Differences in head orientation between speakers and listeners in multi-party conversations

Rutger Rienks*, Ronald Poppe, Dirk Heylen

Human Media Interaction Group
Department of Electrical Engineering, Mathematics and Computer Science
University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands

Abstract

We investigated whether people can identify who is speaking in multiparty conversations solely on the basis of the head orientations of the participants in the conversation. Two experiments were conducted using a virtual meeting room. Participants were shown the head orientations of four virtual humans and were asked to identify the speaker. In the first experiment participants were shown static frames, in the second an animation of an entire speaker turn. Head orientation data for both frame and animation experiment were taken from real-life meetings. Given that the animations provide more context, we assumed better scores for this experiment. However, participants did not score significantly better in the animation experiment (42.36%) compared to the static experiment (40.63%).

Key words: Head orientation, multi-party conversation, gaze behavior, virtual environments

1 Introduction

People move their heads a lot during conversations for all kinds of reasons. One of these has to do with gaze and focus of attention. People orient their heads so that they can see the person or object that is in their focus of attention. It has been shown that there is a high correlation between gaze and head orientation. In studies with four meeting participants at a round table, Stiefelhagen (2002)

* Corresponding author: Tel: +31 53 489 3740; Fax: +31 53 489 3503

Email addresses: rienks@ewi.utwente.nl (Rutger Rienks),
poppe@ewi.utwente.nl (Ronald Poppe), heylen@ewi.utwente.nl (Dirk Heylen).

Preprint submitted to Elsevier Science 4 March 2005
reported that in 89% of the time it could be correctly determined at whom the subject was looking based only on the participant’s head orientation. We might therefore explain a large part of the head movements by the same reasons invoked for explaining gaze behavior.

Kendon (1967) groups the determinants or functions of gaze behavior into five classes: providing visual feedback, regulating the flow of conversation, communicating emotions, communicating attitudes and interpersonal relationships, and improving concentration by restricting visual input. A similar list is also presented in (Argyle et al., 1973).

There are some typical differences between speakers and hearers in the gaze patterns they display. “Hearers give speakers fairly long looks broken by comparatively brief glances away, whereas speakers alternate looks toward their recipients with looks away from them of about equal length” (see (Goodwin, 1981)). Overall, listeners in dyadic conversations look more towards speakers, i.e. 75% of the time, than vice versa which is 41% of the time, as Argyle and Cook (1976) observed. In accordance with this Vertegaal (1998) found that, in a setting with three persons, people gaze much more at the speaker (62.4%) than at others (8.5%). Kendon (1967) also noted a specific distribution of gaze over the course of an utterance. At the beginning of the utterance a speaker looks away but gazes steadily toward his addressee as the utterance approaches termination. A hearer at this point looks away from the speaker. Could these differences in gaze behavior tell us at any point in the conversation who is speaking and who is listening? Given that there are these differences and the correlation between gaze and head orientation, this might imply that head orientations of persons in a meeting are possible indicators for speaker identification.

Our interest in these questions derives from our research in forms of mediated communication in videoconferencing systems and in communication with embodied conversational agents. Evaluation of the effects of gaze on the quality of interactions in mediated conversation have been carried out by Vertegaal (1999), Garau et al. (2001), Colburn et al. (2000), amongst others. In the Gaze Groupware system (Vertegaal, 1999), for instance, gaze information was used to improve the quality of video conferencing. Using the gaze information obtained from an eye-tracker, a 2D image of each participant is rotated to indicate whom the person is looking at. Improving the gaze behavior of virtual humans has noticeable effects on the way communication takes place. In a number of experiments it was shown how gaze behavior leads to increased appreciation of virtual humans since they become more believable or efficient (Heylen et al., 2004; Poggi et al., 2001). Gaze, facial expressions, postures, head movements, gestures and many more cues contribute to the flow of conversation and the perception of persons in social interaction (Argyle et al., 1973). When transferring such results to the domain of mediated communi-
cation one might be able to use the same cues. Given the effects this might have on many different factors, from believability and likability to efficiency, it is important to spend some effort on determining how people perceive these cues.

To find out whether head orientations of participants in a conversation provide information on who is speaking and who is listening, we carried out two experiments both using a virtual meeting room in which a group of 3D humanoids were shown sitting around a table. Only the head orientation of the humanoids was manipulated. Data was taken from recordings of real meetings. Before we discuss the experiments, we present more details on the collection of this data.

2 Meeting data collection

Data of head orientations during multiparty conversations was obtained by recording three meetings with a total duration of 21 minutes in the IDIAP (Institut Dalle Molle d’Intelligence Artificielle Perceptive) smart meeting room in Martigny, Switzerland (Figure 3(a)). Each meeting contained a lively discussion about statements which were presented on a white-board. In each of the meetings four meeting participants were sitting two-by-two, at opposite sides of the table.

![Fig. 1. Azimuth, elevation and roll angles for heads](image)

The audio was recorded using an overhead microphone. Three cameras were used to record the visual setting of the meeting. Apart from the video and audio recordings, head position and orientation of all four meeting participants were tracked by using electromagnetic sensors. We used Flock of Bird sensors (Ascension Technology, model 6DFOB) which allowed us to measure position and orientation accurately at a rate of 50 Hz. The sensor is a small box and mounting it on top of a participant’s head does not cause any distraction. The sensors measure head orientations in three directions (azimuth, elevation and roll, see Figure 1) with an accuracy of 0.5°.
We analyzed both video and head orientation data to discover possible biases due to incorrect mounting of the Flock sensor on the head. We corrected the orientation data for these biases. The azimuth orientation angles for participant 2 and 3 were rotated 180° to ensure that looking straight forward corresponds to a rotation of 0° and looking to the right corresponds to a positive rotation for all participants. The speech was transcribed manually from the audio recordings. We annotated the data collection further by indicating who was speaking at what time. Head orientation data and speaker data were time aligned and all occurrences with non-speech (laughter, silence, etcetera) or with speech overlap were removed from the data set. In this paper we will only be concerned with azimuth directions as these correspond to orientations towards other participants or objects of interest.

The seating arrangement of the participants is shown in Figure 3. An example of a distribution of the azimuth angles for one person (Person 3) in a meeting is displayed in Figure 2(b). Note that Person 4 is sitting opposite to Person 3 (0°), Person 2 is sitting to Person 3’s left (-90°), and Person 1 is sitting at approximately -45°. The white-board is to his right (15°). The graph shows a number of peaks around -60° (Person 2), -30° (Person 1), -10° (Person 4) and there are a further series of peaks between 30° and 50° (white-board). The graph thus clearly shows four peak areas which correspond with the three other participants and the white-board. There are various obvious explanations why the correspondence is not exact. First of all, the real seating arrangement was not as neat as the virtual arrangement in the picture. Also there is clearly a degree of sloppiness in the head orientations which can be compensated by eye-movements. The orientations towards the two people opposite (1 and 4) tend a bit to be in between the persons. The same tendency to minimize movements is clear from the orientation towards Person 2 sitting next to Person 3. Despite this sloppiness, it is obvious from Figure 2(c) that interesting correlations between head orientation and attending to speakers can be made. This figure shows the distribution of azimuth orientation angles of Person 3, when Person 1 is speaking (in meeting 3). It clearly shows that Person 3 is orienting his head towards the speaker. This is precisely what we expect given the previous observations on percentages of gaze. We obtained similar graphs, with different locations of the peaks, for all participants in all meetings. The highest peak each time corresponds to the location of the speaker. This confirms our assumption that azimuth orientation is a good predictor to determine the speaker. Particularly when we combine the azimuth orientation of all participants, in that case we expect three peaks towards the speaker in a significant number of cases.

We used the data of the real meetings to create frames and animations in a virtual meeting room to find out how informative the use of azimuth orientations of the head were with respect to identifying the speaker in multi-party conversations. The use of a virtual meeting room allowed us to control all
the factors, abstracting from all other possible cues. For such reasons, virtual environments have become an important tool in behavioral and perceptual psychology (Loomis et al., 1999).

3 Method

To find out how well humans can identify the speaker among all meeting participants on the basis of head orientations, two experiments were carried out. In both, participants in the experiment were asked to identify the present speaker when shown the head orientations of a fragment of the meeting.

3.1 Stimuli

Figure 3(b-d) shows different views on the Virtual Meeting Room (VMR) that was used in the experiment. The VMR is a replica of the IDIAP smart meeting room (Figure 3(a)). The distances between all participants and the white-board have been properly scaled. The virtual humans in the VMR all
look identical and show a similar pose to make sure posture and physical appearance were of no influence. The VMR enables participants in the experiment to look at the scene from different viewpoints.

![Image](image1.png) ![Image](image2.png)

(a) Real setting of meeting 3  (b) VMR camera view

(c) VMR view from above  (d) VMR view from person 1

**Fig. 3.** Real meeting setting and different views of the VMR

We performed two experiments that differed in the kind of stimuli that were presented. In Experiment 1, single frames (*stills*) were presented. Each stimulus was a view on the VMR that displayed the figures with the azimuth head orientations manipulated in accordance with data from the actual meetings. In the second experiment, a stimulus consisted of an animation showing the head movements along the azimuth axis of all meeting participants during an entire speaker *turn*. Again, the data was randomly selected from the actual, recorded meetings. An experiment session consisted of four parts. The samples of the first two parts of both experiments were taken from meeting 1, the third part from meeting 2 and the last part from meeting 3. The speaker turns displayed in the second experiment varied in length between half a second and 25 seconds.

Given that the 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, we expected the animations to yield better results.
3.2 Participants

In the first experiment 40 persons (6 women, 34 men) took part, and another 40 persons (5 women, 35 men) in the second. The participants were employees and students of our department. The mean age was 32 (range 23-48). In both experiments, a participant was asked to judge 80 stimuli, resulting in a total of 3200 samples per experiment.

3.3 Procedure

For both experiments, sessions consisted of four parts, each containing 20 samples. Participants were asked to complete all 80 samples. There was no time constraint imposed. It was possible to interrupt a session and continue at a later time. The samples were presented via a web interface. The interface presented an option panel where the participants of the experiment could indicate who they thought was the speaker in the frame or animation by clicking on the appropriate button. A forced-choice methodology was abandoned by introducing a ‘no idea’ button. This prevented participants from conveying indifference to the task (Ray, 1990). Participants could not go back to redo their choice. In Experiment 2, participants could replay the entire speaker turn before making a choice. When a choice had been made, the experiment continued with a new sample.

In both experiments half of the participants received feedback on the first session part. The feedback was given by showing a red arrow above the head of the correct speaker directly after the participants had made their vote. We expected humans to score better when given feedback, as they are able to reflect on their choices when presented with the right answer.

4 Results

This section presents the results of both experiments. As a baseline we used 25%, the expected outcome when no a priori probabilities are known. In Experiment 1 participants had to identify the speaker amongst the meeting participants from a static scene (frame). The results of Experiment 1 are shown in Table 1.

In Experiment 2 participants had to identify the speaker amongst the meeting participants from a dynamic scene (animation) where complete speaker turns of head orientations were shown. The results of this experiment are shown in
<table>
<thead>
<tr>
<th>Type</th>
<th>Part 1</th>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>With feedback</td>
<td>49.85%</td>
<td>51.43%</td>
<td>33.59%</td>
<td>26.59%</td>
<td>40.37%</td>
</tr>
<tr>
<td>Without feedback</td>
<td>42.52%</td>
<td>44.93%</td>
<td>36.81%</td>
<td>39.34%</td>
<td>40.90%</td>
</tr>
<tr>
<td>Average</td>
<td>46.19%</td>
<td>48.18%</td>
<td>35.20%</td>
<td>32.96%</td>
<td>40.63%</td>
</tr>
</tbody>
</table>

Table 1
Experiment 1 performance with and without feedback

<table>
<thead>
<tr>
<th>Type</th>
<th>Part 1</th>
<th>Part 2</th>
<th>Part 3</th>
<th>Part 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>With feedback</td>
<td>41.24%</td>
<td>40.89%</td>
<td>38.63%</td>
<td>45.10%</td>
<td>41.47%</td>
</tr>
<tr>
<td>Without feedback</td>
<td>45.47%</td>
<td>42.42%</td>
<td>35.55%</td>
<td>49.53%</td>
<td>43.24%</td>
</tr>
<tr>
<td>Average</td>
<td>43.36%</td>
<td>41.65%</td>
<td>37.09%</td>
<td>47.32%</td>
<td>42.36%</td>
</tr>
</tbody>
</table>

Table 2
Experiment 2 performance with and without feedback

Samples where participants identified no speaker but instead used the ‘no idea’ button have not been taken into account. The ‘no idea’ button was used for 263 and 106 samples, in Experiment 1 and 2 respectively.

The performance for both experiments is significantly higher than the baseline of 25% (t(39) = 15.72; p < 0.0005 for Experiment 1 and t(39) = 16.02; p < 0.0005 for Experiment 2). This means that people are able to interpret head orientations as cues as to whom is speaking in multi-party conversations.

If we look closer at the results of Experiment 1, particularly Table 1, we see that the human performance on the prediction of the speaker is 40.63%. There are two significant differences between the two session types. In the first place the performance on the first two session parts is better with feedback than without feedback (t(38) = 2.29; p < 0.05). Secondly, for the last two session parts we see a significantly worse performance for the sessions with feedback compared to the sessions without feedback (t(38) = −2.95; p < 0.01). When no feedback was given the performance remained much more stable over the different session parts. Note that feedback is given in the first session part only and that both the first and the second session are taken from Meeting 1. Session 3 and 4 are based on data from Meetings 2 and 3 respectively. When all session parts are taken together there is no significant difference between the two session types.

Table 2 contains the performance results of Experiment 2. From this table two things become apparent. First, contrary to the case in Experiment 1, providing the participants feedback does not have any influence on the performance on judging speakers when shown animations of head orientations
during complete speaker turns. There is no significant difference in the first two session parts and the last two session parts between the sessions with and without feedback. The second observation is that when the overall performance is considered, there is no significant difference ($t(78) = 1.17; \text{n.s.}$) between the total performance on the complete turn experiment and the single frame experiment. This suggests, somewhat surprisingly, that the extra information present in the speaker turn experiment does not improve identification results.

5 Discussion

The two experiments showed that people are able to distinguish between speakers and non-speakers in conversations when presented with head orientations only. There are noticeable cues in head orientations of speakers and listeners that enable humans to make a better than random prediction about the current speaker. However, in the majority of cases they fail to identify the correct speaker. One can imagine many configurations during a conversation in which the head orientations do not provide enough cues, for instance, when everyone is looking to a distractor such as the white-board.

We examined the difference between judging stills of head orientations and turn-length animations. However, no significant difference was found between the two experiments. This suggests that the dynamic aspects of head orientations do not provide extra cues. To investigate this further, we considered for each turn the judgment on the frame samples from that turn and took the majority vote on the judgments as the judgment on the turn. The average score in this case was 42.76% which is higher than the average frame and average turn score. This suggests that when people see more than one frame from a turn they might perform slightly better than on a single frame and also better than on the complete sequence. Further qualitative analysis is needed to find out why the complete turn experiment performs worse.

We also looked at the role of giving feedback. We expected that when feedback was given, humans create meeting specific rules or models to predict the speaker. When feedback is given participants perform better, but this holds only for feedback on frame items from the same meeting. There might be several explanations. It could be that humans learn the a priori speaker probabilities because similar configurations occur more often in one session. If a certain meeting participant happens to be the speaker more often than the others during training, it is likely that this person is also identified more often as the speaker in situations where the participant is uncertain. The differences in a priori speaker probabilities are different for each meeting as can be seen in Table 3.
### Table 3
Number of samples and a priori speaker probabilities in each meeting and all meetings.

<table>
<thead>
<tr>
<th>Samples</th>
<th>Meeting 1</th>
<th>Meeting 2</th>
<th>Meeting 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori speaker 1</td>
<td>24.8%</td>
<td>40.4%</td>
<td>26.9%</td>
<td>28.7%</td>
</tr>
<tr>
<td>A priori speaker 2</td>
<td>9.8%</td>
<td>27.3%</td>
<td>23.4%</td>
<td>16.9%</td>
</tr>
<tr>
<td>A priori speaker 3</td>
<td>29.4%</td>
<td>7.7%</td>
<td>8.6%</td>
<td>19.5%</td>
</tr>
<tr>
<td>A priori speaker 4</td>
<td>36.0%</td>
<td>24.7%</td>
<td>41.2%</td>
<td>34.9%</td>
</tr>
</tbody>
</table>

Apart from these possible a priori probability models, the meetings could also exhibit different gaze behavior. There is a noticeable difference in identification performance for different meetings. Difference in gaze/head orientation between meetings can exist because of differences in gaze behavior of individual participants, the contribution of context (the role of a person in the meeting, differences in interpersonal relationships, time, meeting topic, and to what extent an object such as the white-board becomes the center of the focus of attention). This needs to be further investigated.

### 6 Conclusion and Future Research

It appears from these experiments that people can predict who is speaking in multi-party conversations when presented only with information about the head orientations of the meeting participants. It remains to be investigated what other cues might improve these scores. Would information about exact gaze be more informative? To gain more insight into human gaze behavior, more research is needed about how and to what extent modalities such as gestures, facial expressions and poses provide cues about who is speaking and who is listening. Also, the differences in meeting settings such as the presence of a white-board, the number of meeting participants, and the seating arrangement are to be examined. The results of this research can be applied for extraction of information from multiparty conversations and for regeneration of gaze behavior for all modalities in the context of virtual meeting rooms, where virtual humans are to behave naturally (Reidsma et al., In press).
7 Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-47). We would like to thank Job Zwiers, Anton Nijholt, Matthijs Poppe and Lynn Packwood for their comments and assistance.

References