Virtual Meeting Rooms: From Observation to Simulation

Dennis Reidsma, Rieks op den Akker, Rutger Rienks, Ronald Poppe, Anton Nijholt, Dirk Heylen, and Job Zwiers

Twente University, Enschede 7500AE, the Netherlands,
dennisr@ewi.utwente.nl,
www home page: http://hmi.ewi.utwente.nl/

Abstract. Virtual meeting rooms are used for simulation of real meeting behavior and can show how people behave, how they gesture, move their heads, bodies, their gaze behavior during conversations. They are used for visualising models of meeting behavior, and they can be used for the evaluation of these models. They are also used to show the effects of controlling certain parameters on the behavior and in experiments to see what the effect is on communication when various channels of information - speech, gaze, gesture, posture - are switched off or manipulated in other ways. The paper presents the various stages in the development of a virtual meeting room as well and illustrates its uses by presenting some results of experiments to see whether human judges can induce conversational roles in a virtual meeting situation when they only see the head movements of participants in the meeting.

1 Introduction

The state of the art in computer graphics, and embodied conversational agents allows one to build quite realistic 3D virtual environments in which virtual, human-like avatars simulate various behaviors that occur in real meetings. Much working time is spent in meetings, they are subject of various research disciplines. AMI is an European Research Project that aims at developing new technologies for supporting meeting activities: meeting browsers, and technology that makes remote meeting participation easier, more effective and more natural.[1] The Human Media Interaction Research Group (HMI) of the University of Twente is one of the partners of AMI [Nijholt et al., 2004]. UT-HMI has a tradition in research in interaction with embodied conversational agents, research in computer graphics for virtual environments and machine learning techniques for recognition of higher level features (such as dialogue acts, gestures, emotions) from lower level features (such as words, hand arm movements, facial features). Based on our agent platform (implemented in *Java*, using *Java3D*, *X3D* and *XML*-technology, using *H-ANIM* standards for human body animation) we have constructed a 3D virtual replica of one of the meetings rooms that is used for data collection of

[1] AMI - Augmented Multi-party Interaction.

meetings in the AMI project. This *virtual meeting room* (VMR) is useful for various purposes that can be grouped into the following three categories:

1) Presentation of multi-media information about meetings. Information can be directly obtained from recordings of behaviors in real meetings -tracking of head, hand, arm and body movements- or output by some machine model that induces features from directly recorded or annotated meeting data; but one could also think of a 3D summary of real meetings. These presentations can be used for evaluation of results obtained by machine learning methods.
2) Research in human social interaction, recognition, and interpretation of visualized information. Virtual Environments allow to control various independent factors and can be used to study how they influence features of social interaction and social behavior.
3) As an (immersive) virtual environment, a communication means for real-time remote meeting participation. Real-time presentation of a virtual model of a meeting reduces the amount of data that has to be sent to and displayed on client side remote displays. It will be clear that this builds on knowledge about what types of events and behaviors in the real meeting are essential to be presented in the virtual meeting in order to maximize the quality of those impressions that are required by the user given his task and role in the meeting, such as the feeling of presence, and the possibility of mutual gaze.

In this paper we concentrate on the use of the VMR for establishing requirements of their use for remote meeting participation (see item 2) above). We want to know how important the various information "channels" (voice, gestures, head movements, body postures, facial expressions, etc.) in meetings are in order to interpret the conversational situation and the interactions. Results of such studies may add to the insight in what are the technical requirements of virtual meeting rooms as means for remote meeting participation. The way we go from real meetings to virtual meetings is from *observations* of meetings, multi-channel audio, visual and human *recordings* of real meetings, through *annotations* of various aspects of behaviors and actions in meetings -described on different levels depending on the views we have on the meeting- to meetings in a virtual meeting room. These stages are mediated by models describing the relations between the different levels. Based on hand annotated meeting data several machine learning techniques are used for inducing features that are not directly observable and these features of behavior can then be visualized in the VMR as well.
As an illustration of these uses of our VMR we present results of an experiment we performed with human judges who were asked to predict who the speaker is in a virtual meeting situation simulating a real meeting situation. The judges were presented head orientations of participants in the meeting. It turns out that humans have a hard job in doing so. This experiment is part of a series of experiments we are doing in order to compare results for different settings in which we control the selection of types of conversational signals provided to the user.

1.1 Organisation of this paper

The paper is organized as follows. In section 2 we will present our view on the development of the concept of meeting in interaction with technological and societal developments. The aim of that section is to position our research on VMRs in the context of Social Intelligence Design. In section 3 we present a schematic overview of the proces from observation to simulation and we discuss possible uses of virtual meeting rooms in section 4. As an illustration of the scheme we focus in section 5 on some experiments we did with the VMR.

2 Meetings and Technology

In a general sense a meeting is any coming together, willingly or unwillingly, of two or more people at such a close distance of each other that they are aware of each others presence and, willingly or unwillingly, react on that. The concept of distance, and related to that the concept of being in the same meeting room, has strongly been developed and is still being renewed by the development of technology in the last few centuries; in particular by the developments of communication and information technology. We shouldn't forget that this is really a process of conceptual development, in which the content of sharing the same space evolves from physically sharing the same space to sharing mentally the same space, where we identify invariantly a number of central themes: fight for the individual privacy, respecting each others private space, the need of being respected by others, the will to express one self and one's ideas and to realize individual goals. The impact of technology on meetings can not be described adequately in terms of quantitative measurable effects it has on properties of processes that occur in existing forms of meetings; technology develops the very idea of meeting itself, it has impact on how people realize the idea of meeting. Moreover, what is essential for meetings is that technology offers new perspectives on language and communication, on human perception and on social interaction. These new perspectives may help us to gain more insight in the essential qualities of these aspects of social reality. *In a more restricted sense* a meeting is an organized process of people coming together focussing on a common topic or task, something put on the meeting agenda. Meeting in this more restricted sense is one of the characteristics of the modern way we organize our work in all kinds of organizations. Meeting in this sense is work, and more and more work time is spent in meetings. Meetings in this more restricted sense, however professionalized and organized they may be, are still peoples meeting and all the themes that play in the more general sense of meeting can be identified in these meetings as well, be it often in more organized, more conventional forms, mediated by invented rules of good conduct: turn taking behavior, addressing behavior, politeness rules, and dominance relations.

Meetings are important means by which various types of groups try to accomplish their group tasks and at the same time try to maintain the groups coherence. The question here is what are the essential aspects of communication required for performing these tasks and how they can be realized in new ways

of remote meeting participation.
In communication we have the following 'ingredients' that could be more or less important to be visually presented.

- The participants in the meeting that express their ideas, discuss, and make decisions. Head movements, and body postures, facial expressions, emotions, conversational gestures, actions, locations.
- The material that supports the presentation or discussion: documents, presentations, whiteboard, the shared work space.
- The objects the interactants talk about, act on and verbally or non-verbally refer to.

The importance of visualising each of these ingredients strongly depends on the function that the meeting has for the group for which the meeting is a means to perform their tasks. Are there presentations, discussions, brain storming? Are decisions being made, are commitments of individual participants to do certain tasks an issue, is the functioning of the group itself or of individual members an item on the agenda? Do people refer to objects, or designs, or documents, that need to be visualized and that people interact with? For what specific tasks what type of meetings are most effective? For what type of meeting activities does video channel have added value? Some research has been reported in for example [Bailenson et al., 2001, Bailenson et al., 2002] or [Fussell et al., 2000, Kraut et al., 1996]. How important is the feeling of presence in virtual meetings and how can this be obtained? [2]
Technological and social developments interact with philosophical reflection on the social phenomena of our technological society and shed new light on the central concepts involved. It is quite hard to foresee what impact new technologies on everyday live will have and what their chances are to survive. Time will show whether immersive virtual environments that allow humans to interact remotely with other humans as well as with computer generated embodied conversational agents have added value over already existing means for tele-meetings and cooperative work spaces for non-collocated groups.

3 From Observation To Simulation

In this section we describe the process from observation through annotation to simulation and the various models that describe the relations between the annotated features of verbal and non-verbal conversational behavior.

[2] "Presence means that the user constructs a mental spatial model out of virtual stimuli and the perception of the self in the virtual environment." "Presence (..) describes the cognitive process of constructing an environment. As a result of this construction, the user experiences a sense of presence, that is the user feels him- or herself as part of the virtual environment. Since the body is real, the "realness" of the virtual environment is inferred. Users describe that they are "there" and that the virtual stimuli can have actual effects on behaviour and emotions." [Regenbrecht and Schubert, 1997].

3.1 Annotations of Behavior in Meetings

Within the AMI meeting project we see a huge effort in meeting data collection, meeting data annotation and dissemination of these data for various multidisciplinary research purposes inside and outside the project. One hundred hours of meeting recordings are planned for of which about 60% are scenario based meetings with four people meeting four times as part of a design project in which they have to work on a prescribed task to develop a remote tv control unit. Participants have various roles in this play and in order to meet reality as best as possible, external events and information are brought in that may influence the decision making process as well as the outcome of the meetings.

The hundred hours of recordings will be annotated in varying levels of detail for different dimensions. There are several reasons for creating manual annotations of corpus material. In the first place ground truth knowledge is needed in order to evaluate new techniques for automatic recognition of those same aspects. In the second place, as long as the quality of the automatic recognition results is not high enough, only manual annotations provide the quality of information needed to do research on human interaction patterns (see also section 4.2).

The annotations can be organized in layers of increasing complexity. The lowest layers describe mostly the *form* of the interactions, or the observable events. The higher layers describe interpretations of these observable events, giving the *function* of the interactions. Consider for example the situation where a participant raises his or her hand. The form of this gesture can be observed and annotated as 'hand raising'. On an interpretation layer, this event may be annotated with the function of this gesture, such as 'request for a dialogue turn' or 'vote in a voting situation'.

Examples of layers that can be annotated are:

- hand and body postures
- labelled gestures (interpretation of movement and pose)
- speech transcription
- communicative acts
- argument structures / topics
- summaries

3.2 The Virtual Meeting Room

Figure 1 shows a screen shot of the meeting room together with three different views of the HMI Virtual Meeting Room. The annotations described in the previous section can be replayed in the meeting room in different ways. Replay can show all available annotated information (down right in picture, a shot that shows head orientation, recognised body pose, current speaker and addressees of utterance) or only a selection (down left shows for example only head orientation). This section describes the general process. The next section shows

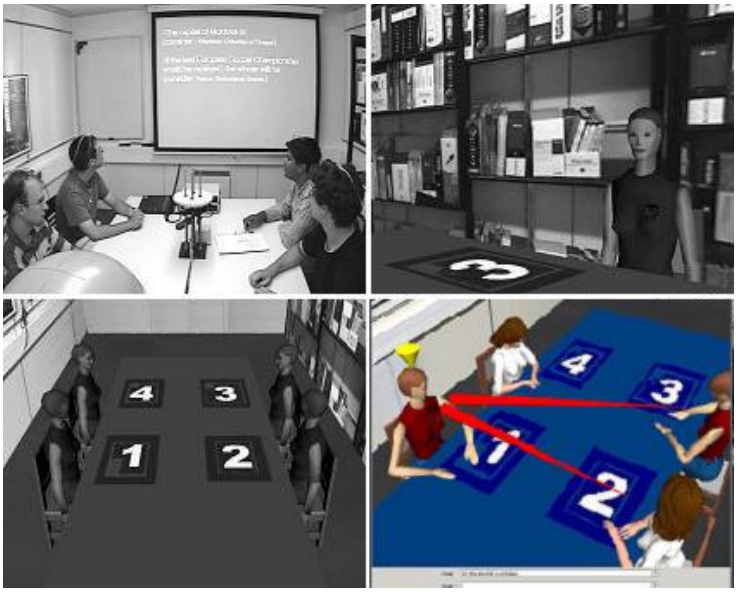

Fig. 1. Real and virtual meeting room. Upper left: the real meeting room with participants with flock of birds sensor for recording exact head movements. The other three pictures show three different views of the virtual meeting room: central view (down left), view from the eyes of a participant (upper right) and an extended view with visualisation of head orientation, recognised body pose, current speaker and addressees of utterance (down right)

how different types of replay are suitable for widely different uses, ranging from remote meeting participation to validation of models of social interaction.

Figure 2 shows an abstract view of the Virtual Meeting Room 'observation to simulation' process. The left hand side depicts the observation and interpretation. Human interactions in meetings are recorded on video and audio. Observation of these videos leads to descriptions of observable events (body movements, joint angles, facial expressions, speech, etc). These observations can be interpreted on progressively more complicated levels (see also [Reidsma et al., 2004]). Examples of these layers, more or less ordered on their level of interpretation, are sound, movements and facial configurations; gestures, words and facial expressions; communicative acts; argument structures; intentions, desires and knowledge.

The right hand side depicts the simulation process. At a certain point, the information from the annotations is used to play back, regenerating the lower level information from models of human interaction (see also section 4.1).

The rules for generation of communication are derived from domain knowledge (models and theories of human interaction) collected through the analysis of large amounts of data from real world examples. Examples are models for

choosing modalities, realizing gestures or speech, formulating sentences, deciding on communicative goals given beliefs and intentions, choosing communicative actions based on goals, etc.

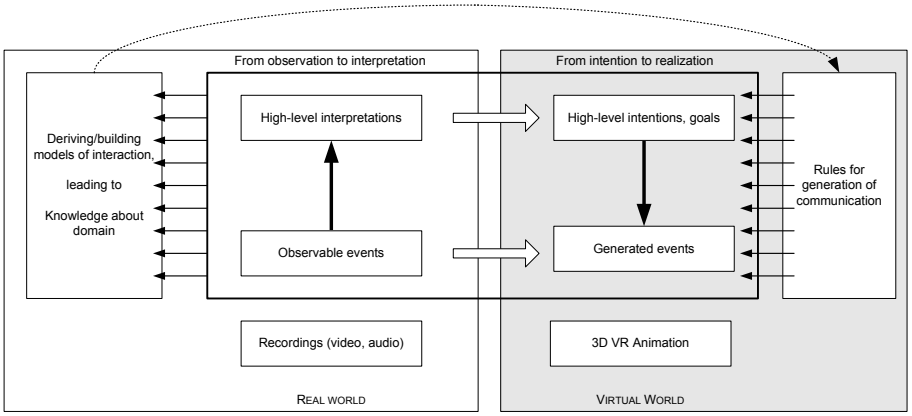

Fig. 2. Schematic overview of the various steps from observations and recordings via annotations to simulations mediated by various models expressing the relations between the aspects of verbal and non-verbal conversational behavior of participants in the meeting.

4 Uses of the Virtual Meeting Room

This section goes into a little more detail concerning some of the uses of the virtual meeting room. In the introduction it was already mentioned that this paper will focus on three categories of VMR applications: visualisation of multi media information from meetings for several purposes, elicitation and validation of models for social interaction and as an environment for teleconferencing that provides a sense of immersion and presence.

4.1 Re-visualization of meetings

Using a general implementation of a VMR it is possible to re-visualize the contents of a recorded meeting. This can be done literally, trying to stay as close to the original recordings as possible, or more conceptually, aiming for a visualization that reflects the meaning of the meeting rather than the actual form. The re-visualization process traces a path through figure 2. This path starts at the bottom left corner (real world / video recordings), and first goes upwards through various stages of observation and interpretation. At a certain point the

transition to the right part of the model is made (in a sense 'copying' the information present on the left side to the right hand box on that level), after which the traject of generation is followed down to produce an animation of the meeting in the virtual meeting room (bottom right).

This transition can be made at many different levels. Doing this at the lowest levels (Figure 3) is already interesting: replaying recognized 3D joint angles in a VMR in parallel with showing the original video offers a kind of quick-and-ready validation of the pose recognition process. If the recognition is good enough to use as input for a gesture labelling algorithm but not good enough to give convincing replay results, the transition can be made at a higher level. After interpreting the movements as labelled gestures, the replay is created from these gesture types rather than directly from the body poses, leading to an animation that is less close to the original video but more clearly expresses the *meaning* of the movements.

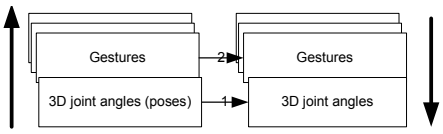

Fig. 3. From observation to simulation on different levels of interpretation

Other levels where the transition can be made are that of communicative actions (the simulation can then use different realizations for the communicative actions to stress an aspect of the meaning, or use another language, including the appropriate different culturally determined gestures) or the layer of summarized arguments and opinions. The last possibility is especially interesting if one wants to achieve *summarized replay* of a meeting or set of meetings. If the discussion about a certain issue was spread over more fragments of several meetings, at a certain level of interpretation the main structure of the arguments are found. Making the transition at this level will result in a new, *interpreted* replay of the discussions. If the models for simulating interaction are good this leads to a coherent direct visualization of a discussion between meeting participants that captures the main points of the original meeting (who proposed what, who was for / against, who used / supported which arguments, etc), without the redundancies of arguments that occurred in reality. The form of the simulation will deviate much from the original video recording in that case.

4.2 Validations of models of social interaction

This section describes applications of the VMR that are much more related to social intelligence. If autonomous agents must display believable social behaviour, there are many communicative aspects to be taken care of. For such aspects

models are needed: in what circumstances are which communicative actions desirable? How does a person show whom he/she is addressing? Does it depend on status differences? What is acceptable behaviour for an Embodied Conversational Agent (ECA) to show that he/she is listening to the user and interested in what the user says? The virtual meeting room provides ways to both elicit and validate such models. The following paragraphs give a few examples of this, of which at least the first will actually be done in the HMI virtual meeting room. A few other experiments that use virtual environments for elicitation and/or validation of models of interaction can be found in [Bailenson et al., 2001] and [Slater and Steed, 2001].

Validate models of addressing behaviour In [Jovanovic and op den Akker, 2004] Jovanovic et al. give an outline of their plans for research on modelling and detection of addressee in conversation. When such models have been developed, based on corpus annotations, these will be validated in the HMI virtual meeting room. A possible way to achieve this is to let an ECA simulate a fragment of conversation, expressing the addressee of utterances in one of the many ways allowed by the model (using vocatives, gaze, etc). A human judge, immersed in the VMR, will then be asked to assess who is the addressee of utterance. This experiment can provide the validation whether a model of addressing behaviour is good enough to use in an ECA, insofar as that a human will understand its addressing cues.

VMR Turing test The VMR Turing test (adapted from [Bailenson et al., 2004] allows one to validate a complex set of models, testing whether they result in convincing, natural social interaction by ECA's.

It works as follows: show a human subject a VMR with avatars controlled by other humans and avatars controlled by an ECA. Remove from the human avatars all communication channels that the ECA doesnt have (for example face expressions). Ask the subject to judge which avatars are controlled by the ECA. This can be done for any level of complexity (for example, you can validate models of listening behaviour by having the subject talk to two humanoids of which one is ECA and one is human and see whether the subject can tell which is which, if both are not allowed to talk back).

Turn taking A experimental setup similar to the one described for validation of addressee models can be used to validate certain patterns of nonverbal behaviour related to *turn taking*, simulated in [Padilha and Carletta, 2003]. The models can be applied to generate turn taking behaviour in a number ECA's; a human judge can be instructed to assess which person is the next speaker, the previous speaker, etc. Models that lead to a better prediction are more suitable for use in ECA's.

4.3 Remote participation and enhancement of meetings

An example of the possibilities offered by the virtual reality aspects of the VMR is based on the fact that different meeting participants need not necessarily

all have the same view of the virtual environment. This means that different participants can have a different perception of the seating arrangements. Since it is known that some positions are more advantageous in terms of discussion impact than others, it might be sensible to give each participant such a view of the seating that he or she never feels to be in the most disadvantageous position, leading to all participants feeling more comfortable during the meeting.

A virtual teleconferencing environment offers the possibility to introduce autonomous agents in a meeting that have the same communicative channels at their disposal as the human participants. This gives opportunities for defining experiments to discover regularities in human social interaction, as already described in previous section (validation). It also facilitates the introduction of *helper agents* into an actual meeting. Existing work has already shown that people can be influenced in their behaviour as well as their assessment of a situation by the presence of autonomous agents and their behaviour, even if they know that the agents are not representing a real human [Pertaub et al., 2002]. The emergence of advanced recognition technology for human interaction, partly developed from extensively annotated corpora, will allow embodied conversational agents to use this fact to influence the course of the meeting. A simple example would be the introduction of a virtual chairman in the meeting room with a regulating task. An enhancement of this chairman would be possible if the recognition technology gets advanced enough to detect potentially tense situations: the virtual chairman could try to defuse such situations by making a joke, or changing the subject of discussion. Another, ethically highly dubious, example is to include a virtual participant who listens very attentive whenever person A is saying something, and grows bored and restless whenever person B is saying something, in order to increase the status and believability of person A.

5 Example of an experiment in the VMR: speaker prediction from head orientation by human judges

Social intelligence is very much related to both the understanding and generation of nonverbal communication. Our VMR allows us to analyze nonverbal communication and its social intelligence aspects between the inhabitants of a virtual meeting room, and it allows us to generate and validate social interaction behavior from models of social intelligence. Since different communication channels (gestures, head movements, facial expressions, etc.) can be controlled in the VMR, we are able to zoom in on the social intelligence properties of one particular modality, to leave out other modalities, and to study any combination of modalities. We are particularly interested in generating believable models of speaker and addressee behavior in meetings. To gain more insight into this behavior, we are studying the impact of various modalities on the prediction of both speaker and addressee. Here we describe an experiment where human judges were asked to predict the speaker, solely given the head orientations of all participants.

A number of researchers reported studies concerning the functions of gaze and mutual gaze in conversations. According to Kendon [Kendon, 1967] gaze serves four functions: visual feedback, regulate conversational flow, communicate emotions and relationships and to improve concentration by restricting visual input. Gaze behavior of speakers, addressees and overhearers is related to turn-taking and turn-giving behavior as well as to addressing behaviors, behaviors that speakers show when they are addressing their speech to one or a selected subgroup of participants in a meeting (see [Jovanovic and op den Akker, 2004]). Since recording eye gaze without being obtrusive is quite hard, see e.g. [Vertegaal, 1998], *head orientation* is often used as indication of gaze and focus of attention. Stiefelhagen showed that head orientation can be used to inform about the participants gaze in meetings [Stiefelhagen and Zhu, 2002], [Stiefelhagen, 2002], [Stiefelhagen et al., 2001]. There is a rather high correlation (more than 85%) between gaze and head orientation. All in all we can expect that head movements of participants in a meeting may be used as an indicator of who is speaking and to whom someone is speaking.

In the meeting room at IDIAP in Martigny three four-person group meetings have been recorded where participants had the task of debating several statements[3]. Participants wore an electro-magnetic sensor on their heads so that the exact head position and orientation could be recorded. The meetings were further audio and video recorded. After recording, biases in head orientation due to incorrect mounting of the sensor on the head were removed. From the obtained data set, all occurrences with non-speech (laughter, silence, etc.) or with speech overlap were removed. Table 1 shows the number of frames for each meeting, the prior speaker probabilities of $P(Sp = p_i)$ for each of the 3 meetings separately and for all meetings in total.

	M1	M2	M3	Total
Samples	11333	13078	28148	52559
A priori p1	40.4%	26.9%	24.8%	28.7%
A priori p2	27.3%	23.4%	9.8%	16.9%
A priori p3	7.7%	8.6%	29.4%	19.5%
A priori p4	24.7%	41.2%	36.0%	34.9%

Table 1. Number of samples and prior speaker probability $P(Sp = p_i)$ in meeting M_j

5.1 Speaker prediction with a Naive Bayes classifier

Speaker prediction based on head orientation essentially boils down to computing the maximum a posteriori probability of a person speaking given the head

[3] In AMI jargon, these meetings do not belong to the *core* meeting data collection; they are *spoke* recordings, for research interests of individual AMI project partners.

orientation of all participants. We determined which person was looked at by person placing boundaries in the azimuth angle range. This discretization is arbitrary in that a person is always looking at another. Whiteboard and other possible places of interest are ignored this way. Function $LA(i)$ gives the person $v_i \in \{1, 2, 3, 4\}$ who is looked at by person p_i. The current speaker is Sp. We obtain our speaker prediction by choosing the person p who yields the highest probability given the situation v_i ($i \in \{1, 2, 3, 4\}$):

$$\underset{j}{\text{argmax}} \; P(Sp = p_j \mid LA(1) = v_1, LA(2) = v_2, LA(3) = v_3, LA(4) = v_4)$$

Assuming conditional independence of the different $LA(i)$ variables given the value of Sp (the independence assumption characteristic for the Nave Bayes classifier) we compute:

$$\underset{j}{\text{argmax}} \; P(Sp = p_j) \prod_i P(LA(i) = v_i \mid Sp = p)$$

We used a Naive Bayes classifier to predict the speaker from the discretized azimuth angles from all four persons. We did this with a subset of 70% over all three meetings. In one case, we balanced the training set, such that the a priori probability of each person being the speaker was 25%. In a second case, we used an unbalanced training set, with the a priori probabilities of the last column of Table 1. We tested our classifiers with all samples that did not appear in the training set, so the other 30% of our samples. In our first case, we obtained a performance of 39.0%, our second case yielded 53.7%. This difference is explained by the fact that in our second case, both training and test set have a similar a priori probability.

5.2 Speaker prediction by human judges

How good are human judges in performing the same task: to decide who is the speaker when presented a view of the VMR? Participants of the experiment were shown the meeting room with the participants, displaying only azimuth head angles. There was an option panel where they were able to choose among the four speakers, being either confident or very confident. To prevent biased unfounded choices, when judges had no idea they could indicate this by pressing a special button. Figure 1 shows the VMR setting.

Each experiment consisted of a session with four parts, each containing 20 samples. There were two types of sessions. Type 1 contained feedback only on the first part whereas in type two the feedback was omitted completely. For the first two parts of both session types two times 20 samples were randomly

chosen from meeting 3, the third part contained 20 randomly chosen samples from meeting 2 and the last part contained 20 randomly chosen samples from meeting 1.

The idea behind this was twofold. In the first place it enabled us to see if the feedback was helpful to the participants. Secondly, we were able to see whether feedback on samples from one meeting influenced the results on samples from other meetings. The feedback was given directly after the participants had judged the sample by showing a red arrow above the head of the speaker. Judges were students and employees of our department. Both two session types were completed 20 times, resulting in a total of 3200 answered samples. The results are shown in Table 2.

	Part 1	Part 2	Part 3	Part 4	Total
Session type 1	47.8%	49.3%	29.8%	24.8%	37.9%
Session type 2	39.3%	42.0%	33.3%	35.3%	37.4%

Table 2. Classification results for humans per session type

The table shows that the human performance is approximately 38%, which is lower than we expected. An interesting thing to note here is that there are significant ($p < 0.05$ using a paired T-test) differences between the two session types. In the first place the results on the first two session parts are better with feedback than without feedback and for the last two session parts we see a significant worse performance. Furthermore it appeared that when no feedback is given the performance remained much more stable over the different session parts.

In session type 2 the human judges were not informed of the prior speaker probabilities. In session type one, the participants 'learned' the a priori distribution for meeting 6. This can explain the differences between the two session types. This way, feedback helps the participants to make an a priori estimation on talkativity for each person. In Table 3 the confusion matrices for the human judges is shown.

Actual	Estimated speaker			
speaker	Sp' = 1	Sp' = 2	Sp' = 3	Sp' = 4
Sp = 1	31.8%	15.4%	24.0%	28.8%
Sp = 2	15.3%	42.8%	24.0%	17.9%
Sp = 3	14.3%	14.7%	51.1%	20.0%
Sp = 4	21.7%	12.7%	22.0%	43.7%

Table 3. Confusion matrix for actual speakers (Sp) and predicted speakers (Sp') for the experiment with humans. Performance is 42.3%

6 Conclusions and Further Research

Virtual meeting rooms may add value to the already existing technological means people have to communicate and meet. What requirements VMRs should obey depends on the type of activities that people do when meeting. A lot of research remains to be done to see how people perceive and interpret meeting situations and how they react on them in a virtual meeting room. Results of such research is necessary to see what information channels and modalities are important to effectively perform the various tasks in a meeting.
Several questions remain concerning research on human performance in recognizing the speaker and more general the flow of conversation. How do human judges perform when they look at the real pictures of the meeting instead of showing them the corresponding situations in the VMR? It appears that also from a picture of the real meeting it is sometimes quite hard to see who is the speaker. We will also perform similar experiments with the VMR where we also show arm and hand movements and body postures according to the recordings and recognitions in the real meetings. Do human judges perform significantly better when provided with this information than on the basis of head movements only? We expect they will because there are typical head and body gestures that distinguish speakers from listeners. Further, we will perform similar experiments to see how good judges are in deciding whos the *addressee* in a given situation showing head movements (with or without real postures and gestures). We will compare these results with machine learning techniques trained on annotated meetings. Then we will pursue our work on meeting modeling and see how we can present real meetings in an effective way by means of a virtual representation that shows the most informative view on the meeting.

Acknowledgements This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-42).

References

[Bailenson et al., 2002] Bailenson, J., Beall, A., and Blascovich, J. (2002). Gaze and task performance in shared virtual environments. *The journal of visualisation and computer animation*, 13:313–320.

[Bailenson et al., 2004] Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., and Turk, M. (2004). Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence*, Vol. 13, No. 4, August:428441.

[Bailenson et al., 2001] Bailenson, J. N., Blascovic, J., Beall, A., and Loomis, J. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence*, 10(6):583–598.

[Fussell et al., 2000] Fussell, S. R., Kraut, R. E., and Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work . In *Proceeding of the Conference on Computer Supported Cooperative Work, CSCW2000.*

[Jovanovic and op den Akker, 2004] Jovanovic, N. and op den Akker, R. (2004). Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 32:1–25.

[Kraut et al., 1996] Kraut, R. E., Miller, M. D., and Siegel, J. (1996). Collaboration in performance of physical tasks: Effects on outcomes and communication. In *Proceedings of the Computer Supported Cooperative Work Conference, CSCW' 96. NY: ACM Press.*, pages 57–66.

[Nijholt et al., 2004] Nijholt, A., op den Akker, R., and Heylen, D. (2004). Meetings and meeting modeling in smart surroundings. In *Social Intelligence Design. Anton Nijholt and Toyoaki Nishida (eds.), Proceedings third international workshop*, pages 145–158. CTIT Series WP04-02, Enschede.

[Padilha and Carletta, 2003] Padilha, E. and Carletta, J. (2003). Nonverbal behaviours improving a simulation of small group discussion. In *Proc. 1st Nordic Symp. on Multimodal Comm.*, pages 93–105.

[Pertaub et al., 2002] Pertaub, D.-P., Slater, M., and Barker, C. (2002). An experiment on public speaking in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments.*, Vol. 11, Issue 1, February:68–78.

[Regenbrecht and Schubert, 1997] Regenbrecht, H. and Schubert, T. (1997). Measuring presence in virtual environments. In *Human Computer Interaction International '97, San Francisco/CA, USA*.

[Reidsma et al., 2004] Reidsma, D., Rienks, R., and Jovanović, N. (2004). Meeting modelling. In *Proceedings of the Martigny Workshop AMI/M2/M4/Pascal*.

[Slater and Steed, 2001] Slater, M. and Steed, A. (2001). Meeting people virtually: Experiments in shared virtual environments.

[Stiefelhagen, 2002] Stiefelhagen, R. (2002). Tracking focus of attention in meetings. In *Proc. of the ICMI2002*.

[Stiefelhagen et al., 2001] Stiefelhagen, R., Yang, J., and A.Waibel. (2001). Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces (PUI01)*.

[Stiefelhagen and Zhu, 2002] Stiefelhagen, R. and Zhu, J. (2002). Head orientation and gaze direction in meetings. In *CHI '02 extended abstracts on Human factors in computing systems*, pages 858–859. ACM Press.

[Vertegaal, 1998] Vertegaal, R. (1998). *Look who's talking to whom. Mediating Joint Attention in Multiparty Communication and Collaboration.* PhD thesis, University of Twente.