i.e., embodied conversational agents) and in which this presentation knowledge is modeled. Obviously, knowledge about verbal and nonverbal communication and how to embed this knowledge into models of the behavior of embodied agents is part of such presentation knowledge.

Currently, animated agents are used as presenters, guides, sale persons, conversation partners, actors, tutors, and to represent human agents in online environments (communities, workspaces, class rooms, etc.) (Cassell, Nakano, Bickmore, Sidner, & Rich, 2000). They are employed on commercial web pages, in educational, training, and simulation environments, and in entertainment applications. In 3D virtual reality environments they move around and they have 3D gestures and pointing actions in order to guide and explain. For this they need knowledge about their domain, their environment (including the user and other agents they interact with), about how to use gestures in relation to their verbal expressions (De Carolis, De Rosis, Carofiglio, Pelachaud, & Poggi, 2001; Theune, Heylen & Nijholt, 2005) and about how to use pointing gestures to indicate objects of interest (Sluis, 2005). In order to make them believable it is useful to equip them not only with intelligence, but ultimately also with models of personality and emotion (Prendinger & Ishizuka, 2004).

In this paper the emphasis is on embodied 3D agents that explain 2D displayed information: a painting, a roadmap, a weather map or a collection of PowerPoint sheets. In our own research we are particularly interested in the latter, that is, agents that are able to give a PowerPoint presentation. In the next section we will look at some previously introduced presenter agents and we look at the research topics that have been tackled in order to have them behave in useful and believable ways. In the third section of this paper the starting point is not so much the presenter agent as the context in which it has to perform. For this we also need to look at human presenters and their audience. This provides a much more general view, taking into account developments and research in the areas of smart meeting and lecture rooms, remote lecture and meeting assistance and virtual and mixed reality environments. It also requires us to look at multimodal and multi-party interaction research. After that we have a short discussion on our own research on virtual presenters performing in multimodal and multi-party interaction environments. Current research activity addresses, in particular, the animation of gestures by a virtual presenter. More recently we also started investigations into adapting a presentation (e.g., the order of slides) to a human presenter by listening to his presentation. It is interesting is to investigate how more general research on generating media presentations (Bocconi, Nack, & Hardman, 2005) and on how to select between modalities for media presentations as done in our own research environment (Bachvarova & Elouazizi, 2005), relates to our research on making virtual presenters intelligent. We end this paper with a section containing the conclusions and some directions for future research.

The research reported in this paper is part of the FP6 European IP on Augmented Multi-party Interaction (http://www.amiproject.org/). This project is concerned with modeling interactions in a smart meeting environment, with the aim to provide real-time support to the meeting partners and to allow off-line multimedia retrieval and multimedia browsing of information obtained from a particular meeting. The technology that is being developed allows us to detect, track and identify people in a particular environment and to interpret their activities and their interaction with other people or with objects or locations in the environment. An FP6 European project that is related to the AMI project is the CHIL (Computers in the Human Interaction Loop) project. This is also a

project on modeling multimodal interactions, but with more explicit interest in the study of presenter activities, including interactions with the audience in a lecture room. Both projects, AMI and CHIL, cover the current EU attempts to stimulate research on multimodal interactions.

Background and Examples

We will say a few words on embodied conversational agents that have been introduced to present information to their human partners. Obviously, there is no clear-cut line between agents that inform users, agents that demonstrate a product, tutor agents or agents that play a role in a simulation.

Sales Persons, Receptionists and Further

The majority of embodied conversational agents and clearly those that can be found in commercial applications deal with tasks that require little intelligence and where the application requires very limited interaction (or no interaction at all) with the user. In these applications we often see a 2D talking face only: there are some facial expressions available and some communication takes place with speech synthesis or with a human voice. The user can make menu choices in order to interact. Obviously, in research environments we see more sophisticated examples of agents, using speech and language modeling, domain knowledge modeling and modeling of emotions and personality (Gebhard, 2001). This is particularly true for agents that act as tutors that have to interact with students in much more subtle ways than a sales person relating the characteristics of a car or a receptionist explaining how to get to a particular office. Non-interruptable news presenters can also be replaced by virtual presenters. Again, there is still a lot to do in order to render newsreaders that make you believe that they understand what they are saying and that express this understanding in a human-like way. Pioneering work on a (3D) virtual newsreader has been reported in (Magnenat-Thalmann & Kalra, 1995). Research issues that need to be tackled in order to obtain sophisticated embodied conversational agents (virtual humans) are discussed in (Gratch et al., 2002). General research on ECAs addresses issues related to facial animation, emotion display, gaze behavior, gestures, text-to-speech synthesis, prosody and, more recently, the ECA perceiving its conversational partner by using natural language processing, speech recognition and computer vision. This technology will allow more natural and social interaction (Bickmore, 2003). A different approach, meant to achieve more lively presentations, uses presentation teams where several agents discuss the properties of a product (Rist, André, & Muller, 1997). Presently there are modest attempts to replace fully script-based approaches by approaches that allow more interaction between presenter agents and users.

Embodied Guides and Presenters in 3D

Pioneering work on 3D virtual presenters has been reported in Noma and Badler (1997). They introduced a virtual human presenter that can make presentations in a 3D virtual environment or on the World Wide Web. It gets its input from speech texts with embedded commands that relate to the presenter's body language. The system aims at simulating a professional presenter, for example a weather reporter on TV. The presenter can interact with a visual aid (a 2D screen). Hence, it knows about the camera and the presentation screen, but does not have an audience in its direct environment. This simplifies the gaze behavior; the presenter will always look either at the screen or into the camera. In general a lecturer will need more complicated gaze behavior to look at its audience properly. Moreover, the animation model that is introduced requires extensions. For example, there are no posture shifts besides those needed to look at the screen and then back into the camera. In 'real' monologues or presentations, however, posture shifts also occur frequently at discourse segments (Cassell et al., 2001).

Figure 1: 3D museum guide (Ciger, 2005)

There are many ways to present information. When we ask directions on the street we can get quite detailed information, about how to go from the current location to a desired location. The explanation consists of verbal and nonverbal utterances, that is, sometimes the verbal utterances support pointing gestures or gestures that explain objects (landmarks) and situations that will be met, sometimes the gestures support the verbal utterances. A museum guide interacts verbally and nonverbally (using gestures and gaze) with his or her audience to explain the interesting parts of a sculpture or painting, addressing one, several or all persons in his or her audience, and using pointing (deictic) gestures to draw attention to details in a painting or an object. And, after having explained a piece of art, the guide will make it clear where to go or look next, again by verbal and nonverbal means of addressing the audience.

Various authors have introduced these 3D museum guides. The guides walk around in virtual environments and explain paintings or other works of art. They are also often used for providing the user with museum information, for giving directions to the visitor or to guide the visitor to a certain location. Interesting recent examples of such work can be found in Kopp, Gesellensetter, Krämer, and Wachsmuth (2005), in particular examples of conversations between such guides and visitors, and in Ciger (2005) on a guide that leads a visitor through a museum to a painting in which the visitor is interested.

Yet a different way of presenting information is to give a lecture or meeting room presentation, using a data or video projector. Clearly, also in this case we can expect that the presenter will use deictic references to pictures, bullets, and texts fragments that appear during a presentation, for example, a PowerPoint presentation, on the screen. Later on we introduce an embodied 3D presenter that explains verbally and nonverbally what is visible to the audience on a 2D object such as a PowerPoint slide or a painting.

Scripting and Mark-up Languages

Presentations by embodied agents rely heavily on scripting and mark-up languages. Obviously, here again, the emphasis is on the presentation contents and the relation between the contents and the artificial presenter, leaving out any possible interaction with an audience. A well-known example is the Multimodal Presentation Mark-up Language (MPML) (Descamps & Ishizuka, 2001) that also allows affective control of the presentation by incorporating models for emotion, mood and personality for the artificial presenter. The language is specifically designed for non-expert (average) users, allowing them to direct the behavior of multiple animated characters when creating web-based (interactive) presentations. The presentation modalities are specified at the level of actions (such as gestures, or speech) that can be synchronized in the script. For every action in the script an emotion that effects the execution of the action can be specified in the script. MPML is focused on helping users design presentations on web pages, using MS-agent commands and 2D avatars. In later work, those MS-agents' commands are used to specify behavior for a robot or a 3D virtual presenter (Nozawa, Dohi, Iba, & Ishizuka, 2004). The robot uses speech synthesis and a laser pointer to present the multimedia content on a computer screen. Later in this paper we will

return to mark-up languages. Notice again, that these scripting languages require the user to manually script the actions of the agents and the way they perform these actions. Automatic derivation of desirable agent presentation behavior from available multimedia sources is an obvious and important research area.

Presentations in Meeting and Lecture Rooms

In meeting and lecture rooms we have presentations. How can we support such presentations and what role can be played by virtual presenters? Moreover, what can we learn from the behavior of human presenters and how can we use that knowledge in designing virtual presenters and automatic, and preferably also interactive, presentations?

Meeting and Lecture Room Technology

A lot of meeting and lecture room technology has been developed in previous years. This technology allows real-time support to physically present lecturers, audiences and meeting participants, online remote participation of meetings and lectures, and off-line access to recorded lectures and meetings. We discuss these issues below:

- Smart meeting and lecture room technology that allows real-time support needs to be based on some level of understanding of what goes on in the smart room. Capturing without understanding is not new. There are some commercial tools, often related to PowerPoint, to capture lectures. Projects that have been devoted to capturing video and audio streams and that have much broader views on integrating, browsing and visualization are VACE (Einhorn, Olbrich, & Nejdl, 2003) and the Classroom 2000 project (Brotherton, Bhalodia, & Abowd, 1998). However, there are other important things that can be captured as well and that require understanding of nonverbal issues and multimodal interaction between people. Different levels can be distinguished and, of course, with increasing knowledge about verbal and nonverbal human behavior in multi-party interaction, it becomes possible to climb from a signal processing level to levels where we have a fusion of signals and, although presently far from comprehensive, semantic and pragmatic interpretation of interactions and events. There is no need to wait until comprehensive understanding can be achieved to design useful support tools that base their real-time support on captured information in the lecture or meeting room.
- Support for online, real-time and remote participation in meetings and lectures, will also be based on computational models of verbal and nonverbal human behavior during lectures and meetings. As an example, a remote participant may be alerted to pay attention because the smart lecture or meeting environment detects that the topic that is being discussed matches the remote participant's interests (as they appear in the user profile). As a second example, tracking a lecturer, audience members or meeting participants using localization techniques allows us to provide input to a virtual cameraman and a virtual director who take care that relevant interaction is captured by the cameras and microphones and that a remote participant is provided with multimedia views (text, video, virtual reality) on the interactions and events in the physical lecture or meeting room, possibly tuned to his interests and the context. Obviously, any information that can usefully be provided, possibly transformed for presentation purposes, can also be useful for the participants who are physically present in the lecture room, including the presenter.
- Providing off-line access to lectures and meetings allows for non-real-time manual and (semi-) automatic annotating (or indexing) of captured audio-visual information. Tools have been developed for manual, semi-automatic and automatic annotation of activities and events in smart environments. The annotations may concern directly observable features (nodding, direction of eye gaze, parts of speech), but they can also concern derived information where

features from different modalities are integrated and knowledge about the context and the history of events and interaction is included. As an example of how different modalities can come together, consider topic shifts during a talk. The text that is obtained from speech recognition (including information about pauses and prosody) can be subject to an analysis trying to distinguish topics and topic shifts in a lecture by looking at keywords, key sentences and topic change or topic introducing utterances, pauses and changes in prosody. Video analysis adds information about posture shifts, facial expressions and gestures related to topic shifts. The available presentation slides are another source for topic-shift information.

Annotations allow the building of tools for retrieval, summarization and regeneration. Presenting the results may involve transformations from particular input modalities to other modalities. That is, there is not necessarily a need to present the retrieved information using the same modalities that were considered in capturing it. Similarly, there is no need to confine ourselves to browsing facilities that only look at the raw data that has been captured. Rather we can look at enriched presentations that require the fission of modalities, for example, enriched regeneration of the lecture (talk, documents, interaction, audience, room) in a virtual reality environment that also gives the user the feeling of being present.

Multimodal Presentation Corpora

In the previous section we already mentioned annotations. In order to annotate we need corpora containing lectures and presentations. In the AMI project a corpus of meetings is being collected. These meetings contain whiteboard and PowerPoint presentations. In this latter case, we have a presenter explaining what is visible on a screen. The presenter's behavior is not essentially different (apart from the content) from a guide in a museum who explains a painting or a sculpture to a group of tourists visiting the museum. Gestures are made, there is pointing to the interesting parts, and there is some interaction with the audience, verbally and nonverbally. The corpus that was available at the start of the AMI project consisted of mock-up meetings, including, for example, someone standing up to deliver a presentation. During the AMI project new corpora will emerge, depending on the research interests of the various partners in this large-scale European project. One unstructured corpus that has been added is a series of thirty videos of presentations during a workshop associated with the project (http://mmm.idiap.ch/mlmi04/). In our own research it has been used to design models, in particular the pointing gestures, for multimodal human presentation. In more comprehensive research projects we are also looking at the development of tools that make it possible to relate spoken content with gestures. In addition, new annotation tools are being developed that also take into account pointing and other gestures that refer to parts of a scene, for example, specific regions on a sheet of a PowerPoint presentation (Welbergen, Nijholt, Reidsma, & Zwiers, 2005).

Video Management and Lecturer Support: Examples

Before exploring more ideas and approaches to on- and off-line access to lectures and meetings we spend a few words on more traditional research issues, that is, research which is not necessarily oriented towards designing virtual presenters, by discussing three examples of real-time support provided to a lecturer by smart agents and smart environments. The research presented in these examples is equally useful for the design of virtual presenters as well.

The first example concerns the audio-visual capturing of lectures and broadcasting audio and video to online and off-line (on-demand) audiences. This is already widely accepted in universities and corporations. A lecture can be captured by one or more cameras and microphones and there may also be other sensors (for example, wearable RFID: Radio Frequency Identification tags) involved that help to track activities of the presenter (Arseneau & Cooperstock, 1999) or members of the audience asking questions that can help to provide the on- and off-line audience

with a better understanding of the lecture. Obviously, when we confine ourselves to audio-video broadcasting, in an online situation we can only influence the broadcast by selecting and combining information coming from different cameras and microphones. For the cameras and microphones we can introduce 'intelligent' virtual cameramen and sound technicians that have knowledge that allows them to decide where to point the camera, when to zoom in and out, when to change the viewpoint, and who to listen to. That is, rather than hiring professional videographers we can build an automated camera and microphone management system (a virtual director) (Rui, Gupta, & Grudin, 2003), that can make decisions using knowledge about videography, lecture content, speaker, and audience members. For an off-line audience (or a virtual presenter selecting content to present) we can of course also introduce possibilities to access the lecture using content-based retrieval techniques for video fragments and video summarization techniques, based on manual or automatic indexing techniques.

The second and third examples concern the development of technology that supports a lecturer. A good example of the development of lecture support technology can be found in the EU-sponsored project FAME (Facilitating Agent for Multicultural Exchange) (Rogina & Schaaf, 2002). This work is about the development of a lecture tracker that tries to synchronize slides with recognized text. Slides and other documents that are used in the presentation are analyzed in advance and this analysis allows improvement of the speech recognizer's vocabulary and language model. During the presentation, slides and other documents can be presented when they correlate with the recognizer output. Hence, in the case of a slide presentation the slides follow the speaker rather than the speaker following the slides. Obviously, questions from the audience can also induce slide changes. A similar approach has been followed in the Jabberwocky system (Franklin, Bradshaw, & Hammond, 2000), where keywords and phrases are extracted from the slides and during the presentation the system listens to the speaker and tries to switch slides at appropriate moments. It would be interesting to see how research on rhetorical annotations, as presented in e.g., Bocconi et al. (2005), can help in generating compelling and enjoyable presentations.

Our third example concerns the research that is being performed in the CHIL project. CHIL - Computers in the Human Interaction Loop - is an Integrated Project on multimodality under the European Commission's 6th Framework Programme. CHIL is looking at human interactions in lectures and one of the research aims is to develop a so-called Attention Cockpit, an agent that tracks the attention of the audience and provides feedback to a lecturer. In this way the lecturer can be notified when the audience becomes disconnected or when a participant intends to ask a question. Potential services for the speaker that are foreseen are among others Audience Monitoring and Info (providing information about the background knowledge of the audience to the speaker and monitoring audience eye contact, facial expressions, body language and questions) and Slide Changing (changing slides automatically based on monitoring both the speaker's speech and questions asked) (Waibel, Steusloff, Stiefelhagen, & the CHIL Project Consortium, 2004). Other potential services take care of microphone control and informing the speaker about the quality of speech. Obviously, although CHIL is not concerned with designing a virtual presenter, research results for the Attention Cockpit can help us to model capabilities of the virtual presenter in our research. Presently CHIL research is in progress. One supporting action has been a survey conducted to collect data from people attending presentations and people giving presentations (Terken, 2004).

Cross-media Mappings of Captured Information

Audio and video are the main input sources for our research on multimodal capturing and fusion of information coming from different modalities. However, rather than assuming mappings from audio to audio and from video to video, possibly transformed and filtered because of audience

interests, audience context and available devices, we prefer to consider (reference-preserving) cross-media mappings of captured information in order to obtain the most effective and enjoyable multimedia presentation of a lecture and related events in a smart environment. As may have become clear from previous sections, whether it is for participants who are physically present (e.g. while being in the lecture room and looking back on part of the presentation or previous related presentations), for remote audience members or for off-line participants, multi-media presentation of captured information needs careful attention (Bachvarova & Elouazizi, 2005).

In previous research the possibilities of including in these multimedia presentations a regeneration of meeting events and interactions in virtual reality has been considered. There are various advantages to having this possibility. A meeting, whether online or off-line, can be attended from different viewpoints (literally), we can get more immersed in a meeting, the virtual reality representation allows better views on what happened or is happening, it allows for multimodal access, it is easy to add enrichments to the visualization and the virtual reality representations are easier to manipulate than the video-recordings. Other reasons to look at virtual reality representations have to do with being able to validate models of multi-party interaction and to do experiments on the role and importance of the different modalities and their combinations (Reidsma et al, 2005). Current scene capturing technology allows a translation from meeting activities into a 3D virtual reality regeneration of these activities. In such a reconstruction meta-information (information about the process, extra resources) can be added and made part of the visualization. Moreover, during reconstruction we can manipulate the information and make translations from one particular presentation medium to another.

Introducing Semi-autonomous Presenters

While in the previous section our starting point was the human presenter or meeting participant, we now look at a semi-autonomous virtual presenter in 3D that is designed to perform in a virtual reality environment. The presenter is assumed to perform for various audiences. That is, the virtual environment can be a lecture room or a museum that can be visited on the World Wide Web by a user using his desktop PC in order to attend a course or to get some explanation about a painting. There is not necessarily any other audience present. However, the user can also share the presentation with others. In a multi-user environment the audience can be visualized, they can see each other and they can communicate with each other during a presentation. Another possibility that we do not want to exclude is that the virtual presenter itself is obtained from real-time capturing of a human presenter, while simultaneously the captured video and audio is real-time regenerated in a virtual reality environment that can be looked at on a desktop PC or visited in a more immersive way (Nijholt, 2005).

We can consider this research as complementary to the research mentioned in the previous sections. There we assume sensors that can capture all kinds of useful information; however, to detect, fuse and interpret the information we need models describing the ‘why’ of activities and interactions. Most of the time this ‘why’ concerns the behavior of the human inhabitants of the environment. Together with these models we can transform the captured knowledge to answers, summarizations, browsing environments or replays using different media and modalities. Similar models are needed in order to generate realistic behavior of embodied agents, in our case a virtual presenter. In addition, behavior animation of a virtual presenter can be improved by adding motion capturing information from real presenters. See for example (Poppe, Heylen, Nijholt, & Poel, 2005) for the capturing of presenter activities in front of a screen or whiteboard with a camera. This will not be discussed here. That is, here we confine ourselves to models and associated algorithms from which to steer the presentation animations of a virtual presenter.

In our case, the presentations are generated from a script describing the synchronization of speech, gestures and movements. The script also has a channel devoted to slides and slide

changes; they are assumed to be an essential part of the presentation. Instead of slides, this channel may be used for the presentation of other material on a screen or wall. The script is described in a newly developed MultiModalSync language for synchronization and timing. This is illustrated in Figure 2.

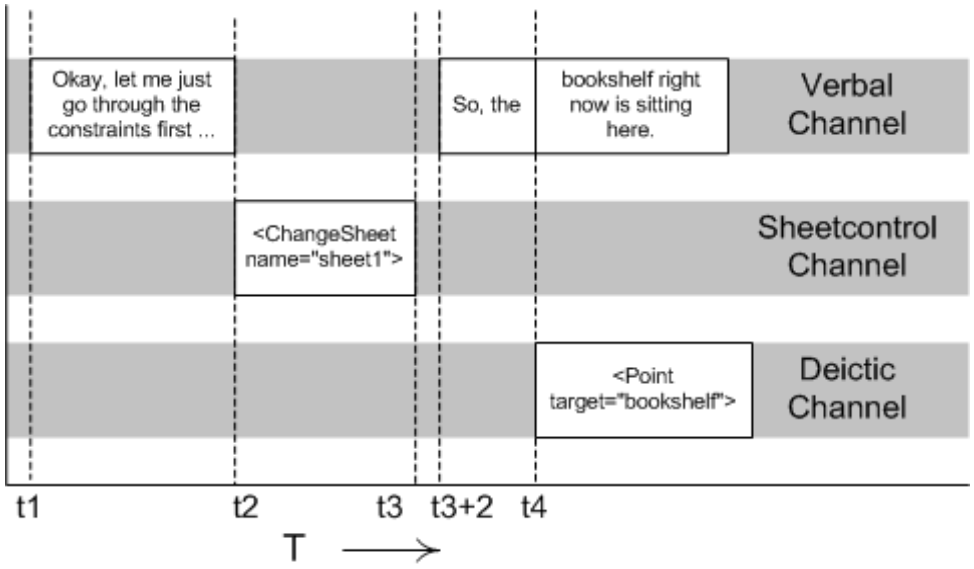


Figure 2: The MultiModalSync script language

The full architecture of the presenter is shown in Figure 3. An animation planner is responsible for the planning and playback of body animation. It makes use of movement models derived from neurophysics and behavioral science to perform this task. Currently, the animation planner is capable of playing deictic gestures, pose shifts and speech (mouth movement) specified in the script. Static speaker characteristics influence how this behavior is executed. The architecture can easily be extended to execute other gesture types. The verbal planner and the sheet planner regulate the text-to-speech generation and the sheet changes, respectively.

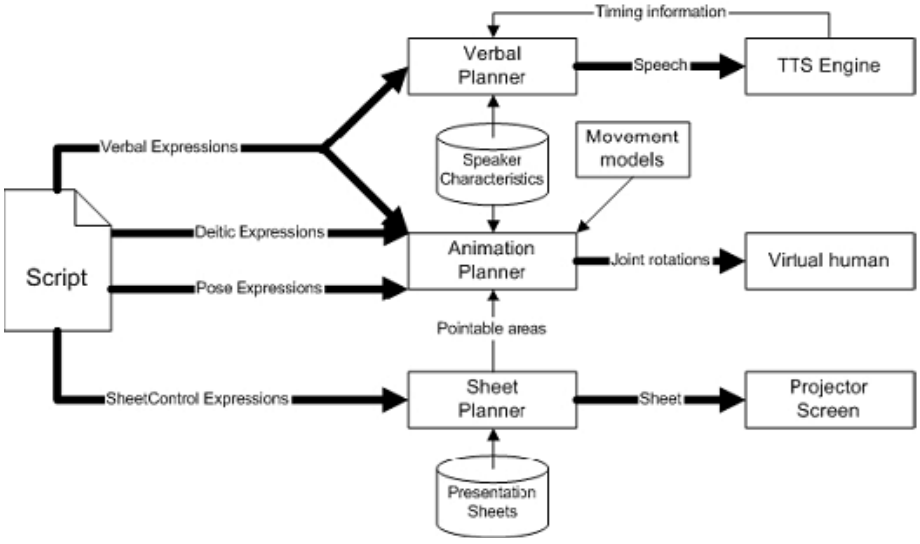


Figure 3: Architecture of the virtual presenter

In Figure 4 we have both the human presenter and our virtual agent presenting the layout of a room to an audience. As may become clear, we have not yet put efforts into having a good looking virtual presenter. The emphasis has been on designing and implementing presenter behavior. For technical details we refer the reader to Welbergen et al. (2005).

Figure 4: Human and virtual presenter explaining a room layout

Evaluating Virtual Presenters

The emphasis is this paper has been on 3D presenters. We looked at 3D presenters from different viewpoints. How do they represent human presenters that are being tracked by cameras and other sensors? How do they act as scripted embodied agents or as embodied agents that can act because their ability to perceive and interpret their environment? Because of these different viewpoints it is not possible to give a comprehensive discussion on how to evaluate and measure the success of virtual 3D presenters.

From a very global point of view we can ask how to compare the presentation behavior of a virtual presenter with the presentation behavior of a human presenter. Well, maybe it is not that much the behavior as well as the effect this behavior has on our understanding of the presentation contents and our enjoyment while attending this presentation. What did we learn from the presentation and how did it stimulate our interest in the topic and our willingness to explore the topic more deeply ourselves? If we want to make comparisons based on such criteria and want to evaluate virtual presenters in a similar way as we evaluate human presenters we need questionnaires in which we can find questions about the presenter's knowledge of the topic and, more importantly for the discussion here, questions that relate to the presenting skills of the presenter. In the latter case we can look at all kinds of questionnaires that are currently being used to evaluate the effect of presentations by human presenters. Questions are concerned with the content of the presentation, with the structure of the presentation, and with the delivery of the presentation. Evaluation can also take into account the response of the audience. From the point of view of this paper, we are in particular interested in the way a presentation is delivered by virtual and human presenters.

In current (human) presenter evaluation forms we are required to answer questions about how responsive, how organized, how clear and how timely the presenter was. Was the presenter well prepared, what did he do to keep participant attention, did he make good eye contact with the audience, did he take care he was visible for the audience, did he have distracting habits, did he interact with the audience, was he responsive for questions, et cetera. There is nothing wrong with having the same kinds of questions for virtual presenters. However, from the point of view of re-

search on embodied conversational agents (in our case, acting as virtual presenters) these questions can be considered as questions that relate to the application domain, and more questions and related research still have to be asked about, for example, the believability of an embodied agent acting as a presenter (Theune, Ham & Heuvelman, 2005), the social and intelligent skills such an agent has (and how they are displayed), the interaction modalities that are made available, et cetera. What is the effect of a male presenter versus a female presenter? How is this effect related to the task domain, the context of use, and the user group?

These questions and observations translate of course to the behavior of an embodied presenter in its environment and in the presence of an audience. That is, they need to be translated to animations (movements of the body, pose, head orientation, gestures, facial expressions, and speech characteristics) and rational and emotional content of feedback and pro-active behavior of a virtual presenter. Evaluation from the point of view of tasks to be performed in an application domain is important, but there are still many questions to be answered that relate to more theoretical and technical problems in the area of animations, verbal and nonverbal communication, visualization and artificial intelligence (Welbergen et al., in press). Moreover, many of these issues are intrinsic to the multi-disciplinary nature of embodied agent research (Ruttkay & Pelachaud, 2004).

Conclusions and Future Research

We surveyed the research issues in designing virtual presenters that are, in our view, the most important to tackle. Clearly, graphics and animations are important; more important, however, is to have behavior and derived animations generated from what has to be presented and the available multimedia resources.

It is necessary to look at human presenters to see what we can learn from them. This requires capturing the presentation behavior of human presenters. In this way we can collect a corpus of presentation behaviors that can be analyzed and from which models of behavior can be constructed and verified. The corpus can also be used to distinguish training and test sets and to employ machine learning approaches to the task of learning models. Obtained models and algorithms allow us to replace manual annotation of presenting behavior by semi-automatic or automatic annotation. On the one hand this allows us to provide real-time support to a human presenter; on the other hand, it allows us to define relations between presented material and presentation behavior. This is also essential information for our virtual presentation agents.

In current research projects corpora of presentations are being collected. More effort is needed in order to develop annotation schemes and to perform the actual annotations. We also see efforts to annotate (index) media items that can be used during presentations. Automatic scripting is a returning aim.

Apart from the research already mentioned we are also looking at ways to support a presenter in a situation where, instead of using a detailed presentation, during his presentation, the presenter gets the freedom to follow new thoughts requiring the retrieval of slides, sound fragments, photos or videos, triggered by words or phrases used by the presenter and obtained from speech recognition.

Acknowledgements

We are grateful to Herwin van Welbergen who designed the MultiModalSync language (together with Job Zwiers) and is now working on providing our presenter agents with smooth animations. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication 159).

References

Arseneau, S., & Cooperstock, J. R. (1999). Presenter tracking in a classroom environment. *IEEE Industrial Electronics Conference (IECON'99)*, *Session on Cooperative Environments, 1*, 145–148.

Bachvarova, Y. & Elouazizi, N. (2005). A Conceptual Argument for Modality Ontology to Support Automatic Modality Assignment. Proceedings of the *ICMI 2005 Conference workshop Multi-modal Interaction for the Visualization and Exploration of Scientific Data,* Trento.

Bickmore, T. (2003). Relational agents: Effecting change through human-computer relationships. PhD Thesis, Media Arts & Sciences, Massachusetts Institute of Technology.

Bocconi, S., Nack, F. & Hardman, L. (2005). Using rhetorical annotations for generating video documentation. Proceedings of the *IEEE International Conference on Multimedia and Expo (ICME) 2005*, July 2005. Available at http://www.icme2005.org/

Brotherton, J. A., Bhalodia, J. R. & Abowd, G. D. (1998). Automated capture, integration, and visualization of multiple media streams. Proceedings of the *IEEE International Conference on Multimedia Computing and Systems, 54* – 63.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C. & Rich, C. (2001). Annotating and generating posture from discourse structure in embodied conversational agents. In *Workshop on representing, annotating, and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents,* Autonomous Agents 2001 Conf. Montreal, Quebec.

Ciger, J. (2005). Collaboration with Agents in VR Environments. Ph.D. Thesis, EPFL, Lausanne.

De Carolis, B., De Rosis, F., Carofiglio, V. Pelachaud, C. & Poggi, I. (2001). Interactive information presentation by an embodied agent. Proceedings of the *Class Worksho*p, Verona.

Descamps, S. & Ishizuka, M. (2001). Bringing affective behavior to presentation agents. Proceedings of the 3rd *International Workshop on Multimedia Network Systems (MNS2001)*, IEEE Computer Society, Mesa, Arizona, 332-336.

Einhorn, R., Olbrich, S. & Nejdl, W. (2003). A Metadata Model for Capturing Presentations. Proceedings of the 3rd IEEE *International Conference on Advanced Learning Technologies (ICALT 2003)*, 110-114.

Franklin, D., Bradshaw, S. & Hammond, K. (2000). Jabberwocky: You don't have to be a rocket scientist to change slides for a hydrogen combustion lecture. Proceedings of the *Intelligent User Interfaces 2000 (IUI-2000)*, 98-105.

Gebhard, P. (2001). Enhancing embodied intelligent agents with affective user modeling. In J. Vassileva & P. Gmytrasiewicz (Eds.), *UM2001, 8th International Conferenc*e (271-273). Berlin: Springer.

Gratch, J. Rickel, J., André, E., Badler, N., Cassell, J. & Petajan, E. (2002). Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, 54-63.

Kopp, S., Gesellensetter, L., Krämer, N. & Wachsmuth, I. (2005). A conversational agent as museum guide - Design and evaluation of a real-world application, *The 5th International Working Conference on Intelligent Virtual Agents (IVA'05)*, Springer, 329-343.

Magnenat-Thalmann, N. & Kalra, P. (1995). The simulation of a virtual TV presenter. Proceedings of the *Pacific Graphics 95*, World Scientific (9-12), Singapore.

Nijholt, A. (2005). Meetings in the virtuality continuum: Send your avatar. In T.L. Kunii et al. (Eds.), *Proceedings of the 2005 International Conference on CYBERWORLDS* (75-82). Los Alamitos: USA; Singapore: IEEE Computer Society Press.

Noma, T. & Badler, N. I. (1997). A virtual human presenter. *Proceedings of the IJCAI-97 Workshop on Animated Interface Agents*, 45-51.

Nozawa, Y., Dohi, H., Iba, H. & Ishizuka, M. (2004). Humanoid robot presentation controlled by multimodal presentation markup language MPML. Proceedings of the *13th IEEE Int'l Workshop on Robot and Human Interactive Communication (RO-MAN2004)*, Kurashiki, Japan, No.026.

Prendinger, H. & Ishizuka, M. (Eds.). (2004). *Life-like characters. Tools, affective functions, and applications*. Berlin Heidelberg: Cognitive Technologies Series, Springer.

Poppe, R., Heylen, D., Nijholt, A. & Poel, M. (2005). Towards real-time body pose estimation for presenters in meeting environments. In V. Skala (Ed.), Proceedings of the *13-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005*: Short Papers, University of West Bohemia, Plzen, Czech Republic, 41-44.

Reidsma, D., Akker, R. op den, Rienks, R., Poppe, R., Nijholt, A., Heylen, D. &. Zwiers, J. (2005). Virtual meeting rooms: From observation to simulation. In R. Fruchter (Ed.), Proceedings of the *Social Intelligence Design 2005*, Stanford University, Stanford, CA, USA, CD Proceedings.

Rist, T., André, E. & Muller, J. (1997). Adding animated presentation agents to the interface. *Proceedings of Intelligent User Interfaces*, 79-86.

Rogina, I. & Schaaf, T. (2002). Lecture and presentation tracking in an intelligent meeting room. Proceedings Fourth *IEEE International Conference on Multimodal Interfaces*, 47-52.

Rui, Y., Gupta, A. & Grudin, J. (2003). Videography for telepresentations. *CHI Letters*, *5*(1), 457- 464.

Ruttkay, Zs., & Pelachaud, C. (Eds.) (2004). *From brows to trust. Evaluating embodied conversational agents*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Sluis, I. van der. (2005). Multimodal reference. Studies in automatic generation of multimodal referring expressions. Ph.D. Thesis, Tilburg University, the Netherlands.

Terken, J. (Ed.). (2004). Report on observation studies with requirements for CHIL services. Deliverable D7.1 of the Project CHIL (Computers in the Human Interaction Loop) IP 506909.

Theune, M., Heylen, D. & Nijholt, A. (2005). Generating embodied information presentations. In O. Stock & M. Zancanaro (Eds.), *Multimodal Intelligent Information Presentation* (Chapter 3, 47-40). Kluwer Series on "Text, Speech and Language Technology", Vol. 27, Kluwer Academic Publishers.

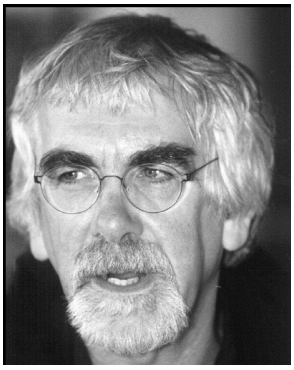
Theune, M., Ham, R. ter, & Heuvelman, A. (2005). Can you teach an old dog new tricks? How older users react to embodied agents. *Dutch SIGCHI conference,* The Hague, 2005.

Waibel, A., Steusloff, H., Stiefelhagen, R. & the CHIL Project Consortium. (2004). CHIL - Computers in the human interaction loop. 5th Intern. Workshop on *Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal. See also http://chil.server.de/

Welbergen, H. van, Nijholt, A., Zwiers, J. & Reidsma, D. (2005). Presenting in Virtual Worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. In M. Maybury, O. Stock & W. Wahlster (Eds.), Proceedings of the *Intelligent Technologies for Interactive Entertainment (INTETAIN'05)* (203-212), Lecture Notes in Artificial Intelligence 3814. Berlin Heidelberg: Springer-Verlag.

Welbergen, H. van, Nijholt, A., Reidsma, D. & Zwiers, J. (in press). Presenting in virtual worlds: Towards an architecture for a 3D Presenter explaining 2D-Presented information. *IEEE Intelligent Systems*, September/October 2006.

Biography



Anton Nijholt is full professor of Computer Science and chair of the sub-department Human Media Interaction of the Department of Computer Science of the University of Twente. His main research interests are multi-party and multimodal interaction, virtual environments and social and intelligent (embodied) agents. Before joining the University of Twente he held positions at the Vrije Universiteit Brussel and several other universities in the Netherlands and Canada. He received his Ph.D. from the Vrije Universiteit of Amsterdam and did his M.Sc. Thesis at Delft University of Technology. Contact him at the University of Twente, Human Media Interaction, PO Box 217, 7500 AE Enschede, The Netherlands; anijholt@cs.utwente.nl.