Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays

Samer Al Moubayed, Gabriel Skantze

Department of Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm, Sweden
sameram@kth.se, gabriel@speech.kth.se

Abstract

In a previous experiment we found that the perception of gaze from an animated agent on a two-dimensional display suffers from the Mona Lisa effect, which means that exclusive mutual gaze cannot be established if there is more than one observer. By using a three-dimensional projection surface, this effect can be eliminated. In this study, we investigate whether this difference also holds for the turn-taking behaviour of subjects interacting with the animated agent in a multi-party dialogue. We present a Wizard-of-Oz experiment where five subjects talk to an animated agent in a route direction dialogue. The results show that the subjects to some extent can infer the intended target of the agent's questions, in spite of the Mona Lisa effect, but that the accuracy of gaze when it comes to selecting an addressee is still significantly lower in the 2D condition, as compared to the 3D condition. The response time is also significantly longer in the 2D condition, indicating that the inference of intended gaze may require additional cognitive efforts.

Index Terms: Turn-taking, Multi-party Dialogue, Gaze, Facial Interaction, Mona Lisa Effect, Facial Projection, Wizard of Oz

1. Introduction

The function of gaze for interaction purposes has been investigated in several studies [1, 2]. Gaze direction and dynamics have been found to serve several different functions, including turn-taking control, deictic reference, and attitudes [3]. In a multi-party or situated dialogue, gaze may be an essential means to address a person in a crowd, or pointing to a specific object out of many. These functions have been investigated in experiments and models have been proposed on how to control gaze movements in, for example, robots and embodied conversational agents [4]. However, very little research has been done to investigate the perception of these movements by observers, especially in situated, multi-party settings.

It is known that perception of three-dimensional objects that are displayed on two-dimensional surfaces is guided by, what is commonly referred to as, the Mona Lisa effect [5]. This means that the orientation of the three-dimensional object in relation to the observer will be perceived as constant, no matter where the observer is standing in the room. If the portrait of a face is gazing forward, a mutual gaze will be established between the portrait and the observer, and this mutual gaze will hold no matter where the observer is standing. Accordingly, if the the portrayed face is gazing to the right, everyone in the room will perceive the face as looking to their left. Thus, either all observers will establish mutual gaze with the portrait, or none of them, and no exclusive eye-contact between the portrait and one of the observer is possible. This principle, of course, extends to all objects viewed on 2D surfaces, such as the direction of hands pointing.

This effect has important implications for the design of interactive systems, such as embodied conversation agents, that are able to engage in situated interaction, as in pointing to objects in the environment of the interaction partner, or looking at one exclusive observer in a crowd. The purpose of this study is to investigate how gaze may be used to control turn-taking in a multi-party human-computer dialogue, depending on the use of 2D or 3D displays.

2. Background

In a previous study in [6] and [7] we have measured the agreement of gaze direction between observers using an animated agent. In a perception experiment, five subjects were simultaneously seated around an animated agent, which shifted the gaze in different directions (see Figure 1). After each shift, each subject reported who the animated agent was looking at. Two different versions of the same head were used, one projected on a 2D surface, and one projected on a 3D static head-model (see Figure 2). The results showed a very clear Mona Lisa effect in the 2D setting, where all subjects perceived a mutual gaze with the head at the same time for frontal and near frontal gaze angles. While the head was not looking frontal, none of the subjects perceived any mutual gaze. In the 3D setting, the Mona Lisa effect was eliminated and the agent was able to establish mutual and exclusive gaze with any of the subjects.

While that study provides important insights and proves the principal directional properties of gaze through a 2D display surface, it does not show whether this effect will hold during interaction, or whether people are able to cognitively compensate for the effect, and correctly infer the *intended* direction of gaze.

In [8] and [9], a virtual receptionist is presented, which is able to communicate with multiple interlocutors, and to address them individually using gaze. The system uses a flat screen for the animated head, which would theoretically give rise to the Mona Lisa effect. Still, experiments on the use of gaze for controlling turn-taking in a multi-party conversational setting shows that the system may successfully address the different users to some extent. An accuracy of 86.2% between intended addresse and the next person to speak is reported [8]. However, it is not clear to what extent the results are due to the fact that there were only three users, which would make it possible for them to learn and infer the intended direction of gaze, with the cost of an extra cognitive effort. Also, it is not clear whether only gaze was shifted,or whether also head pose was used (as indicated in supplementing video material), which may ease the task for the subjects.

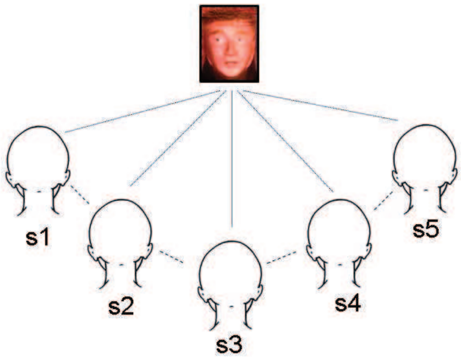

Figure 1: *Schematic setup and placement of the subject and stimuli point.*

In this study, we explore the interactional effects of gaze in a multi-party conversational setting, and investigate the difference between 2D and 3D manifestations of the animated agent. Unlike the previous perception experiment [7], which focused on the *perceived* gaze, this experiment will investigate how gaze may affect the turn-taking *behaviour* of the subjects. Thus, it is possible for the subjects to infer the intended gaze (which may however require extra cognitive resources). Another difference is that the subjects in the perception experiment did not share their vote on where the projected head was looking. In this experiment, the subjects' decisions may affect each other and confusion about the gaze target may have effects on the fluency of the interaction.

3. Method

We used a experimental setup similar to the previous experiment reported in [7]. Five subjects were seated at fixed positions at an equal distance from each other and from the stimuli point, as shown in Figure 1. An animated agent were presented which addressed the subjects by directing its gaze in their direction. Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface (using the same 3D computer model), as shown in Figure 2. The conversational behaviour of the animated agent was controlled using a Wizard-of-Oz setup, as explained further down. For each new question posed by the agent, the gaze was randomly shifted to a new subject.

The subjects were given the task of watching a first-person video from a camera navigating around the city of Stockholm, after which the animated agent asked them to describe the route they had just seen. After each video was finished, the animated agent started to ask the subjects about directions on how to reach the landmark the video ended with, starting from the point of view the video started with. The dialogues were in Swedish.

3.1. Subjects

Two sets of subjects were asked to take part in the experiment. Each set consisted of five subjects. The ten subjects were not informed with the purpose of the experiment beforehand. All subjects were fluent speakers of Swedish and had normal, or corrected to normal vision. During each session, the five subjects were randomly seated on the five seats. To record as many agent-user turn-taking shifts as possible, the subjects were asked not to consult each other about the answer, and the Wizard interrupted the users if their answers were too long. Furthermore, they were asked to try to provide some kind of response if they felt addressed, even if they did not know the correct answer (for example by saying ”I don't remember”).

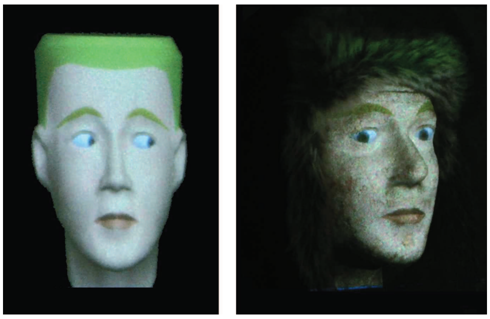

Figure 2: *A snapshot of the animated agent displayed on a 2D white board (left), and on a 3D head model (right).*

3.2. Conditions

Two conditions were used in the experiment for comparing the effects of gaze on the subjects. In the 3D condition, a micro laser projector (SHOWXX Pico Microvision projector) was used to project an animated head on a 3D head model (Figure 2, right). This surface was found previously to eliminate the Mona Lisa effect and to deliver accurate, and absolute perception of gaze direction [7]. In the 2D condition, a flat board was inserted in front of the 3D head model, in order to keep the two settings as similar as possible (Figure 2, left). This condition was assumed to represent a general 2D surface display (such as flat screens, walls, etc.).

3.3. Gaze Control

The five gaze targets were manually calibrated using the 3D surface projection so that each gaze angle would result in a mutual gaze with one of the five subjects. The same calibration was used in the 2D setting. The gaze shift was temporally aligned with the first stressed syllable of the last word of the question, and the exact timing of the gaze shift was logged. The randomization of the gaze shifts was controlled so that no subsequent gaze shift would be identical, that is to avoid conditional dependency on the subject who answered the previous question (if no change in the direction of the gaze is made, it is possible that the same subject will take the turn again even if the subject was not the intended one). Note that the agent's head pose was fixed for both conditions during the whole experiment.

3.4. Videos

A set of eight videos were used. The videos were recorded using a handycam for the purpose of route direction dialogue studies. In each video, the holder of the camera starts walking in the streets of Stockholm until reaching a specific landmark. The average length of each of the videos was two minutes. The eight videos were played for each set of subjects, and the 2D/3D condition was changed after each video. The first session started with the 2D condition, and the second session started with the 3D condition. Thus, each video was used in both conditions. The videos were projected on the wall in front of the subjects, behind the animated agent. An HD video camera was placed

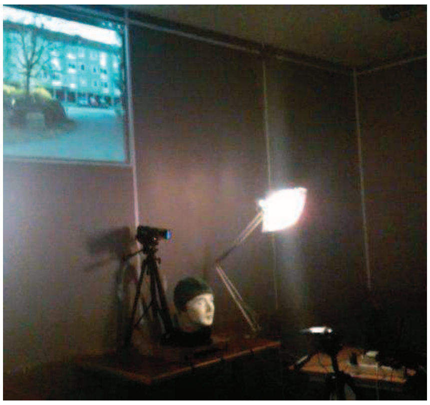

Figure 3: *A snapshot of the experimental setup showing a video playing on the wall, and an animated agent being projected on the static head model (3D condition).*

above the display surface to record the subjects. Figure 3 shows the setup of the experiment.

3.5. Wizard-of-Oz setup

After a video had ended, the dialogue with the animated agent was started by the Wizard. The Wizard was seated outside the room and guided the dialogue with the help of a microphone that was placed in the room. The subjects were not informed beforehand that the dialogue was guided by a Wizard. The Wizard's task was to select the type of response provided by the subjects, which automatically triggered a new question from the agent. The dialogue flow was defined in a state chart. The system's speech was synthesized using a diphone synthesizer synchronized with the lip movements of the animated agent. An example interaction is shown in Table 1.

Table 1: *The beginning of one of the recorded interactions translated from Swedish.*

Agent	Hi, can you help me to find my way to the Finnish church?
Subject 1	Yes
Subject 2	Yes
Agent	Should I continue in this direction?
Subject 3	Yes, a little bit at least
Agent	How far should I go then?
Subject 4	I think it was about 20 meters, maybe a bit longer
Agent	Is there some landmark I should look for?
Subject 2	You should look for the Great Church, when you come to Västerlånggatan you will see the Great Church on your left, then should pass it
Agent	Do you agree?
Subject 5	Yes

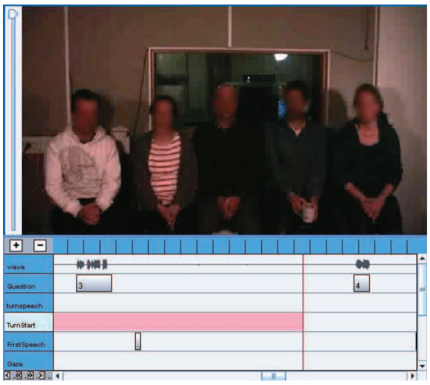

Figure 4: *A snapshot of the video that was captured from the agents perspective, along with the ANVIL annotation tool used for annotating the videos.*

4. Analysis

During the experiment, all events and their timings were logged into data files to automatically annotate the video files. These events include the change of session, video and condition, the questions asked by the agent, and the shift of the agent's gaze.

The recorded videos were annotated using the ANVIL video annotation tool [10]. For each answer from a subject, the timing, duration and the subject id of that answer were manually labeled. Figure 4 shows a snapshot of the video annotation process. Turn takes were labeled to start with the first vocal action made by any of the subjects. This provides a consistent way of labeling turn takes. Using the logged data and the manual labels, the response time was calculated as the time between the gaze shift, and the beginning of the succeeding utterance.

The first video of each of the sessions was considered as training and was therefore removed from the analysis, resulting in a total of 14 videos for both sessions. The full annotation resulted in 57 turn shifts (question-answer pairs) for the 2D condition, and 56 turn shifts for the 3D condition.

5. Results

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. Table 2 shows the confusion matrix between the intended gaze target and the speaker for the 3D condition. Rows in the table represent the intended target of gaze and columns represent the speaker who took the turn. Table 3 represents the confusion matrix in the same format, but for the 2D condition. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. In this measure of accuracy, the result is counted as a hit if the intended subject took the turn, and a miss if the wrong subject took the turn.

Another possible measure of accuracy is to count the number of seats between the intended subject and the actual turn-taker. The error is calculated by dividing this number by the maximum possible distance. Using this measure, the accuracy for the 2D condition is 80% and for the 3D condition 96%.

By looking at the confusion matrix of the 2D condition, it is interesting to see that the Mona Lisa effect is present to some extent. This is represented by the fact that all subjects at some point took the turn when the head was looking frontal (gazing at subject 3, who is seated in the middle). This effect is not present in the 3D condition. It is also interesting to see that when the agent was targeting subject 2 and subject 4, the subjects who were seated further away (subject 1 and subject 5 respectively) tended to take the turn. This can be explained by the fact that when the gaze is directed to the side on a 2D surface, the direction is perceived by all subject to be directed to their side but not at them. However, subject 1 and subject 5 have no more subjects seated further away, which will make them the most likely subjects to actually take the turn.

Table 2: *A confusion matrix of the 3D condition between the gaze intended target (t) and the actual speaker (s) that took the turn for that gaze target.*

Speaker/Target	s1	s2	s3	s4	s5
t1	11	3	0	0	0
t2	2	11	0	0	0
t3	0	0	8	0	0
t4	0	0	0	8	4
t5	0	0	0	0	9

Table 3: *A confusion matrix of the 2D condition between the gaze intended target (t) and the actual speaker (s) that took the turn for that gaze target.*

Speaker/Target	s1	s2	s3	s4	s5
t1	11	4	0	0	0
t2	2	6	1	1	1
t3	1	1	4	3	1
t4	0	0	0	0	12
t5	0	0	0	0	9

Figure 5 shows a plot of the response time for each condition. Each curve represents the sorted values for that condition. This plot illustrates the distribution of the values of the interval between the gaze shift of the question and the time it takes for one of the subjects to take the turn.

A two sample ANOVA analysis was applied, with the response time as a dependent variable, and the condition as an independent variable. The results show a significant main effect [$F(1)$= 15.821, $p<0.001$], with a mean response-time of 1.85 seconds for the 2D condition, and of 1.38 seconds for the 3D condition. No significant difference was found in response-time for any of the following factors: subjects, videos, and gaze targets. Neither is there any significant correlation with time (Pearson Correlation = -0.094), which means that there is no learning effect on how to perceive the gaze of the agent for either condition.

6. Discussion & Conclusion

The results show that the use of gaze for turn-taking control on 2D displays is limited due to the Mona Lisa effect. The accuracy of 50% is probably too low in settings where many users are involved. By using a 3D projection, this problem can be avoided to a large extent.

However, the accuracy for the 2D condition was higher than what was reported in a previous perception experiment in a similar setting [7]. A likely explanation for this is that the subjects in this task may to some extent compensate for the Mona Lisa effect – even if they don't "feel" like the agent is looking at them, they may learn to associate the agent's gaze with the intended target subject. This comes at a cost, however, which is indicated by the longer mean response time. The longer response time might be due to the greater cognitive effort required to make this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

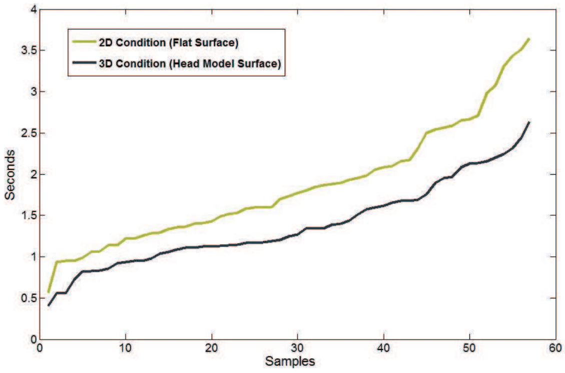

Figure 5: *A plot of the sorted response time (time interval between gaze shift and turn take) for the 2D and 3D condition*

7. Acknowledgements

This work has been carried out at the Centre for Speech Technology at KTH, and is supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314, as well as the SAVIR project (Situated Audio-Visual Interaction with Robots) funded by the Swedish Government (strategic research areas). We would like to thank Martin Johansson for producing the videos and the subjects for participating in the experiment.

8. References

[1] M. Argyle, R. Ingham, F. Alkema, and M. McCallin, "The different functions of gaze," *Semiotica*, vol. 7, no. 1, pp. 19–32, 1973.

[2] P. Mirenda, A. Donnellan, and D. Yoder, "Gaze behavior: A new look at an old problem," *Journal of Autism and Developmental Disorders*, vol. 13, no. 4, pp. 397–409, 1983.

[3] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychol.*, vol. 26, pp. 22–63, 1967.

[4] O. Torres, J. Cassell, and S. Prevost, "Modeling gaze behavior as a function of discourse structure," in *First International Workshop on Human-Computer Conversation*. Citeseer, 1997.

[5] D. Todorovic, "Geometrical basis of perception of gaze direction," *Vision research*, vol. 46, no. 21, pp. 3549–3562, 2006.

[6] J. Beskow and S. Al Moubayed, "Perception of gaze direction in 2D and 3D facial projections," in *Proceedings of the ACM/SSPNET 2nd International Symposium on Facial Analysis and Animation*. ACM, 2010, pp. 24–24.

[7] S. Al Moubayed, J. Edlund, and J. Beskow, "Taming Mona Lisa - communicating gaze faithfully in 2D and 3D facial projections (in press)," in *ACM Transactions on Interactive Intelligent Systems*. ACM, 2011.

[8] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 2010, p. 5.

[9] ——, "Computational models for multiparty turn-taking," MSR-TR-2010-115, Microsoft Research, Tech. Rep., 2010.

[10] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in *Seventh European Conference on Speech Communication and Technology*, 2001.