Implementation issues regarding eye gaze

in synthetic faces.

Gunilla Svanfeldt and Preben Wik
Department of Speech, Music and Hearing
KTH, Lindstedtsv.24, 100 44 Stockholm, Sweden
{gunillas, preben}@speech.kth.se

Abstract

The aim of this study is to investigate people's sensitivity to directional eye gaze, with the long-term goal of improving the naturalness of animated agents. Previous research within psychology have shown the importance of eye gaze in social interactions, and should therefore be vital to implement in virtual agents . In order to test whether we have the appropriate parameters needed to correctly control gaze in the talking head, and to evaluate users' sensitivity to these parameters, a perception experiment was performed. The results show that it is possible to achieve a state where the subjects perceive that the agent looks them in the eyes, although it did not always occur when we had expected.

1 Introduction

Today the default metaphor used in human-computer interaction (HCI) is the desktop metaphor, where the computer is likened with a desk containing a desktop, drawers and file holders. Once animated agents perform satisfactorily well, an important shift of metaphor used in HCI may occur, using instead a person metaphor. Before this can take place a number of unresolved issues must first be addressed.

Well-synchronised lip-articulation in the animated agent is a necessity for a natural communication, and has been provided through the work of Jonas Beskow (Beskow, 2003). The lip-reading support that the articulating synthetic faces can provide in noisy environments or to hearing impaired is well established (Beskow *et al.*,1997, Agelfors *et al*., 1998). The next step is to equip the agent with the capacity of being more expressive by means of adjusted articulation and other facial movements, so that turn-taking signals and attitudes can be conveyed. What must not be forgotten in this context is the importance of the eye gaze. So far the eyes in the synthetic faces have been more or less neglected, partly because of lack of control facilities. However, with the new MPEG-4 model that will be described below, the possibility of tailoring the eye gaze behaviour provides new opportunities to use this channel for information.

The aim of this paper is to investigate how to introduce more natural-like eye behaviour in animated agents. In order to test whether we have the appropriate parameters to elaborate with when adjusting the eye gaze in the synthetic face, and to evaluate the users' sensibility to these parameters, an experiment has been performed.

2 Eye gaze

Eye gaze has been studied quite extensively in human-human interaction within psychology. It has also gained attention in the area of animated characters. However, despite the discussion about the function and importance of gaze, it is rarely properly implemented. For example, in (Cassell *et al.*, 1994) a rather detailed description of four categories of gaze is given, but they do not differentiate between head and eye movements in the implementation of the system.

One of the goals in the development of a virtual agent is to make it natural to interact with, besides other important issues such as good articulation and relevant acoustic output. In order to achieve effective interaction, it is necessary for the user to feel that the agent is interested in and focussed on the conversational exchange, and that it displays relevant signals for intentions, understanding and turn-taking. A prominent visual cue for this, except head or eyebrow movements, is the eye gaze.

According to Argyle & Cook (1976), speech related gazes have three main functions:

1. to send social signals
2. to open a channel to receive information
3. to control the synchronizing of speech

As to the first function, certain rules about the amount of gaze apply to different situations. If these rules are broken, people are likely to be confused, or even offended. The amount of gaze transmits impressions of the temporary or permanent state of the user. Kleck and Nuessle (1968) found that persons (on film) with only 15% gaze were perceived as cold, pessimistic, and defensive, while those with 80% gaze were considered friendly, self-confident and sincere. Argyle (1988) reports about how people with higher levels of gaze are seen as more attentive, and that lack of eye contact indicates passiveness or inattentiveness. It is however not only a matter of total amount of gaze. There are norms concerning how long a glance should last, and the amount of mutual gaze also depends on the distance between the two persons, since the gaze can be seen as a regulator of intimacy. Another determining factor is the sex of the two conversational parts – women tend to have less eye contact with men than with other women.

If the agent is not capable of acting according to the social rules, it will not be considered trustworthy in the users' eyes or might induce unwanted reactions of the user. In a study by Park Lee *et al.* (2002), a gaze tracking device is used to acquire data. After data processing and implementation in a synthetic face, a comparison of three different types of eye movements is performed. It showed that with no eye movements at all, the character was perceived as lifeless, but with statistically based eye movements the face character looked more natural and friendly. With random eye movements, the quality of the character was unstable.

In a study performed by Garau *et al*. (2001), similar results were found when comparing dyadic conversations. There were four different conditions: video, audio-only, and two avatar conditions, where the avatar's head and eye movements were either randomly induced or based on research on face-to-face dyadic conversations. The video condition got the highest overall scores, and the inferred-gaze condition got better scores than the random case in similarity to real face-to-face conversation, involvement, co-presence and partner evaluation.

The second function – to open a channel to receive information – is not implemented in the talking head used for the study in this report. The agent cannot receive any visual information, although it hopefully will in the future. It may be discussed whether it still should simulate this ability or not in order to keep a fluent interaction with the user. However, if the agent signals that it can read the users' gestures, and then do not react to these, it might confuse the user. It is well known that a good interface should be clear with regards to what it can and what it cannot perform.

The possibility of controlling the synchronizing of speech, which is the third function of speech related gaze, is very appealing to explore with a talking head. Even though the signalling will only be in one direction, it may still lead to important improvements of the interaction with agent. If the animated agent is able to signal turn-taking, there is likely to be less uncertainty and interrupts in the conversation. It will also mean less cognitive load for the user if the basic signals for conversation regulation are employed.

3 The talking head

Animated talking heads capable of producing lip-synchronised speech have been developed at CTT (Beskow, 2003). The acoustic speech can be either synthetic or natural, and the model can also convey extra-linguistic signs such as frowning, nodding, and eyebrow movements.

To gain knowledge about how to animate the agents in terms of verbal and non-verbal behaviour, 3D facial data collected by means of an optical motion tracking system (*MacReflex)* from Qualisys[1] was used. Reflective markers attached to the speaker's face were registered with infrared cameras and the system provided the 3D coordinates of those markers.

A new generation of talking heads are currently being developed at CTT using the MPEG-4 standard. MPEG-4 is known for being a high compression standard for coding audio and video, but the MPEG-4 (Version 2) standard also describe channels for face and body animation in a very low bitrate coding. The standard defines 66 low-level facial animation parameters (FAPs) that describe the animation of a face model (Ostermann , 2002).

Distances in the MPEG-4 models are, as in many 3D models, not described in metrics, but in units. This is because the size of the model does not correspond to anything absolute in the physical world. There are however relative distances between different parts of the model that need to be described. A generic face has been modelled and measured and a set of standard sizes has been given. Deviations from this can then also be described with ease. For example, in the MPEG-4 face used in this experiment, the distance between the mouth and the tip of the nose is 5.3 units, the distance between the eyes is 12 units, and the diameter of the iris is 2.1 units.

In order to regulate the gaze of the synthetic face, the distance to the virtual viewpoint (or camera analogy) is used as basis for the eye positioning. The idea is to have the agent look into the camera, just as a person on a photograph seems to look at the observer, if the person was looking into the camera when the picture was taken. Having the virtual viewpoint as baseline, the point of focus can then be varied in a three dimensional space.

In the manual texture mapping of the synthetic face, textures risk being asymmetrically set, since there is no automatic procedure. This means that even though the gaze is calculated depending on the camera, some skewness in the model may damage the effect of eye gaze. The parameters that can be manipulated for eye gaze control are: rotation of the eyeballs in two directions, and the combined set-up of the eyes' direction will yield the focus point in space.

4 The experiment

4.1 Aim of the experiment

The primary aim of the experiment was to evaluate if subjects – for any of the conditions – perceive that the agent looked them in the eyes. If so, to what parameters were the subjects sensitive: how much displacement of the eyes was needed for loosing the gaze, and was the sensitivity different depending on dimension or position of the head? This knowledge is crucial for future studies and implementation of eye movements.

[1] http://www.qualisys.se

4.2 Method

There were 15 participants in the experiment, 8 women and 7 men. The stimuli that were presented to the subjects consisted of static pictures with different eye gaze and headposition. The eye gaze was varied in three dimensions – laterally (x), vertically (y) and in the depth (z) dimension. As described earlier, the face model is defined in units rather than ordinary metrics. Due to the lack of previous studies of this kind, the steps and limits were chosen to cover a reasonable range of angles. In the x- and y-dimensions, the total variation for the focus point was 40 units (which corresponds to an angle of approximately 20°), and each step was 5 units. In the z-dimension, the steps were not linear. For focus points between the agent and viewpoint, the steps were 30 units. For distances beyond the observer increasingly larger steps were taken. The variation in the three dimensions is illustrated in figure 1.

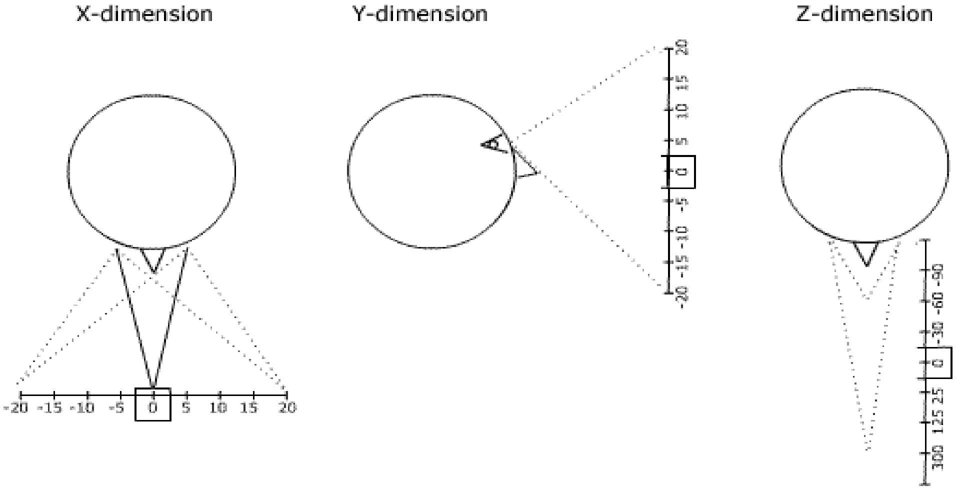
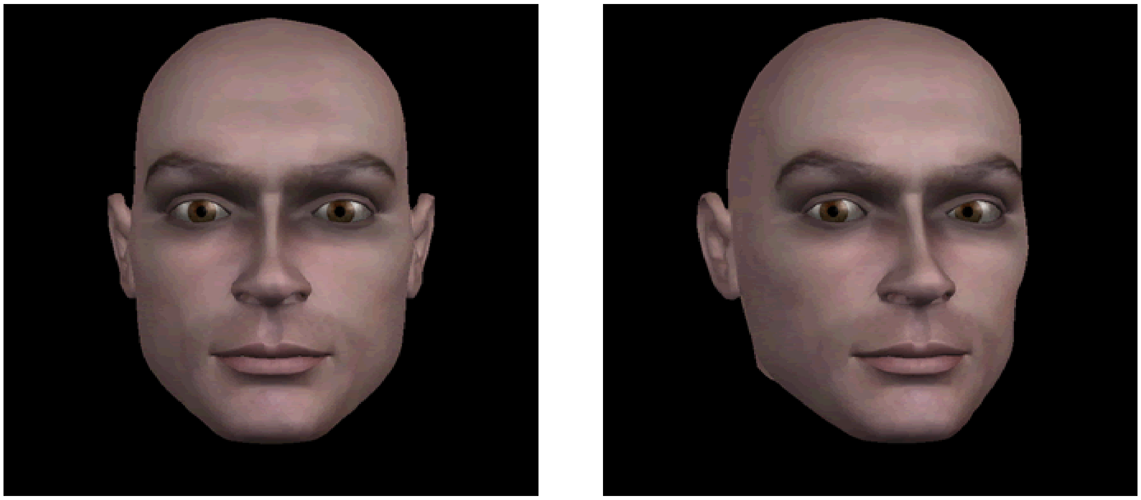


Figure 1. Illustration of the three dimensions in which the gaze point varies.

Two different head positions were evaluated, one front view, and one with the head turned to the side (see figure 2). The idea was to find a head position that typically could occur in a dialogue situation, yet large enough to make a perceivable difference. We found an angle of approximately 11° to be suitable. The two different head positions, together with 23 eye gaze variations gave 46 different conditions. Each condition was presented to the subjects twice, in randomised order. Six additional pictures – the same for all the subjects – were inserted in the beginning of the test as dummies and were removed before the analysis. In total 98 stimuli were presented to each subject. Some examples of eye gaze are shown in Figure 3.

The introduction to the test was presented by another talking head with synthetic acoustic speech, where the aim of the experiment was explained, and instructions to the test were given. For each stimulus, the subject was asked to answer yes or no to the question “Does this man look you in the eyes?”.



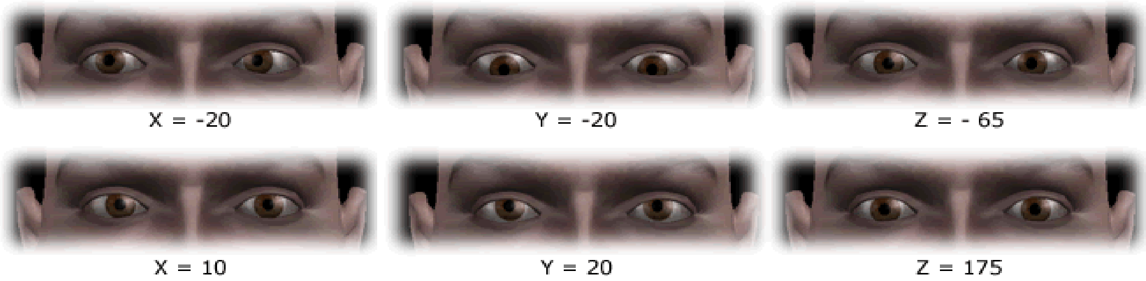


Figure 2. The two head positions used in the experiment. Both have eye gaze that according to calculations should look straight into the viewpoint, and thus look the observer in the eyes.

After the self-instructed sequence of 98 stimuli, four additional pictures were shown where the subject was asked to more qualitatively describe where the agent was focusing its gaze. The subject was also asked about any difference in difficulty of determining the gaze direction of the agent when the face was in the frontal view as compared to the side view.

Figure 3. Examples of gaze direction variation in each dimension.

Finally, the subject was invited to give his or her opinion to what the most prominent defect of the synthetic face was. The aim of this last question was to give us an idea of what possible distractions there might have been during the test, besides getting a hint of what is most urgent to improve.

4.3 Results

The results show that it is possible to produce eye gaze in the synthetic face that observers think meets their gaze. However, this did not always occur when we had expected. Some subjects were also more acceptant when judging the gaze than others, which can be seen in figure 4, where the total amount of positive answers is displayed.

The front view obtained more positive responses than the side view, as seen in figure 5. It received almost twice as many “yes”-answers as the side view, and this trend remained for all three dimensions in space, see figure 6,7 and 8.

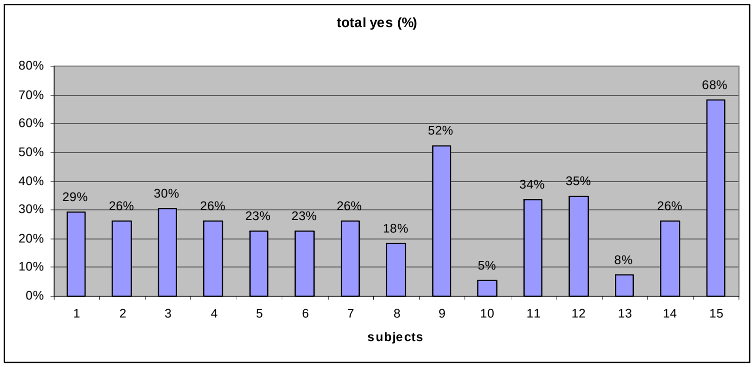


Figure 4. The graph shows the total percent of positive answers for each subject in the experiment. The question they responded to was "Does this man look you in the eyes?".

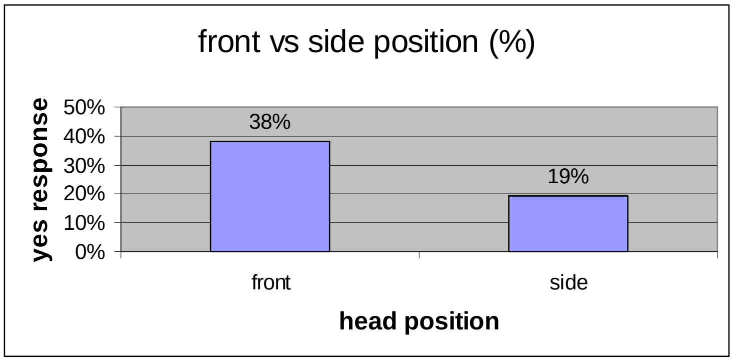


Figure 5. The diagram shows, for all subjects, how many pictures (in percent) that were perceived as looking the subjects in the eyes, separating the front view and the side view.

Both the x- and y-dimensions show that there is an asymmetry in the responses. The positive responses are not centred on 0 (corresponding to the virtual viewpoint), which would have been expected. This trend is more striking for the side view than for the front view. In the x-dimension, the 10 unit displacement got the highest score, and in the y-direction it was the 15 unit case that obtained the most positive answers.

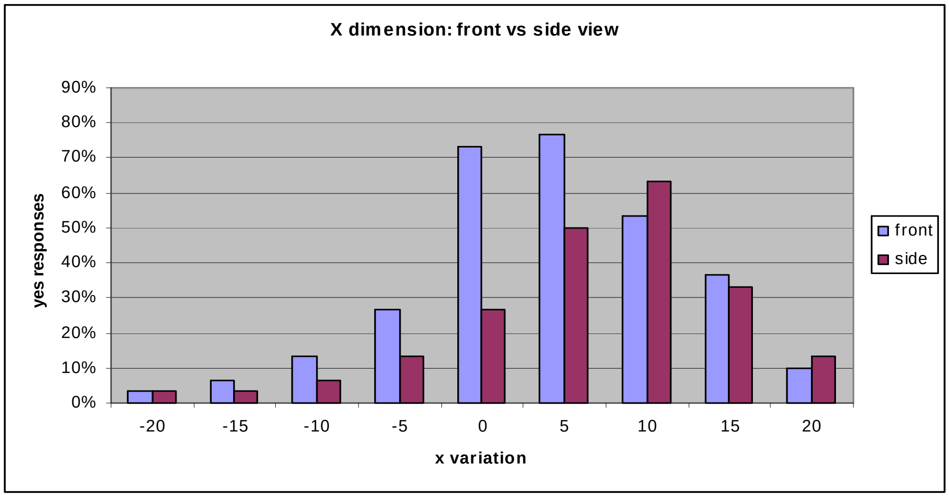


Figure 6 . Illustration of the distribution of responses confirming that the agent looked the subjects in the eyes. The steps on the x-axis are in units, the total range from -20 to 20 corresponds to an angle of 20°.

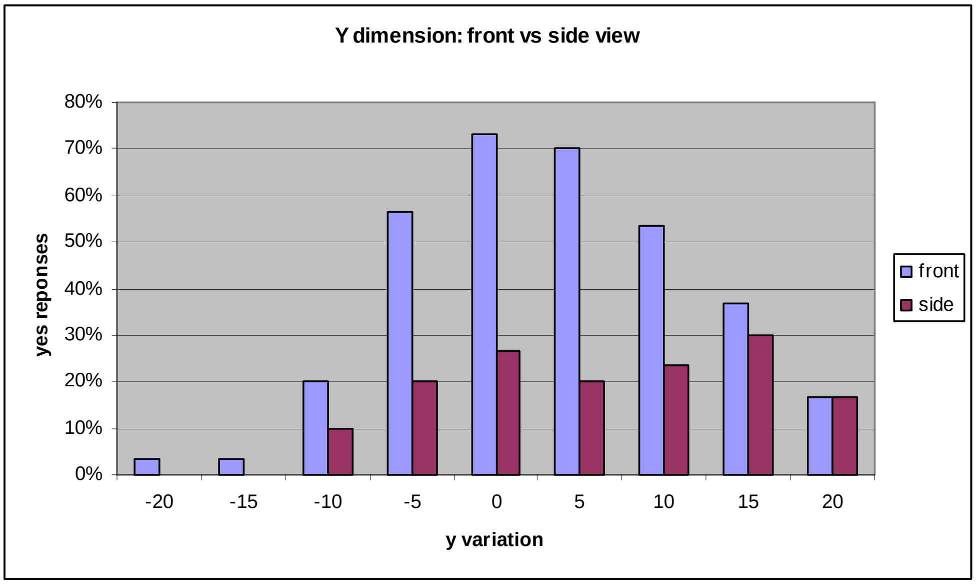


Figure 7 . Illustration of the distribution of responses confirming that the agent looked the subjects in the eyes. The steps on the x-axis are in units, the total range from -20 to 20 corresponds to an angle of 20°.

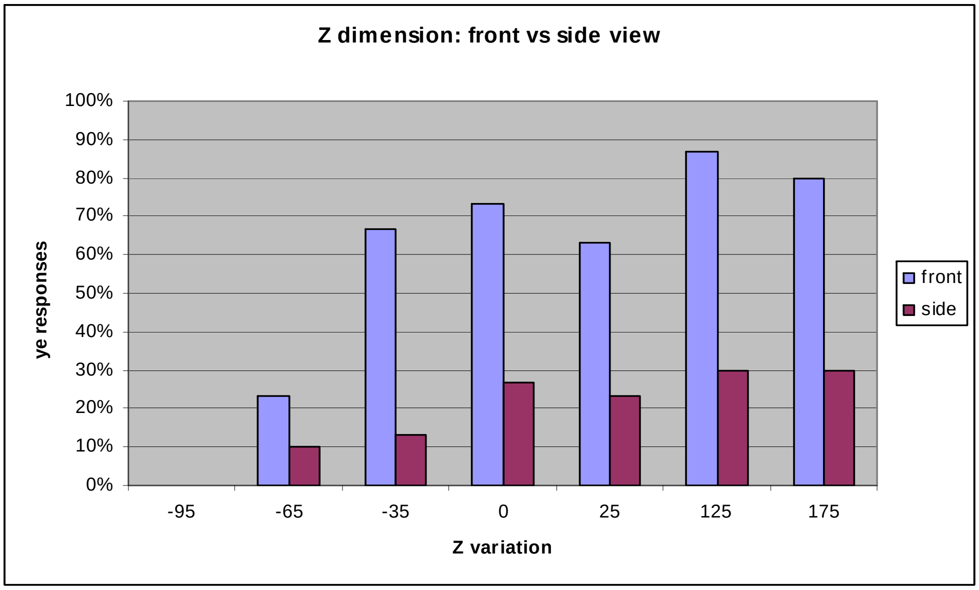


Figure 8 . Illustration of the distribution of responses confirming that the agent looked the subjects in the eyes. The steps on the x-axis are in units.

By the shape of the graphs in figure 6, 7 and 8, it can also be concluded that the subjects were less sensitive to changes in the depth dimension than in the other two dimensions. The sensitivity also diminished when the head was turned to the side, especially in the y-dimension.

The result of the four pictures that were shown after the self-instructing test was interesting in that a rather wide range of answers were given to where the agent focussed its gaze. The first picture had the focus point at 65 units in front of the virtual viewpoint, and thereby meant to be perceived as in between the screen and the subject. Out of the subjects, there were 7 who reported that this was the case, 3 subjects remarked that the agent looked to the left of the subject (which was not intended), another 3 thought the focus was on their chin or nose, and to one subject the agent seemed absent-minded or just unfocused. One subject considered the agent to look him in the eyes.

In the second picture, the agent was supposed to look beyond the subject, so the focus point was set at 175 units behind the viewpoint. As few as 3 subjects said the agent was fixating a point behind them, and another 3 thought the gaze was unfocused. Although there was no such intention, 6 subjects believed the agent looked to the left of the subject, in some cases in combination with behind. 2 subjects thought the agent was looking them in the eyes, and perhaps most surprisingly – since the aim was the opposite – 2 subjects thought the fixation point was in front of the subject.

The opinion about the third picture was more unanimous. The fixation point was set 5 units below the virtual viewpoint, and 11 of the 15 subjects agreed. 6 of the subjects thought that the gaze focus was in between the screen and themselves. However, this may be a sequence effect, since the preceding picture had a fixation point beyond the subject.

Finally, the last of the four pictures had the focus point set 10 units (approximately 5 degrees) to the side if the virtual viewpoint, which would produce a gaze direction to the right of the subject. This was also reported by 8 of the subjects, and 6 of the subjects said that the agent had the gaze focus behind them. 8 subjects also reported that the gaze was above their own eyes, and curiously 2 subjects still thought the agent looked to the left of them (despite that the opposite was intended).

The difference in results between the front view and the side view could be supported by the outcome of the question the subjects were asked about whether there was any difference in difficulty in determining the eye gaze between those two conditions. The front view got more yes-answers, while the subjects reported that it was easier to discriminate different fixation points in the side view. Many of the subjects reported that the variations in gaze direction in the front view were vague. They felt more certain about whether the agent looked them in the eyes or not when the face was turned away. The reason for this may be that the position of the iris in the eye contour becomes more easily determined.

5 Discussion

There is a problem with having yes/no-questions on a material that is not equally distributed in that sense. The subject may subconsciously strive to get a 50/50 distribution of their responses. (According to the implementation strategy of the agent only 4% of the stimuli should have a yes-response, or 30% if broadening the categories one step). Therefore, it is likely that the subjects have been too acceptant in their judgement. The overall distribution of yes- and no-responses is shown in figure 9.

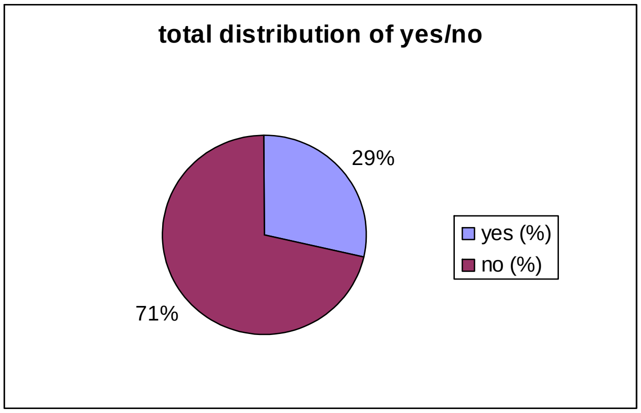


Figure 9. Total distribution of yes- and no-responses for all subjects and all conditions.

A common reaction to the cases where the focus point of eyes was set beyond the virtual viewpoint was that the agent looked absent-minded, not specifically focussing at anything else. It seems that the eye gaze behaviour is interpreted as in a dialogue context, rather than a specific estimate of a focus point in the three-dimensional space. (We will take that as evidence for that the subjects very well can consider the agent a qualified interlocutor.) It was considered especially confusing when the agent focussed somewhere in between itself and the virtual viewpoint (interpreted as between the screen and the subject), since there was nothing to focus on there – it is highly unlikely that a human will fixate a point in the air (unless a spider is hanging there).

The striking asymmetry in the x- and y-direction has several possible explanations. One is that the illumination of the agent was stronger from one side, so the shadowing might have influenced the perception of the gaze.

Another factor that may have contributed to the asymmetry is the manual mapping of texture. When carefully studying the face it can be noticed that one of the eyes (iris and pupil) is larger than the other, which is a texture mapping defect. A combined problem was that the larger iris and pupil was on the brighter side, which might have enhanced the problem. Normally the pupil gets smaller when the light increases, so the effect is likely to be confusing for the observer.

Notable is that some subjects reported about a (to begin with) unconscious tendency to look only at one of the agent's eyes, and that when noticing this, and thus changing strategy, found that they were more unsure of the direction of the gaze. The two eyes were thus not consistent, it was like if the agent was squinting (strabismus). This may also be a result of the texture mapping problem.

Concerning the asymmetry in the y-dimension, the problem may be in the design of the eyes. Compared to photos of real eyes, it can be stated that the synthetic eyes show more of the iris than real eyes tend to do, and also more of the whites (see figure 10). Either the iris should be larger, or the eyes should be more closed. Probably the latter, or maybe a combination.

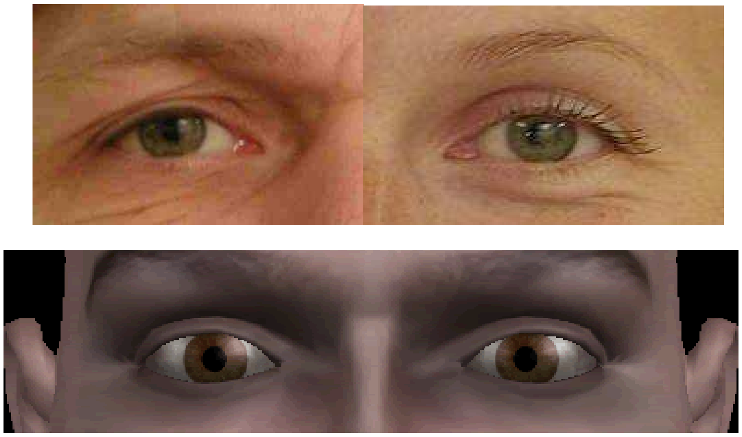


Figure 10. Above the authors' eyes are shown as examples of real eyes. Below the default setting of the agents eyes. The proportion differences in how much iris that is shown, and how much of the whites that are visible illustrate possible clues to the asymmetry in results in the y-direction.

6 Conclusions and future work

The test results showed that we managed to produce eye gazes where the subjects perceived that the agent was looking them in the eyes, but in some cases this happened when we thought it would not, according to the calculations of eye gaze in relation to the virtual viewpoint. The subjects were less sensitive to changes in the depth dimension than in the other two dimensions.

The sensitivity also diminished when the head was turned to the side.

It is worth to stress the interesting fact that the scores were not very high despite that the set-ups were done manually. This highlights the need for more studies of this kind, to thoroughly investigate how to manipulate the parameters that we have in our use in order to control the eye gaze. It is possible that some adjustments of the model are needed to ensure that not small mistakes during texture mapping or illumination risk to disturb the obviously very fine tuning that is required for eye gaze.

It is also possible that the perfect focus point is somewhere else, since we did not test combinations of the three dimensions. In a future study it would be interesting to narrow the range of each dimension and instead allow for combinations as well as smaller steps. That would give more detailed information about the sensitivity of the subjects in this respect.

Another approach that would be interesting to combine with the method described above, is to – instead of just answer yes or no – mark on a scale where the subject perceives that the fixation point is.

In the experiment in this report, static pictures were used, but for obvious reasons, animated sequences are of high interest for us. As soon as eye gaze in static pictures can be achieved, producing realistic eye movements will be the next step. One challenge is the collection of data in order to accomplish natural and trustworthy eye gaze behaviour. But before that kind of implementation can be meaningful, we must learn how to control the eyes, and how different eye gazes are perceived by the users. When introducing eye movements, there are other aspects that become increasingly important, such as the use of the muscles surrounding the eyes, blinks, and other facial movements as well as head movements. To use static pictures, as in this experiment, was a way of factoring out these features.

As with the rest of the facial movements in the talking head, it is desirable to have data driven methods for the eye gaze control. That means we have to collect data that is appropriate for a data driven animation method. The system that will be used for data collection, is the Tobii system[2]. The Tobii system uses video images of the person's face in combination with infrared light in order to track the 3D position of each eye, and to determine the target that each eye gaze is directed towards. This will permit studies of eye gaze behaviour during listening, during talking, for turn-taking signals and other communicative characteristics.

7 Acknowledgements

We wish to thank Cid - Pär Bäckström and Yngve Sundblad for demonstration and promise to lend us their eye tracking facilities. We also wish to thank all the participants of the perception experiment. This research was carried out at the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

[2] http://www.tobii.se/

8 References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman ,T. (1998): Synthetic faces as a lipreading support, *Proc of ICSLP'98.*

Argyle, M. (1988). Bodily Communication. *New York: Methuen & Company.*

Argyle, M. and Cook, M. (1976). Gaze and Mutual Gaze. Cambridge University Press, Cambridge.

Beskow,J. (2003). Talking heads – models and applications for multimodal speech synthesis. *Ph.D. dissertation*, KTH, Stockholm, Sweden, 2003.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G. (2004). Expressive Animated Agents for Affective Dialogue Systems. *In Proceedings of ADS'04.*

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T. (1997). The Teleface project - Multimodal Speech Communication for the Hearing Impaired. *Proceedings of Eurospeech '97*, Rhodos, Greece.

Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Becket, T.; Douville, B.; Prevost, S.; and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *In Proceedings of ACM SIGGRAPH '94*

Garau, M., Slater, M., Bee, S., Sasse, M.A. (2001).The Impact of Eye Gaze on Communication Using Humanoid Avatars. *CHI 2001*.Vol.3, Issue No.1.

Kleck, R.E. and Nuessle, W. (1968). Congruence between the indicative and communicative functions of eye-contact in interpersonal relations. *Brit. J. soc. Clin. Psychol.*, 7, 241-6.

Ostermann, J. (2002). Face Animation in MPEG-4. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 17-56.

Park Lee, S,; Badler, J. B.; Badler, N. I. (2002). Eyes Alive, *ACM Transactions Graphics21 (3), July 2002, ACM:NY*, 637-644.