Differences in Head Orientation Behavior for Speakers and Listeners: An Experiment in a Virtual Environment

RUTGER RIENKS, RONALD POPPE, and DIRK HEYLEN
University of Twente

An experiment was conducted to investigate whether human observers use knowledge of the differences in focus of attention in multiparty interaction to identify the speaker amongst the meeting participants. A virtual environment was used to have good stimulus control. Head orientations were displayed as the only cue for focus attention. The orientations were derived from a corpus of tracked head movements. We present some properties of the relation between head orientations and speaker–listener status, as found in the corpus. With respect to the experiment, it appears that people use knowledge of the patterns in focus of attention to distinguish the speaker from the listeners. However, the human speaker identification results were rather low. Head orientations (or focus of attention) alone do not provide a sufficient cue for reliable identification of the speaker in a multiparty setting.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Psychology; I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis

General Terms: Experimentation, Human factors, Measurement, Performance

Additional Key Words and Phrases: Head orientation, multiparty conversation, focus of attention, gaze behavior, virtual environments, perception of gaze

ACM Reference Format:

Rienks, R., Poppe, R., and Heylen, D. 2010. Differences in head orientation behavior for speakers and listeners: An experiment in a virtual environment. ACM Trans. Appl. Percept. 7, 1, Article 2 (January 2010), 13 pages.
DOI = 10.1145/1658349.1658351 http://doi.acm.org/10.1145/1658349.1658351

1. INTRODUCTION

Gaze, facial expressions, posture, head movements, gestures, and many other nonverbal cues contribute to the flow of conversation and play an important role in communicating affect and affiliation in social interaction [Argyle 1988]. All of these communicative behaviors occur in the visual channel. In technology-mediated forms of conversation, such as videoconferencing or interactions with avatars in virtual worlds, these visual signals may not always be available. It is also possible that they are not

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access).

Authors' address: Human Media Interaction Group, University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands, {rienks,poppe,heylen}@ewi.utwente.nl.

DOI 10.1145/1658349.1658351 http://doi.acm.org/10.1145/1658349.1658351

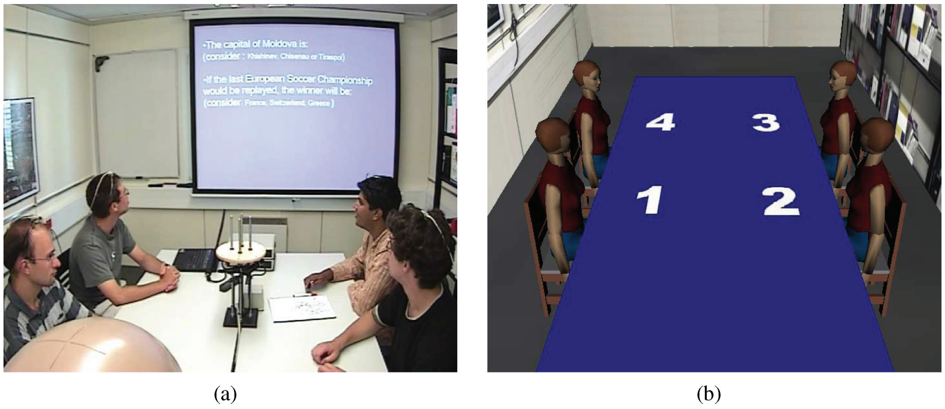
transmitted completely or represented faithfully, either due to limitations of the technology [Whittaker 2002] or because of conscious manipulations [Bailenson et al. 2004].

Several studies have looked at the question whether the quality of mediated interactions can be improved when such nonverbal cues are rendered. Studies such as those by Sellen [1992], Colburn et al. [2000], Vertegaal et al. [2000], and Garau et al. [2001] have investigated, for instance, whether the quality of videoconferencing can be increased when information about gaze and head orientation of the participants is displayed accurately. Most of them report an increased appreciation and effectiveness of the interaction. Also communication between humans and virtual agents can be improved when the head of the agent moves in a more natural way and when it shows appropriate gaze behaviors. The agents have been shown to become more believable, efficient, lifelike, and helpful [Cassell and Thórisson 1999; Poggi et al. 2000; Heylen et al. 2002].

A precise understanding of the relation between the rendition of communicative behaviors and the quality of communication is hampered by the fact that one communicative behavior may serve several different functions and one function can be marked by several alternative behaviors [Whittaker 2002]. It is, therefore, not obvious how to tease the influence of all the behaviors apart. Also, with the traditional tools that are used for research into conversational behavior such as video, it is difficult to focus on one single modality (e.g., speech, facial expression, gesture, or gaze), while ignoring all others. This limitation makes it hard to study in detail how humans judge the effect of a specific modality on the conversation. The use of a virtual environment (VE) may offer a solution as it allows good control over the stimuli [Loomis et al. 1999].

In this article, we report on a specific study that investigates the role of the focus of attention of participants in a multiparty setting. In particular, we look at how focus of attention as manifested through head orientation affects attributions of speaker and listener roles. The focus of attention of participants is an important parameter in conversations. It is mainly manifested through gaze and well approximated by head orientation [Stiefelhagen 2002; Beall et al. 2003]. The role of focus of attention (or gaze) has been well studied and analyzed and has led to the formulation of a number of systematic patterns related to conversational functions [Kendon 1967; Argyle and Cook 1976]. One of the parameters that codetermines where people are orienting their attention toward is participation status. This involves, among others, the question whether a person is speaking or listening, and in the latter case, if the person is being addressed or not. In typical conversations, a speaker displays a specific pattern alternating between looking toward addressees and looking away [Argyle and Dean 1965; Goodwin 1984]. Listeners, on the other hand, look at the speaker most of the time, unless they are distracted by some task [Fussell et al. 2005]. These differences in orientation in natural conversations, one might conjecture, may function as a clue for outside observers to infer who is speaking and who is listening. If this is the case, then rendering the focus of attention of participants in mediated communication may be an informative feature. Direct transfer of these cues is but one way to render focus of attention. Alternatively, the Transformed Social Interaction [Bailenson et al. 2004] paradigm alters this notion by transforming nonverbal cues in form, while retaining the meaning. Such an approach can be used to display cues that are well interpreted by observers.

For our study into the relation between focus of attention and judgements of participation status, we have collected a corpus of multiparty meetings in which the head orientations of the participants were tracked by electromagnetic sensors. This allowed us to study the differences in head orientations for the various roles with more precision. For the study on the judgements we report on in this article, we used a virtual environment (Figure 1(b)) in which the meeting room is replicated in 3D [Reidsma et al. 2007], and the participants are replaced by avatars, following a suggestion by Symons et al. [2004]. The data collected from the corpus was used to animate the head movements of the avatars. We conjecture that if the pattern of head orientations is systematically correlated enough with speaker and listener


-The capital of Moldova is:
(consider : Khishinev, Chisenau or Tiraspol)
-If the last European Soccer Championship
would be replayed, the winner will be:
(consider: France, Switzerland, Greece)
4
3
1
2

(a) (b)

Fig. 1. (a) Real meeting setting and (b) Virtual Meeting Room setting.

status, then observers should be able to correctly identify the speaker in a significant number of cases. Section 4 reports on this experiment.

This article is structured as follows. First, we provide a short overview of some of the literature on focus of attention, gaze, and head orientations. In Section 3, we analyze our recorded data for differences in gaze behavior between speakers and listeners. In the experiment, described in Section 4, we verify whether human observers use the existing differences to correctly identify the speaker among the meeting participants.

2. HEAD ORIENTATIONS IN CONVERSATIONS

Head movements in conversations are determined by a great variety of factors and play a role in several conversational systems (see Heylen [2006] for an overview). Head movements also serve as accompaniments of the rhythmic aspects of speech and typical head movement patterns can be observed marking uncertain statements and lexical repairs. Movements such as shakes and nods can signal negation and affirmation. Postural head shifts can mark switches between direct and indirect discourse [McClave 2000].

Head movements and orientations play a particulary important role in controlling and organizing the interaction. Listeners turn their heads toward speakers to enhance communicative attention. Typical patterns can be found related to turn taking. Certain movements may anticipate an attempt to capture the floor or signal the intention to continue. Duncan and Niederehe [1974], for instance, observed that shifting away the head direction acts as a signal indicating that a person shifts from the listener to the speaker state, and thereby prevents others from taking the turn.

Head orientations are involved in the same conversational processes as gaze. This is not surprising, as the direction of gaze is determined by a combination of head orientation and eye orientation [Kleinke 1986]. Head orientations can thus be used as a deictic signal, indicating the current focus of interest [Langton 2000]. Several studies have indicated that head orientations by themselves are a good indication of the focus of attention (see also Perrett and Emery [1994] and Otsuka et al. [2005]). In a four-person setting that is comparable to ours, participants rotated their head and their eyes in the same direction in 87% of all cases [Stiefelhagen 2002]. The focus of interest could be determined solely by the head orientation 88.7% of the time. Based on the fact that the head orientation component of gaze is so prevalent, we expect to see the same systematic patterns that occur in gaze behavior when looking at head orientations alone. We will validate this conjecture by analyzing a corpus consisting of head orientations and speaker data.

]

Table I. Number of Samples and Turns Per Meeting

	Meeting 1	Meeting 2	Meeting 3	Total
Samples	28,148	13,078	11,333	52,559
Turns	214	85	92	391

When focusing on the distinction between gaze behaviors of speakers and listeners, some notable differences have been found. For example, in dyadic conversations, the speaker spends less time looking at listeners than vice versa [Nielsen 1962; Exline 1963]. This observation is supported by Argyle and Cook [1976], who estimated that listeners in dyadic conversations look at the speaker for 75% of the time, whereas the speaker only looks at the listener for 41% of the time. For multiparty settings with four people, a similar trend was observed by Vertegaal et al. [2000], who also found that listeners gaze much more at the speaker (62.4%) than at a listener (8.5%).

In the next section, we analyze the data on head orientations in our corpus. We will verify the findings in the literature regarding differences in the time spent looking at speakers and listeners. Given the systematic patterns in head orientation and gaze and the differences between speakers and listeners, one might wonder to what extent these behaviors can also function as a signal for observers of conversations, indicating who is speaking and who is listening. We expect that people are aware of these systematic patterns. Consequently, when presented with the set of head orientations of all participants in a four-person setting, we expect observers to be able to identify the speaker in a significant number of cases. This conjecture will be validated with a user experiment using the VMR presented in Section 4.

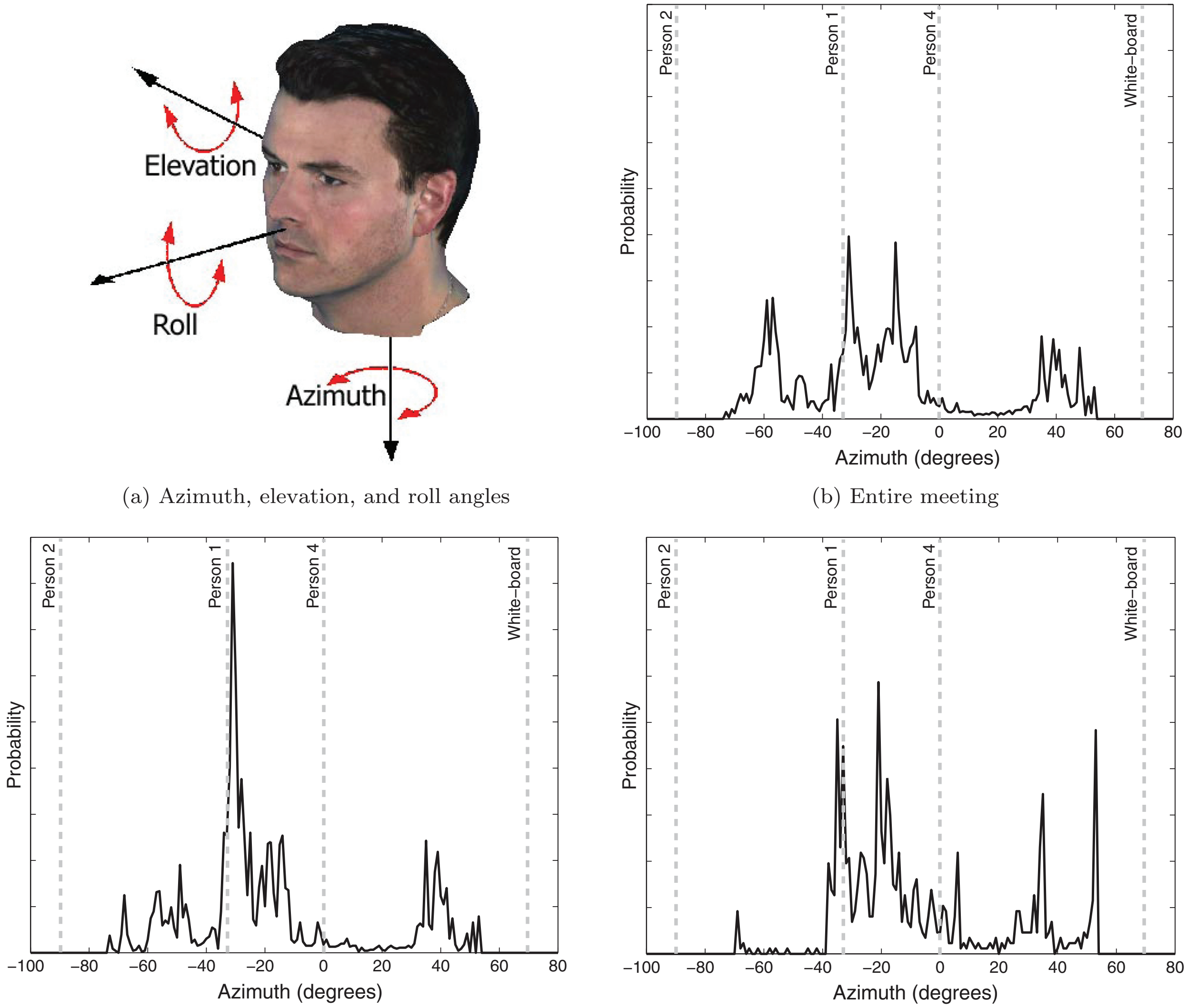
3. CORPUS RECORDING AND ANALYSIS

The corpus that we describe in this section was recorded for two purposes. First, we are using it to verify some findings about differences between speakers and listeners in multiparty conversations. Second, we are using the data to animate avatars. The recording of the corpus is discussed in Section 3.1. We analyze the corpus in Section 3.2.

3.1 Recording the Corpus Data

Three meetings with a total duration of 21 minutes were recorded in the Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) smart meeting room in Martigny, Switzerland (Figure 1(a)). Each meeting consisted of debates about three issues. These were projected onto the whiteboard. The four meeting participants were sitting two by two, at opposite sides of the table. Three cameras and an overhead microphone were used to record the audio and video. The head positions and orientations of all meeting participants were tracked using electromagnetic sensors at a rate of 50Hz. The Flock of Bird sensors we used, Ascension Technology 6DFOB, have an orientation accuracy of 0.5°. Each sensor is only a small box, and when mounted on top of a participant's head, it does not cause any distraction during the meeting.

We analyzed head orientation and video data to discover possible biases due to incorrect mounting of the Flock sensor on the head. We manually corrected the orientation data for these biases, which were all within the $\{-10^\circ, 10^\circ\}$ interval. For each participant, an azimuth orientation angle (Figure 2(a)) of 0° corresponds to looking straight forward and looking to the right corresponds to a positive rotation. The speech was transcribed manually from the audio recordings, allowing us to determine who was speaking at anytime. Head orientation data and speaker data were time aligned and all occurrences with nonspeech or with speech overlap were removed from the dataset. The dataset thus consists of samples taken at distinct time instances. Each sample contains the head orientations of the four meeting participants together with the speaker label. The number of samples and turns per meeting is summarized in Table I.



(a) Azimuth, elevation, and roll angles

(b) Entire meeting

(c) When Person 1 is speaking

(d) When Person 3 is speaking

Fig. 2. Distribution of azimuth angles for Person 3 in Meeting 3.

3.2 Analyzing the Corpus Data

As all the participants are located within the same elevation-roll plane, the azimuth angle is the most informative rotation to distinguish the different focus of attention targets. When one plots the azimuth angle distribution for each of the people over a whole meeting, one can see that the peaks in the orientations of the head correspond more or less with items of interest. An example of such a plot is displayed in Figure 2(b), where the distribution of azimuth angles of Person 3 is given. The locations of the others and the center of the whiteboard are indicated with dotted lines. The seating arrangement of the participants is presented in Figure 1(b). The graph shows that the four peak areas correspond more or less with the three other participants and the whiteboard.

Table II. Percentage of Time that a Speaker and Listeners are Being Looked at

Role	Speaker	Listener	Other	Total
Speaker	N.A.	88.32 (29.44)%	11.68%	100%
Listener	46.39%	42.64 (21.32)%	10.97%	100%

Numbers between brackets are averages for an individual listener.

The fact that the correspondence is not exact has several explanations. First of all, we solely use head orientation and ignore the head position. By leaning forward or backward, the relative positions between the people change. Also, part of the gaze direction is constituted by eye orientation. We expect this to be the main reason for the fact that the orientations toward the two people sitting at the opposite side of the table tends to be a bit in between the people. Also, our data does not only contain "fixations," where a person looks at another person or target, but also those time instants where a person is moving his head from one target to the next.

Figure 2(c) correlates the orientation of the head with information about the person who is speaking. It shows the distribution of azimuth orientation angles of Person 3, when Person 1 is speaking. The figure shows a clear indication for the expected correlations between head orientation of listeners toward the current speaker. The highest peak reveals that Person 3 is directing his head mostly toward Person 1, when Person 1 is speaking. We obtained similar graphs for all the other speaker–listener combinations. In these plots, the highest peak corresponds to the location of the speaker. Given this observation, and the literature discussed in the previous section, we expect that, in multiparty interaction, a speaker is being looked at by more people than any listener is.

Results of a quantitative analysis of where speakers and listeners rotate their heads toward are given in Table II. We defined a person as being looked at by another person if the head orientation of the latter was within a range of $\{-15^\circ, 15^\circ\}$ from the angle between them, as calculated from the mean position of the head during a meeting.

The table shows that listeners orient their heads toward the speaker 46.39% of the time and to any of the other listeners individually 21.32% of the time. We find that the average amount of gazes that a speaker receives is 1.39 (3×0.4639), whereas a listener only receives 0.72 gaze, on average (0.2944 from the speaker and 0.4264 from the other two listeners). From this, it follows that the amount of time that listeners look at the speaker is approximately 2 times higher than the time spent looking at each of the other two listeners. For each participant, we looked at the average number of gazes that was received at each time instant, either when speaking or when listening. The difference between the two cases proves significant ($t(11) = 5.6478$; $p < 0.0001$). It should be noted that our findings differ from those obtained in Vertegaal et al. [2000], where listeners were found to be gazing over 7 times more at the speaker than at any of the other listeners individually. Although the situation used was a comparable setting with four persons, the differences can be explained in part by the criterion that was used to determine who was looked at. Vertegaal et al. [2000] used an eyetracker and reported that all gazes within the face were counted as eye gazes. However, the measurement of such eye gazes was different and appears to be less strict. It should be granted that differences in the conversational setting, the task, or between individuals will have an influence as well. Vertegaal et al. [2000] used a round table, with symmetric relative seating positions. In this setting, each participant was seated directly opposite another participant. This is also the case in our setting, but the different relative positions of the other meeting participants resulted in an average head orientation that was in between the two participants at the opposite side of the table. This effect is visible in Figure 2(c). Vertegaal and colleagues did not find significant differences between the time spent looking at the participants at different relative locations, but this could be due to their measurement criterium. The area where gazes were counted as eye gaze was smaller for the participant sitting directly opposite. Also, there are differences in the room set-up.

Table III. Percentages of Time that a Certain Number of Meeting Participants had their Heads Oriented to either a Speaker or a Listener

	Looked at by number of people			
Role	3	2	1	0
Speaker	13.26%	32.16%	35.04%	19.53%
Listener	2.01%	12.68%	40.68%	44.62%

For example, the distance between participants was larger in our setup. Also, in our setting, there was a whiteboard and some other visible targets that might have invoked different looking behavior.

Another difference between our findings and those reported in Vertegaal et al. [2000] is the amount of time that speakers look at listeners and vice versa. We found that speakers look at an individual listener 29.44% of the time, whereas a listener looks at the speaker 46.39% of the time. Vertegaal et al. [2000] found somewhere between 17.2% and 19.7% (as calculated from the data by the authors) and 62.4% for these situations, respectively. It appears that these findings, when using gaze, are not entirely in line with ours. Again, the differences in methodology and setting could be explanations for this discrepancy.

Table III shows how many heads were oriented toward a listener or the speaker. In the case that three participants look at the same person, the probability that this person is the speaker is 68.75% $(13.26\%/(13.26\% + 3 \times 2.01\%))$. So even though speakers are looked at by three people relatively often, the probability that a person who is looked at by three people is a speaker is much lower.

This analysis allows us to conclude that there are differences between speakers and listeners with respect to the head orientations in the azimuth plane that indicate the focus of attention, which are similar to the findings in earlier studies that looked at gaze. In the next section, we examine whether human observers can identify the speaker. For this, we use a virtual environment.

4. EXPERIMENT: IDENTIFYING SPEAKER AMONG MEETING PARTICIPANTS

We investigate whether observers can infer who is speaking when presented with the head orientations of the participants in a conversation. Given the different patterns in head orientation behavior, we expect the observers to have some clue about who is the speaker when being shown the set of azimuth angles of participants' head orientation on avatars in a reasonable number of cases. Our first hypothesis is:

H1: Human observers are able to identify the speaker in a significant number of cases in a meeting with four people, when shown only their head orientations.

The use of a virtual environment allows us to present human observers with different types of stimuli. Specifically, we use either stills or animations. In the still condition, a static scene with head orientations is shown. The animation condition shows the head orientations over an entire speaker turn. The observer is presented with more context, for example, the length of the turn. Moreover, different gaze patterns can be related to the progress within a turn, whereas this information cannot be obtained in the still condition. We formulate our second hypothesis as:

H2: Human observers are better able to identify the speaker from animations of whole speaker turns, compared to the condition where only stills are shown.

Otsuka et al. [2005] identify three looking regimes for conversational settings: convergence (there is one person attracting the others' gaze more than any of the others), dyad-link (the situation where two people look at each other), and divergence (gaze patterns that do not match the other two). Given the distribution of gaze behavior of speakers and listeners reported in Section 2, we find that the convergence regimes hold for most of the cases in our data. Also, we expect the convergence regimes to be more informative, since there is clearly a difference in gaze distribution among the meeting participants. A convergence-n regime is the situation where the person who is looked at most, is looked at by n individuals. We expect the best speaker identification performance for the convergence-3 regime. Our third hypothesis is thus:

H3: The performance of the convergence-3 regime on speaker identification will be higher than for any of the other convergence regimes.

4.1 Method

Stimuli. We used the VMR with the setting, as described earlier, with four avatars (Figure 1(b)). This setting corresponds to the setting in the real recorded meetings, with the distances between all participants and the whiteboard properly scaled. The azimuth head angles of the avatars were the only parameters that were varied.

The use of a virtual environment for this kind of perception task is novel. Poppe et al. [2007] investigated the appropriateness of a virtual environment for head direction perception research. The virtual meeting room setting that is used in this work is similar to the one Poppe et al. [2007] used in their work. One of their findings was that accuracy of perception for head directions, as observed in a virtual environment, is sufficient for showing the orientation toward individual meeting participants. This finding is in line with Sagiv and Bentin [2001], who found that schematic faces are capable of producing similar effects to real faces, and Wilson et al. [2000] who found that perception of head orientation was high, even for low-resolution images.

Traditional research in perception of gaze and head orientation mainly focused on dyadic situations [Gibson and Pick 1963; Cline 1967; Kleinke 1986]. In these situations, a sender looks at receivers, or slightly next to them. The task of the receivers is to report either whether they are being looked at, or quantitatively determine where a sender is looking. A triadic setting differs in that an observer has to determine where a senders is looking, not relative to himself. This has been found to be a more difficult task, due to the more unfavorable position of the observer [Krüger and Hückstedt 1969]. In Poppe et al. [2007], the accuracy for perception of head orientations in triadic situations was assessed under a number of viewing conditions for the observer. It was found that observers could determine the focus of attention targets with approximately 5° accuracy. This is sufficiently accurate to distinguish between the different avatars in our experiment.

We used two stimuli types: stills and animations. In the still condition, we provided the observers with a static scene. In this scene, the heads of the avatars were oriented in the azimuth plane in precise correspondence with a scene of a real meeting. We randomly selected samples from a meeting. For each sample, there was exactly one speaker and observers had to identify the avatar who they thought was the speaker.

Animations of complete turns provide more context and display the dynamics of head orientations during a turn, with typical differences in speaker and listener behavior. In the animation condition, we displayed the head orientations of the meeting participants during an entire turn, which was derived from the speaker annotation of the data. The speaker turns, randomly chosen from a meeting, varied in length between half a second and 25 seconds. The animation was played at the same rate as the original data. Again, observers had to identify the speaker.

Procedure. Each observer completed 4 session parts of 20 samples each, all from the same stimulus condition. The samples of the first two parts were taken from Meeting 1, the third part from Meeting 2,

Table IV. Speaker Identification Performance in Still and Animation Condition

Condition	Part 1	Part 2	Part 3	Part 4	Average
Still	44.13%	44.69%	37.88%	38.32%	41.25%
Animation	45.52%	42.52%	35.62%	49.37%	43.27%

and the last part from Meeting 3. There was no time constraint imposed. In the animation condition, observers could replay the animation as a whole, as often as they wanted.

The observers were asked to press the button with the number that corresponded with the number on the table before the speaker (see Figure 1(b)). After pressing the button, the experiment advanced to the next sample. A forced-choice methodology was abandoned by introducing a “no idea” button to prevent participants from conveying indifference to the task [Ray 1990].

Participants. A total of 40 people (6 female and 34 male) took part in the experiment. These observers were students and employees of our department between 21 and 48 years of age. One half (20 people) of the observers was presented with the still condition, the other half with the animation condition.

4.2 Results and Discussion

A total number of 3,200 samples was collected, half of which used the still stimulus and half of which used the animation stimulus. The results are shown in Table IV. Samples where observers identified no speaker but instead used the “no idea” button have not been taken into account. This button was used for 148 (9.25%) and 40 samples (2.5%) in the still and animation condition, respectively.

First, we examined if any learning effects occurred. Because we expect these effects, if present, to be most salient in the first number of samples, we compared the performance of the first and second session part. Both parts contain samples from Meeting 1. A paired samples t-test showed no significant improvement of part 2 over part 1, in neither still nor animation condition. This leads us to believe that learning did not play a significant role. The mean duration of the experiment was approximately 9 and 21 minutes for the still and animation conditions, respectively. Given this moderate duration and the fact that participants could take breaks at any given time during the experiment, we do not expect fatigue had an effect on the participants’ performance.

The baseline for performance is 25%, the expected outcome when no a priori probabilities are known. Our findings, summarized in Table IV, are significantly higher, which supports Hypothesis H1. However, the overall percentage of correct guesses (slightly over 40%) is rather low, indicating that it is difficult for human observers to identify the speaker among the meeting participants.

To see what factors had a significant effect on the performance scores, we performed a repeated measures analysis of variance (ANOVA) with the stimulus condition (still or animation) as between-subject variable and the session part as within-subject factor. The dependent variable is the percentage of correct speaker identifications per person, per part (the mean over 20 samples). It was found that there are no significant differences between the still and animation conditions ($F(1, 38) = 0.36$, n.s.), which means that we have to reject Hypothesis H2. This is somewhat surprising, since the animations contain much more context information about the speaker turn, such as duration and begin and end of the turn. Also, when shown an entire meeting turn, human observers can relate certain gaze behavior to the progress of the turn. The lack of improvement over the still condition suggests that human observers are either unable to interpret this extra information, or find it too ambiguous to result in a better identification of the speaker.

We observe that a turn is made up of a series of consecutive frames. We analyzed the samples from the still condition and looked where they occurred within the turn. This allowed us to to see whether differences in speaker identification performance exist in different phases of the turn. In Figure 3, we divided all samples from turns with a length between 1 and 15 seconds (90.82% of all turns) into

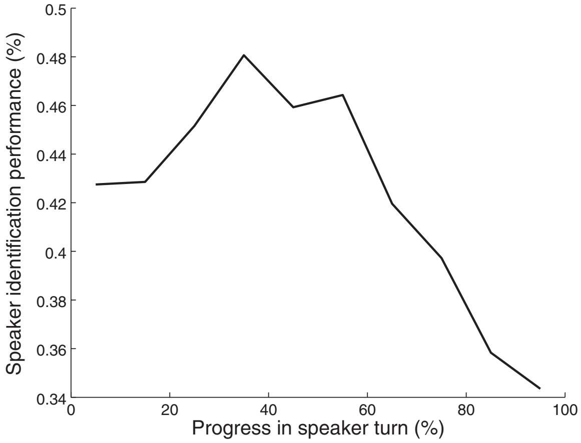


Fig. 3. Speaker identification performance as a function of progression within the speaker turn.

10 bins. Each bin covers a 10% range of progress within the turn, that is, the first bin contains samples that are within the first 10% of the turn length. The number of samples per bin is approximately 130. From the graph, it is clear that the speaker identification performance differs in the different phases within a turn. We observe that the best performance is halfway the turn. The start, and especially the end of the turn are scored worse. This could be explained by the fact that, when a new turn starts, meeting participants shift their gaze from the previous speaker to the current one. A similar pattern occurs at the end of the turn [Weisbrod 1965; Duncan and Niederehe 1974]. We informally verified that this is often the case in our data.

While these observations do not explain our low speaker identification performance, they do give insight in what factors play a role. Note that we did not include the samples from turns that were shorter than 1 second. We expected that these turns were too short to display different phases within the turn. Along the same line, we expect longer turns to display different phases more clearly. To see whether turn length has an effect on the speaker identification performance, we obtained Figure 4. All turns with a length shorter than 15 seconds (97.21% of all turns) were divided into 15 bins. The n^{th} bin contains samples from turns with a length in the range $[n-1, n\rangle$ seconds. We see large differences between bins, but due to the unequal spread of samples over the bins, these differences are not significant. To give an indication, the last five bins contain approximately 25 samples per class, whereas the first five bins contain on average 160 samples. When looking at the first five bins, we see an increasing trend, which suggests that performance is lower for shorter turns. A larger number of samples for the longer turns is required to determine if this trend is significant.

We looked at the variations of performance between the different parts of the session. A significant difference was found for the different parts (F(3, 114) = 3.296, $p < 0.05$). Part 3 scored lowest (36.80%), and Part 1 had the highest identification rates (45.04%). The differences in the scores between the various parts may be explained by uncontrolled variables such as topic of the meeting, meeting participants, and atmosphere. To study the effects of these factors on the speaker identification performance, a larger corpus should be used, with possible factors of influence controlled, or at least annotated. No significant interaction effect was found between condition and the session parts.

We tested our hypothesis that observers' performance would be best in the convergence regimes. Recall that these are the situations where there is one person attracting the others' gaze more than any of the others. We used all the samples from the still condition and calculated the scores for convergence regimes convergence-n. The scores for the animation condition were left out, since a single turn can

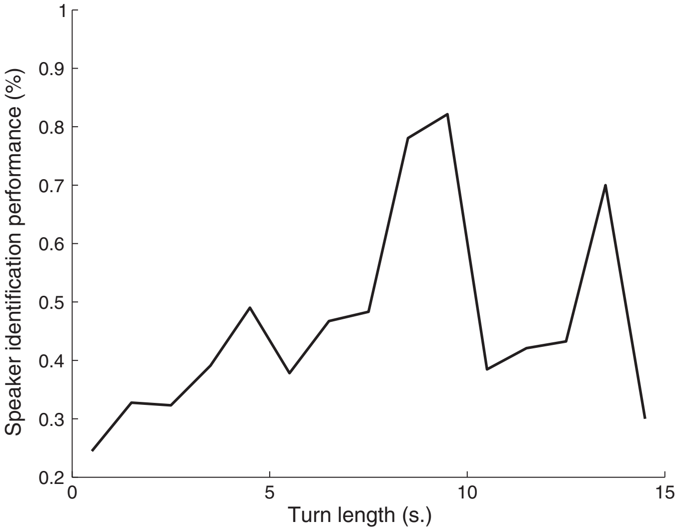


Fig. 4. Speaker identification performance as a function of speaker turn length.

Table V. Performance for Convergence-2 and Convergence-3 Regime

Condition	Occurrence	Performance Score	Most Looked at is Guessed	Most Looked at is Speaker
Convergence-2	865	36.88%	51.10%	42.89%
Convergence-3	317	55.84%	75.39%	70.03%

contain multiple regimes. We looked at the convergence-2 and convergence-3 regimes, convergence-1 occurred only four times and is not further analyzed. The results are shown in Table V.

The table shows that convergence-3 regimes scored significantly better (55.84%) than the convergence-2 regimes (36.88%) (t(547) $= -5.851$, $p < 0.01$). This confirms Hypothesis H3. To find out whether observers indeed used the differences in head orientation behavior as a means to predict the speaker, we calculated how often the person where most heads were oriented to was identified as the speaker. Table V shows that observers identified the person that was looked at most in 51.10% and 75.39% for the convergence-2 and convergence-3 regimes, respectively. This demonstrates that observers indeed seem to think, or at least applied the heuristic, that speakers are generally being looked at more than a listener. These results confirm our expectations that people use the systematic differences in head orientation behavior between speakers and listeners as clues to who has the turn. The relatively low performance scores for the two regimes can be explained by the observation that although the speaker is guessed often to be the one who is being looked at the most, this only is the case in 42.89% and 70.03% for the convergence-2 and convergence-3 regimes, respectively. These percentages are similar to the observations made in Section 3. The slight differences are due to the fact that these are samples taken from the entire corpus.

5. CONCLUSIONS AND FUTURE WORK

In this article, we have shown that the differences in focus of attention behavior between speakers and listeners in a multiparty setting, as previously described in the literature, are also reflected in the data if one takes head orientations as the only clue. By analyzing a corpus of four-person meetings, it appeared that speakers are generally being oriented to by more people than listeners are.

In an experiment, conducted using a virtual environment, we found that human observers can use these systematic differences when asked to identify the speaker if they are shown the head orientations of the meeting participants. This virtual environment proved to be a suitable tool for research in perception of human behavior, since it allows for good stimulus control.

The fact that people make use of the head orientations to infer aspects of the flow of conversation suggests that in forms of mediated communication, it might prove wise to pay attention to capturing and representing the head orientations of the participants correctly. Alternatively, one could manipulate the rendition of the head orientations and use a regime where all listeners look at the speaker to indicate who is speaking.

Although human observers appear to use knowledge about the differences between speakers and listeners regarding head orientation behavior, the speaker identification results are rather low (slightly more than 40% over a 25% baseline). Apparently, head orientations alone do not provide a sufficient cue for reliable identification of the speaker in a multiparty setting. Moreover, no significant differences were found between a still condition and a condition where an animation of an entire speaker turn was shown. We investigated different factors to get more insight in what affects the identification performance. It was found that progress within the turn and turn length were of influence, but more data is needed to reliably determine how these factors play a role.

Future work will aim at researching what meeting characteristics effect the identification performance. We plan to look at different participant characteristics, dominance relations between participants, and meeting topic. Also, we plan to conduct experiments with other modalities to determine which cues human observers use to identify the current and the next speaker. We will use these cues and the patterns in which they are exhibited by humans to animate avatars in a more natural way.

ACKNOWLEDGMENTS

The authors further wish to thank the reviewers for their valuable comments.

REFERENCES

ARGYLE, M. 1988. *Bodily Communication* 2nd Ed. Routledge, London.

ARGYLE, M. AND COOK, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, UK..

ARGYLE, M. AND DEAN, J. 1965. Eye-contact, distance and affiliation. *Sociometry 28*, 3, 289–304.

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 13*, 4, 428–441.

BEALL, A. C., BAILENSON, J. N., LOOMIS, J., BLASCOVICH, J., AND REX, C. S. 2003. Nonzerosum mutual gaze in collaborative virtual environments. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, Berlin, 1108–1112.

CASSELL, J. AND THRISSON, K. R. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Appl. Artif. Intell. 13*, 4, 519–538.

CLINE, M. G. 1967. The perception of where a person is looking. *Am. J. Psych. 80*, 1, 41–50.

COLBURN, R. A., COHEN, M. F., AND DRUCKER, S. M. 2000. The role of eye gaze in avatar mediated conversational interfaces. Tech. rep. MSR-TR-2000-81, Microsoft Research.

DUNCAN, S. AND NIEDEREHE, G. 1974. On signalling that it's your turn to speak. *J. Exp. Soc. Psych. 10*, 234–247.

EXLINE, R. V. 1963. Explorations in the process of person perception: Visual interaction in relation to competition, sex, and need for affiliation. *J. Personality 31*, 1–20.

FUSSELL, S. R., KRAUT, R. E., GERGLE, D., AND SETLOCK, L. D. 2005. *Other Minds.* Guilford Press, New York, 91–105.

GARAU, M., SLATER, M., BEE, S., AND SASSE, M. A. 2001. The impact of eye-gaze on communication using humanoid avatars. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'01).* ACM, New York, 309–316.

GIBSON, J. J. AND PICK, A. D. 1963. Perception of another person's looking behavior. *Am. J. Psych. 76*, 3, 386–394.

GOODWIN, C. 1984. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK, 225–246.

HEYLEN, D. 2006. Head gestures, gaze and the principles of conversational structure. *Int. J. Hum. Rob. 3*, 3, 241–267.

HEYLEN, D., VAN ES, I., VAN DIJK, B., AND NIJHOLT, A. 2002. Experimenting with the gaze of a conversational agent. In *Proceedings of the Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. ACM, New York, 93–99.

KENDON, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica 26*, 22–63.

KLEINKE, C. L. 1986. Gaze and eye contact: A research review. *Psych. Bull. 100*, 1, 78–100.

KRÜGER, K. AND HÜCKSTEDT, B. 1969. Die Beurteilung von Blickrichtungen. *Zeitschrift fur experimentelle und angewandte Psychologie 16*, 452–472.

LANGTON, S. R. 2000. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Q. J. Exp. Psych. 53A*, 3, 825–845.

LOOMIS, J. M., BLASCOVICH, J. J., AND BEALL, A. C. 1999. Immersive virtual environment technology as a basic research tool in psychology. *Behav. Res. Methods, Instrum. Comput. 31*, 557–564.

MCCLAVE, E. Z. 2000. Linguistic functions of head movements in the context of speech. *J. Pragmatics 32*, 855–878.

NIELSEN, G. S. 1962. *Studies in Self-Confrontation*. Munksgaard, Copenhagen, Denmark.

OTSUKA, K., TAKEMAE, Y., YAMATO, J., AND MURASE, H. 2005. A probabilistic inference of multiparty-conversation structure based on Markov switching models of gaze patterns, head directions, and utterances. In *Proceedings of the International Conference on Multimodal Interface (ICMI'05)*. ACM, New York, 191–198.

PERRETT, D. I. AND EMERY, N. J. 1994. Understanding the intention of others from visual signals. *Curr. Psych. Cognition 13*, 683–694.

POGGI, I., PELACHAUD, C., AND DE ROSIS, F. 2000. Eye communication in a conversational 3D synthetic agent. *Euro. J. Artif. Intell. 13*, 3, 169–181.

POPPE, R., RIENKS, R., AND HEYLEN, D. 2007. Accuracy of head direction perception in tryadic situations: Experiment in a virtual environment. *Perception 36*, 7, 971–979.

RAY, J. J. 1990. Acquiescence and problems with forced-choice scales. *J. Soc. Psych. 130*, 3, 397–399.

REIDSMA, D., OP DEN AKKER, R., RIENKS, R., POPPE, R., NIJHOLT, A., HEYLEN, D., AND ZWIERS, J. 2007. Virtual meeting rooms: From observation to simulation. *AI Soc. 22*, 2, 133–144.

SAGIV, N. AND BENTIN, S. 2001. Structural encoding of human and schematic faces: Holistic and part-based processes. *J. Cognitive Neurosci. 13*, 7, 937–951.

SELLEN, A. J. 1992. Speech patterns in video-mediated conversations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'92)*. ACM, New York, 49–59.

STIEFELHAGEN, R. 2002. Tracking focus of attention in meetings. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02)*. IEEE, Los Alamitos, CA, 273–280.

SYMONS, L. A., LEE, K., CEDRONE, C. C., AND NISHIMURA, M. 2004. What are you looking at? Acuity for triadic eye gaze. *J. Gen. Psych. 131*, 4, 451–469.

VERTEGAAL, R., SLAGTER, R., VAN DER VEER, G., AND NIJHOLT, A. 2000. Why conversational agents should catch the eye. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'00)*. ACM, New York, 257–258.

VERTEGAAL, R., VAN DER VEER, G. C., AND VONS, H. 2000. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface*. ACM, New York, 95–102.

WEISBROD, R. M. 1965. Looking behavior in a discussion group. Tech. rep., Cornell University, Ithaca, New York.

WHITTAKER, S. 2002. *The Handbook of Discourse Processes*. Lawrence Erlbaum Associates, Mahwah, NJ, 243–286.

WILSON, H. R., WILKINSON, F., LIN, L.-M., AND CASTILLO, M. 2000. Perception of head orientation. *Vis. Res. 40*, 5, 459–472.

Received November 2006; revised November 2008; accepted December 2008.