In the Eye of the Wizard

Effects of (mutual) gaze on an avatar mediated conversation.

Student:	Bart van Gennep
Student number:	0020931
Institution:	University of Twente
Program:	Human Media Interaction
Mentors:	dr.ir. R.W. Poppe
	dr.ir. D. Reidsma
	dr. M. Poel
Assignment type:	Final Project
EC:	30
Date	19-12-2013

Summary

In this report we look at the effects of gaze on avatar mediated conversations. It is theorized that gaze has an effect on the efficiency of turn-taking, perceived dominance and perceived rapport. Would this effect also occur if one were to communicate via an avatar that copies gaze? And does a delay in copying gaze behavior to the avatar have an effect?

To test this we expand the AsapRealizer [1] with an eye motion capture and animation module using data from the Kinect [2] and video analysis. The eye module mirrors iris movements and detects blinks.

We perform an experiment where participants communicate with each other via this avatar. The experiment is executed in ten pairs with two sessions of roughly six minutes per pair. In one of the sessions participants see gaze behavior of the other participant mirrored directly, in the other session this behavior would be delayed by four seconds. The participants were given a questionnaire on perceived rapport, dominance and user satisfaction. The sessions were also recorded on video.

After the experiment we perform both a system and a user evaluation. For the system evaluation we analyze generated log files on eye and head behavior by using MATLAB [3]. This shows that the head module copies user behavior within 200ms 88% of the time. Standard deviations are also fairly low indicating few extremes. The eye module however does not perform as well. It copies user behavior within 200ms only 66% of the time, though 81% within 500ms. The eyes are closed too often, especially when participants look down.

The capabilities of the avatar are somewhat limited. It sometimes loses track of the participant for a moment. Also the avatar has a fixed appearance and was a woman. So she did not look much like the participants. It does however give a reasonable impression of where a participant was looking at.

The user evaluation was split in an analysis of the questionnaire and an analysis of video captured of the participants. There are no significant differences between the delay conditions in the questionnaire results.

For the video analysis we manually annotate participant speaking time using annotation tool ELAN. [4].We then have ELAN derive the gaps and overlaps of the session. We look at the percentage of conversation time taken up by speech, gap and overlap. Further we look at number of utterances per minute and average duration of speech, gaps and overlap. Turn taking is considered more efficient with less overlap and gaps. The video analysis showed no significant differences between the delay conditions.

In conclusion, we found no differences between the delay conditions with this avatar. This could be because there simply is no effect of gaze behavior in avatar mediated conversation. It could also be because of the limited capabilities of the avatar and the lack of likeness. Future work will have to investigate this further.

Contents

1 Introduction

People do not necessarily need to speak with words. Even only their use of gaze can speak volumes. Blinking, making and breaking eye contact, distractedly gazing away towards a sudden movement behind their conversational partner, an unlimited amount of different behaviors related to gaze and eyes contribute to interaction between humans. Should these behaviors also be employed by an embodied conversational agent? What effect does it have when this gaze behavior is out of sync with other behaviors?

Bailenson & Lee [5] demonstrated the ability to use automatic, indiscriminate mimicking (i.e., a computer algorithm blindly applied to all movements) to gain social influence. In this study participants interacted with an embodied artificial agent in immersive virtual reality. The agent either mimicked a participant's head movements at a 4 second delay or utilized prerecorded movements of another participant as it verbally presented an argument. Mimicking agents were more persuasive and received more positive trait ratings than non-mimickers, despite participants' inability to explicitly detect the mimicry.

In this paper we would like to look at the effect of modifying gaze behavior of humans communicating via an avatar. It is hypothesized that gaze has an effect on the efficiency of turn-taking, perceived dominance and perceived rapport. Would this effect also occur if one were to communicate via an avatar that copies gaze? And does a delay in copying gaze behavior to the avatar have an effect?

We must first formulate hypotheses based on what role gaze plays in a conversation. Heylen [6] performed a survey about the role of gaze in conversations. This survey shows that gaze has a role in indicating the addressee, effecting turn transitions and requesting listeners to provide backchannels.

H1: A delay of gaze behavior decreases efficiency of turn-taking.

Gaze has also been shown to influence rapport. Rapport arises when participants exhibit mutual attentiveness, positivity and coordination. [7] Mutual gaze is one of the most important indicators of mutual attention. [8] The impression of favorableness is a linear function of the amount of gaze. [9] And staring can dramatically lower perceptions of rapport in avatar communication. [8]

H2: A delay of gaze behavior reduces feelings of rapport.

Another aspect that is influenced by gaze behavior is dominance. Gazing away may reflect hesitation, embarrassment or shyness. [6] And a steady gaze and upright posture signal dominance, while the opposite can signal submission. [10]

H3: A delay of gaze behavior reduces perceived dominance.

Lastly how well the system mirrors the eyes behavior should influence user satisfaction and perceived realism.

H4: Perceived realism and user satisfaction are higher without a delay of gaze behavior.

In this study we use a similar method to look at the effects of (mutual) gaze in an avatar mediated conversation as Bailenson & Lee [5]. We perform an experiment where two people interact with each other via an avatar that mimics their behavior. In one condition the gaze behavior of the avatar is delayed by 4 seconds. We are interested in the efficiency of turn-taking, rapport, dominance, perceived realism and user satisfaction. For this experiment we use an avatar based on the system used by Poppe et al [11]. Because no satisfactory existing gaze mimicking system was found, we had to add our own gaze module. Because of this the results also include a system evaluation.

First we discuss the related work in chapter 2. Next in chapter 3 we discuss the methodology. After this we discuss the technical setup in chapter 4 and the experiment setup in chapter 5. Further we talk about the measurements and processing of data in chapter 6 and lastly we have the results in chapter 7 and the conclusions in chapter 8.

2 Related work

Much research has been done involving gaze, interaction control, avatars and mediated interaction. We first discuss the effects of gaze on interaction control, rapport and dominance. Next we talk about ECAs and avatars including their role in mediated interaction and the technology used. Much of the related work is at least a few years old. More recent research into gaze seems to focus more on specific groups such as children or the deaf. [12]

2.1 Functions of Gaze

The function of gaze in human-human face-to-face dialogues has been studied quite extensively. According to psychological studies [13] [14] [15] there are three functions of gaze. These are sending social signals, to open a channel to receive information and to regulate the flow of conversation.

Heylen [6] performed a survey about the role of gaze in conversation. This survey shows that gaze has a role in indicating the addressee, effecting turn transitions and requesting listeners to provide backchannels.

Here we first look at interaction control, which includes regulating the flow of conversation. Next we discuss the role of gaze in rapport and dominance. This last point is related to the sending of social signals.

2.1.1 Interaction Control

People take a great number of things into consideration in order to manage the flow of the interaction when conversing face-to-face. This is known as interaction control or interaction management. Examples of features that play a part in interaction control include auditory cues such as pitch, intensity, pause and disfluency, hyper- articulation, etc.; visual cues such as gaze, facial expressions, gestures, and mouth movements; and cues like pragmatic, semantic and syntactic completeness. [16]

Gaze behavior as a mechanism for interaction control is a widely studied topic within human–human communication. Kendon showed that subjects spent less than 50% of their speaking time and more than 50% of their listening time looking at their partner. [13] Kendon also found that speakers tend to look away at the beginning of utterances and interpreted it as a way for a new speaker to signal that he or she is taking the floor and doesn’t want to be interrupted. Furthermore it was found that speakers looked at their interlocutants at the end of utterances to signify that they had finished speaking and to make the floor available. Nakano et al found that gaze not only signals a request to take the turn, but that gaze behavior depends on the type of conversational act, and suggested that gaze was a positive evidence of grounding. [17]

2.1.2 Social Signals

Heylen [6] also observed that gaze is involved in signaling interpersonal attitudes. People look more at those they like, people high in dominance look more in competitive situations, people high in affiliative needs look more in a cooperative situation, negative attitudes may be signaled by looking away. Gaze is also said to be a cue for intimacy. Two of the interpersonal attitudes that can be signaled are rapport and dominance.

2.1.2.1 Rapport

Communication is more effective and persuasive when participants establish rapport. Tickle-Degnen and Rosenthal [7] argue rapport arises when participants exhibit mutual attentiveness, positivity and coordination.

One of the most important indicators of mutual attention is gaze. As we grew up, we were taught by our parents to "look someone in the eye" when we speak. During initial interaction, mutual gaze signals interest, a precondition to the continuation of the interaction. Later, gaze signals the unity of the dyad members, both in terms of the unity of their experience and the mutuality of their relationship goals. [8] Furthermore rapport is strongly tied to mimicry. [18]

2.1.2.2 Dominance

Dominance can also be signaled by gaze. Previous research on impressions of viewed behavior suggests that more gazing is perceived as indicating more power. Research has also found that perceived dominance varies as a function of whether one is gazing while listening versus speaking. [10]

2.2 ECAs and Avatars

Now that we have seen related work on the function of gaze we continue with ECAs and avatars. An Embodied Conversational Agent is a humanlike visual representation of the system, but animated personas are used to represent humans in human–human interaction as well. In games, such a representation is sometimes called an avatar, and we shall use that terminology here to separate representations of the system from representations of a human interlocutor. A talking head acting as an avatar can prove useful as a mediator in human–human communication. [16]

Avatars are digital models that may look or behave like the humans they represent. In virtual environments, avatars are often rendered dynamically, in real time, to reflect at least some user behavior or movements. [19]

Avatars are suited for research into nonverbal behavior, as we can completely control their behavior. This allows us to focus on a limited set of modalities or behaviors.

2.2.1 Avatar Mediated Interaction

Over time, our mode of remote communication has evolved from written letters to telephones, email, internet chat rooms, and video-conferences. Similarly, virtual environments that utilize digital representations of humans promise to further change the nature of remote interaction. [20]

In an avatar mediated environment interactants see the verbal and nonverbal behaviors of one another rendered onto digital avatars in real time. A powerful consequence of using this type of system is enabling transformed social interaction, the strategic decoupling of signals (about appearance and behavior) performed by one interactant from signals received by another interactant. [21]

Because this information is all digital, the frequency (how many times during an interaction the behavior is mimicked), the thoroughness (how many types of gestures or movements are mimicked at once), and the intensity (whether the mimicry is an exact mirror or only an approximation of the original gesture) of the transformation can all be precisely and automatically regulated. For that reason, such an environment can be used effectively for online manipulation of human–human interaction, although via an avatar. [21]

Bailenson & Lee [5] demonstrated the ability to use automatic, indiscriminate mimicking (i.e., a computer algorithm blindly applied to all movements) to gain social influence. In this study participants interacted with an embodied artificial agent in immersive virtual reality. The agent either mimicked a participant’s head movements at a 4 second delay or utilized prerecorded movements of another participant as it verbally presented an argument. Mimicking agents were more persuasive and received more positive trait ratings than non-mimickers, despite participants’ inability to explicitly detect the mimicry.

2.2.2 Interactive Human-ECA systems

An important final step in interactive human-ECA systems is to animate generated or observed behavior on a virtual agent and display this to the human conversational partner. Several systems have been introduced that combine online observation and behavior generation. [11]

Huang et al. [22] implemented a virtual agent with the aim of maximizing the feeling of rapport between the agent and a human conversational partner by producing speech utterances, smiles and head nods. Several authors have investigated mediated conversations in which the representation of the conversational partner is (systematically) controlled.

MushyPeek is a real-time system where the lip synchronization and head orientation of a virtual agent are generated based on detected voice activity [16]. Evaluation of these systems is carried out over entire conversations by looking at the amount of speaking [16] or subjective ratings of rapport [22].

2.2.3 Head & Eye Tracking

Head pose estimation methodologies and accuracies typically depend on the image resolution at hand and view point. High performance can reasonably be achieved with 2D or 3D Active Shape or Appearance models when dealing with high resolution images and near frontal head poses. The task is much more difficult when handling mid-resolution head video sequences and people with natural head movements that can be extremely fast and have a significant amount of profile or worse looking down head poses. [23]

Recently, depth sensors have become both affordable (e.g., MS Kinect, ASUS Xtion) and accurate. The additional depth information proves key for overcoming many of the problems inherent to 2D video data. [24]

The most widely used current eye tracking designs are video-based eye trackers. Video-oculography (VOG) systems obtain information from one or more cameras (Image data). The eye location in the image is detected and is either used directly in the application or subsequently tracked over frames. Based on the information obtained from the eye region and possibly head pose, the direction of gaze can be estimated. [25]

Eye tracking setups vary greatly; some are head-mounted, some require the head to be stable (for example, with a chin rest), and some function remotely and automatically track the head during motion. Most use a sampling rate of at least 30 Hz. Although 50/60 Hz is most common. [25]

Each setup has its advantages and disadvantages. For example head-mounted display have an easier time tracking the eyes because there are no issues with viewing angles. On the other hand remote tracking is unobtrusive but has a harder time tracking the eyes.

3 Methodology

In order to test the hypotheses on modified gaze behavior we must be able to modify gaze behavior while leaving other behaviors unchanged. The technical setup of this system is described in chapter 4. We also do a preliminary evaluation there. It was also possible to work with pre-recorded avatars, or a wizard on one side of the conversation. But then we would lose the real-time interactive component.

We perform an experiment where two people interact with each other via an avatar that mimics their behavior. In some cases the gaze behavior of the avatar is delayed by 4 seconds. We are interested in the efficiency of turn-taking, rapport, dominance, perceived realism and user satisfaction. The specifics of the experiment are described in chapter 5.

In the experiment we collect several types of data. The system generates log files, we record video & audio of the participant and we ask the participants to fill out questionnaires on rapport, dominance and user satisfaction. For a more detailed description of the measurements and processing of the data see chapter 6.

The same data that is used to animate the avatar can also be used to analyze its performance without additional recording effort. We use the log files of this data to do an extensive evaluation of the system. Further we use the questionnaires to test the subjective hypotheses on rapport, dominance and user satisfaction and we analyze the videos to test the more objective hypothesis on turn-taking efficiency. We do this video analysis by annotating the videos by session on speaking time, overlap time and gap time. Turn-taking is considered less efficient when there is relatively more overlap and gap time.

We discuss the results in chapter 7 and conclude in chapter 8.

4 Technical Setup

In order to study the effects of (mutual) gaze on an avatar mediated conversation we need to have an avatar that copies the head, face and eye behaviors of a real person. An avatar that copies head and face behaviors using the Kinect has previously been developed by ter Maat [26] and used in Poppe et al [11].

There are a few requirements on the system.

1. The system should be real-time. (It continuously captures and animates head, face and eye behaviors. There is a special focus on head and eye movements.)
2. The system should be robust. (It should be able to continue copying head, face and eye behaviors for an extended period without crashing.)
3. The system should be unobtrusive.

We first considered using an existing webcam-based program to capture eye behavior. Though the focus in these systems lies in determining where people look on the screen it can potentially be used to mirror eye movements. Unfortunately the performance of webcam-based eye trackers such as the ITU gaze tracker [27] was disappointing. It seemed to need high resolution input of the eyes for a good performance, which could not be provided by a simple webcam. Also it required calibration, which can be obtrusive to the participant.

Because of this we have created a pupil tracker using image processing of the eye regions retrieved via the Kinect. An addition advantage is that there is no need for special equipment other than the Kinect. Next we discuss the whole system, starting with the framework on which it is built. Following that we talk about the different components that handle data collection, data conversion and animation.

4.1 Framework

The system used to control the avatar is based on the SEMAINE API [28], an open source framework for building emotion-oriented systems. By encouraging and simplifying the use of standard representation formats, the framework aims to contribute to interoperability and reuse of system components in the research community. By providing a Java and C++ wrapper around a message-oriented middleware, the API makes it easy to integrate components running on different operating systems and written in different programming languages.

The SEMAINE API uses a message-oriented middleware for all communication in the system. As a result, all communication is asynchronous, which decouples the various parts of the system. The actual processing is done in "components", which communicate with one another over "Topics". Each component has its own "meta-messenger", which interfaces between the component and a central system manager. When a component is started, its meta-messenger registers with the system manager over a special meta-communication channel.

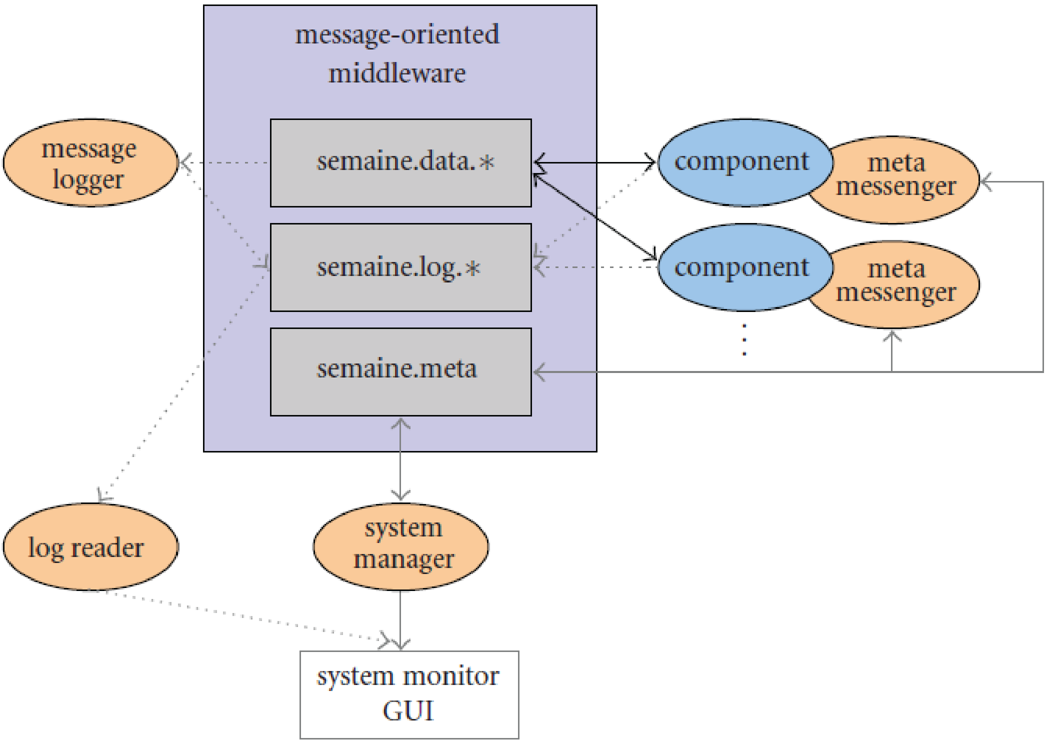


Figure 1 SEMAINE API system Architecture [28]

Figure 1 illustrates the SEMAINE API system architecture. Components communicate with each other via Topics (indicated by black arrows). Meta-information is passed between each component's meta-messenger and the system manager (grey arrows). Optionally, components can write log messages, and a message logger can log the content messages being sent; a configurable log reader can receive and display a configurable subset of the log-messages (dashed grey arrows).

4.2 Components

For the capturing and rendering of behavior we use three components, the KinectComponent, FaceMirrorComponent and ElckerlycComponent. The KinectComponent is responsible for capturing the head, face and eye behavior. The FaceMirrorComponent converts the data for use in the ElckerlycComponent. The delay in gaze behavior is also applied here. Then finally the ElckerlycComponent is responsible for rendering the behavior in the avatar.

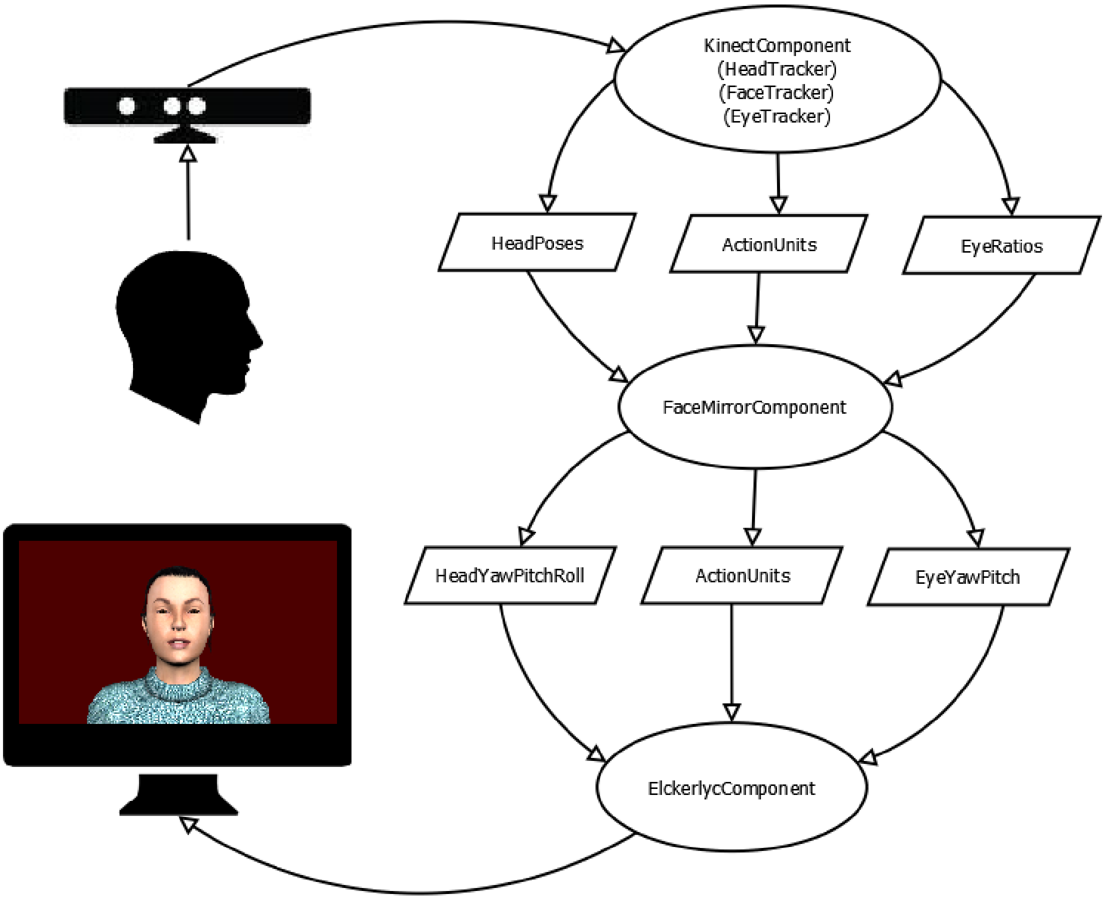


Figure 2 Dataflow between the different components

The data is sent via topics, the KinectComponent sends Head Poses, Action Units and Eye Ratios. This is converted into HeadYawPitchRoll, Action Units and EyeYawPitch. This process is illustrated in Figure 2. We will now go into more detail for each of the components where most focus lies in the motion capture of the eyes.

4.2.1 KinectComponent

The KinectComponent is responsible for gathering data about head, face and eye behaviors. We approach the Kinect using the Kinect for Windows SDK. [2] Of particular interest to us is the face tracking library [29]. The Face Tracking SDK, together with the Kinect for Windows SDK, enables us to create applications that can track human faces in real time. The KinectComponent is written in c++ with a java JNI wrapper.

The Face Tracking SDK's face tracking engine analyzes input from a Kinect camera, deduces the head pose and facial expressions, and makes that information available to an application in real time. After the data has been collected it is sent to the FaceMirrorComponent. First we will discuss the Head Poses, then the Action Units and last the Eye Ratios.

4.2.1.1 Head Poses

The Kinect calculated the head pose based on three angles: roll, pitch and yaw. The directions of these angles are illustrated in Figure 3. 0° is the neutral position. -90° is down for pitch and right for roll and yaw. +90° is up for pitch and left for roll and yaw.

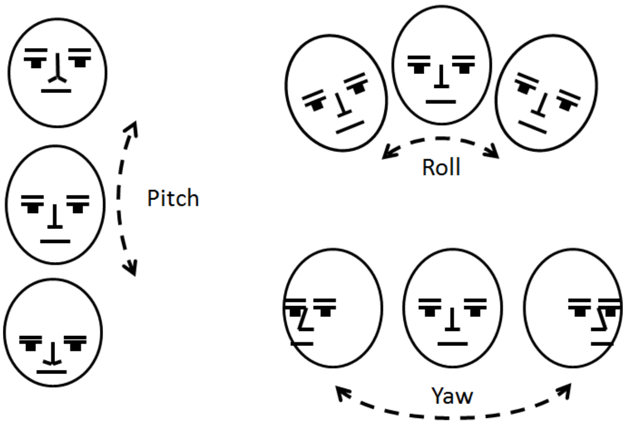


Figure 3 Head Pose Angles [29]

Face tracking tracks best when the user's head is turned towards the Kinect. For pitch it is tracks when less than 20°, but works best when less than 10°. For roll it tracks when less than 90°, but works best when less than 45°. And for yaw it tracks when the user's head yaw is less than 45°, but works best when less than 30 degrees. [29]

4.2.1.2 Action Units

The Face Tracking SDK results are also expressed in terms of weights of six AUs and 11 SUs, which are a subset of what is defined in the Candide3 model [30]. The SUs estimate the particular shape of the user's head: the neutral position of their mouth, brows, eyes, and so on. The AUs are deltas from the neutral shape that you can use to morph targets on animated avatar models so that the avatar acts as the tracked user does.

We are only interested in the AUs. Each AU is expressed as a numeric weight varying between -1 and +1. The Face Tracking SDK tracks the following AUs: Upper Lip Raiser (AU0), Jaw Lowerer(AU1), Lip Strectcher (AU2), Brow Lowerer(AU3), Lip corner depressor(AU4) and outer brow raiser(AU5). See a more detailed list of these AUs including illustrations in Appendix A.

4.2.1.3 Eye Ratios

Details on the eyes are not directly given by the Kinect and thus some processing is required. The Kinect gives a 1280x960 color image of the scene as well as the 2d coordinates of 86 face points on this image. In Figure 4 we see the positions of the different coordinates on the face. We are mainly interested in the eye coordinates, these are points 0 till 7 for the right eye and points 8 till 15 for the left eye.

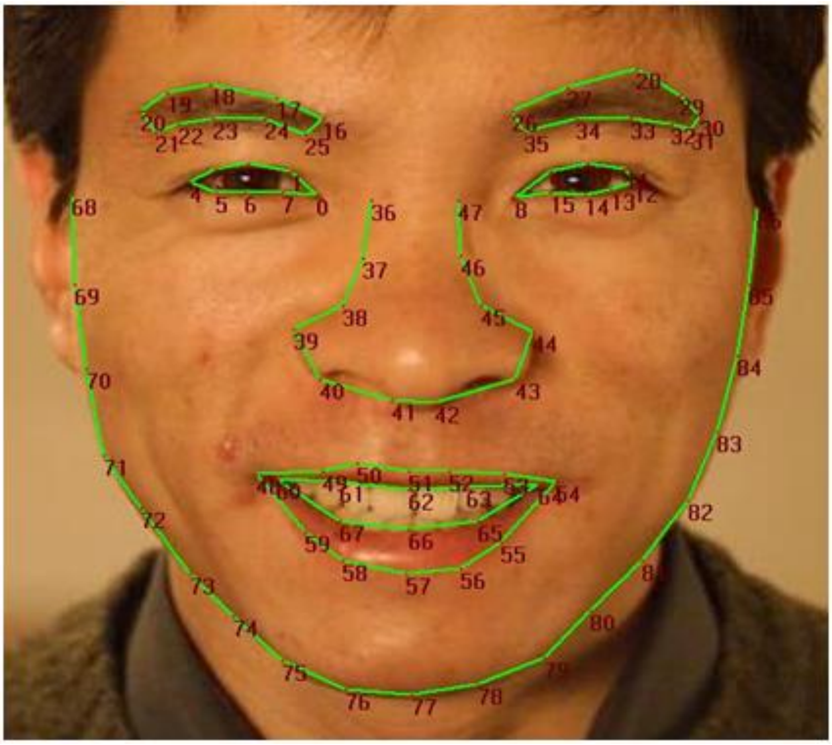


Figure 4 Kinect 2d face coordinates. Source: Microsoft [29]

We can use this data to extract and further analyze images of the eyes. For this process we use the following algorithm:

1) Get image of each eye
   a) Rotate image so eyes are on one horizontal line
   b) Also rotate 2d face points of the eyes
   c) Cut out eye image
2) Process eye image
   a) Invert it
   b) Convert to grayscale
   c) Increase contrast
   d) Convert to binary image
   e) Find biggest blob
   f) Analyze blob
   g) If eye open, find centroid of blob

This process is discussed in more detail in the next sections.

4.2.1.3.1 Get image of each eye

The eyes have to be extracted from the image given by the Kinect. For good results it is best if the eyes are upright, therefore the image needs to be rotated so that the eyes are in a horizontal line. In Figure 5a we see an example of a face with the face points indicated. In Figure 5b the image is rotated so that the eyes are on a horizontal line.

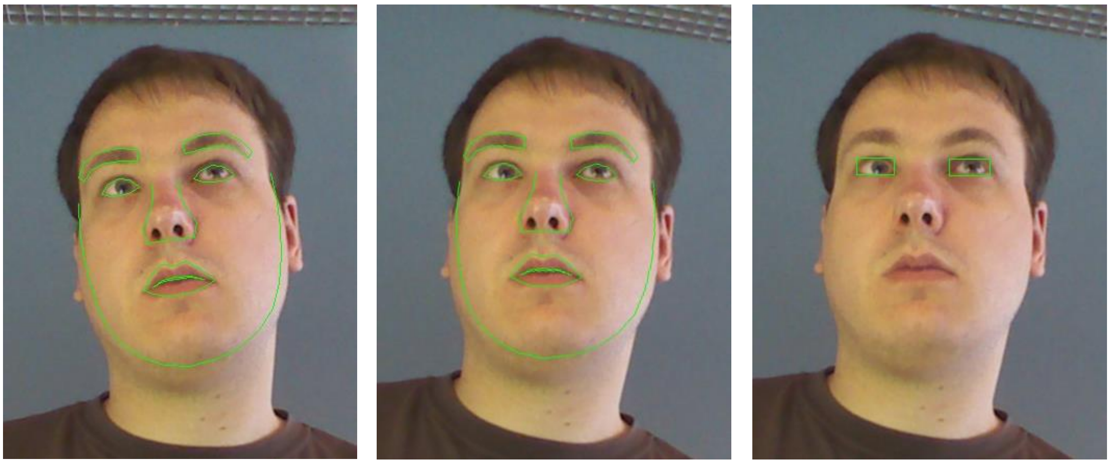


Figure 5 (a) face with 2d face points, (b) rotated so eyes are on a horizontal line, (c) eye bounding boxes
***These images have been cropped, see full images in Appendix A**

The angle of rotation is determined by getting the angle between the line going through face coordinate 0 and 8 (corners of the eye) and the x-axis. The image is rotated around the center point of the image. You could also rotate around the mid-point between the eyes, either way the rotation puts the eyes on a horizontal line. The 2d face points are rotated in the same way. Then the rotated 2d face points are used to find a bounding box around each eye as seen in Figure 5c.

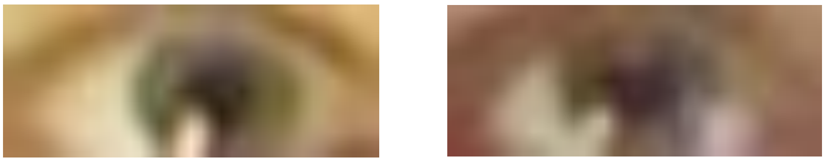


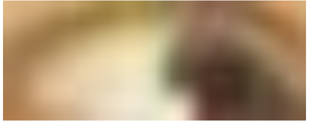
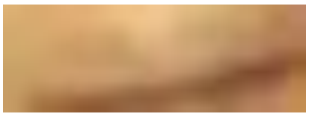
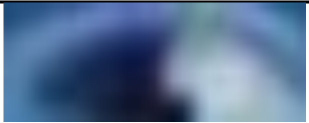
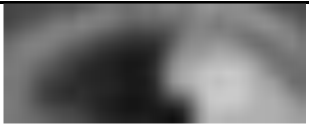
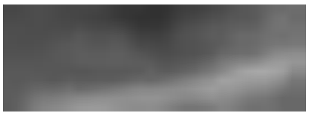
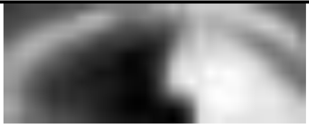
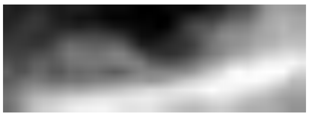
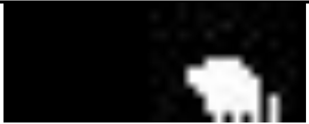
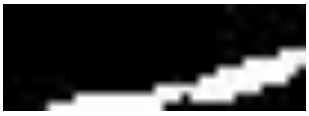
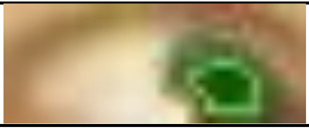
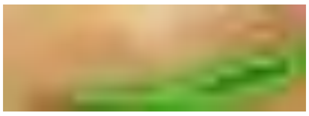
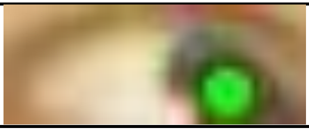
Figure 6 Right and left eye images

Using these bounding boxes the eye image can be retrieved as seen in Figure 6. What you will notice is that the resolution of the eye image itself is quite low, in this case 37x15. As the Kinect films a large area the actual eye region is small. But even with these low resolution images it is possible to determine pupil direction and detect blinks.

4.2.1.3.2 Process Eye Image

Now that we have the image, it needs to be processed. We have used the OpenCV [31] library for the image processing. The algorithm for processing of the eye image itself is based on an example from a blog about OpenCV vision apps. [32]The example uses a high resolution image, therefore the algorithm is modified slightly.

Table 1 Processing the eye image, algorithm and examples.

	Algorithm	Example open eye	Example closed eye
1)	Load the source image.		
2)	Invert it.		
3)	Convert to grayscale.		
4)	Increase contrast by equalizing histogram.		
5)	Convert to binary image using a threshold of 210.		
6)	Find biggest blob.		
7)	Analyze blob	Roundish blob -> eye open	Wide thin blob -> eye closed
8)	If eye open, find centroid of blob.		CLOSED

The algorithm is illustrated in Table 1. First we load the source image from the previous section. Next we invert the image and convert it to grayscale (3) so that dark areas (iris or eyelashes) on the image become light.

Then the contrast is increased by equalizing the histogram with the OpenCV function `equalizeHist`. A histogram quantifies the number of pixels for each intensity value. Equalization involves mapping the given histogram to a wider and more uniform distribution of intensity values so that the intensity values are spread over the whole range. [33]

Now we convert the image to binary using a threshold. The example uses 220, however because we also want to detect the closed eye, which is not as dark as the pupil, we use a value of 210. This gives us the blobs of the iris or of the eye lashes if the eye is closed. Next we find the blobs using OpenCV function `findContours` and select the biggest of the blobs using function `contourArea`. [34]

Then we divide the width of the blob by the height of the blob. If the value smaller than 1.5 the blob is considered roundish and the eye open. If it is larger than 2 it is considered wide and the eye closed. If the value lies between 1.5 and 2 it cannot be determined whether the eye is open or closed and no data is passed on for this eye. Lastly if the eye is open we find the centroid of the blob using OpenCV Moments. [34]

After the coordinates of the pupil in the eye image are found they are converted into a ratio in relation to the size of the eye image. The x coordinate is divided by the width of the eye image and the y coordinate is divided by the height. If both eyes are open the average ratio for both eyes is calculated. If the values vary, the value will nonetheless end up somewhere in the middle. This does mean that there is no vergence in this avatar. The eye tracker outputs a float array consisting of four values, namely {left eye open, right eye open, x ratio, y ratio}. The first two values are 0 for closed and 1 for open. The latter two values are a value between 0 and 1.

4.2.2 FaceMirrorComponent

The FaceMirrorComponent transforms the input data from the KinectComponent for use in the ElckerlycComponent. In addition there are a few options that can be set. It is possible to mirror behavior and to set a head pitch offset to correct for the position of the Kinect. In addition this is where the gaze delay is implemented. When a gaze delay is set, head, eye and eyelid behaviors are put in a queue which are later send to the ElckerlycComponent. Also the head, eye and eyelid behaviors can be logged to a file from this component. This component is written in Java. The configurable options are set in a configuration file. Next we discuss the head data, action units and eye data send by the FaceMirrorComponent.

4.2.2.1 Head Yaw Pitch Roll

The head yaw and pitch is received from the KinectComponent (section 4.2.1.1).Control of the avatar head joint is also done with these angles, therefore relaying the data is fairly straightforward. The pitch is modified by the set offset and the yaw and roll are mirrored if this option is set. There is also some smoothing done on head movements. Smoothing is done by taking the average of the last two frames. Lastly the head yaw, pitch and roll is send to the ElckerlycComponent in the form of a Float array.

4.2.2.2 Action Units

The action units cannot be transferred directly as the Action Units used by the ElckerlycComponent are slightly different than those given by the Kinect. Kinect uses action units from Candice [30] while Elckerlyc uses the Facial Action Coding Scheme by Ekman. [35] While Candice uses values between -1 and 1, FACS only uses positive values. Therefore in some cases one action unit from Candice related to two in FACS. The conversion values have first been determined by Poppe, Ter Maat & Heylen. [26] An overview of this conversion can be seen in Table 2. As you can also see in the table some action unit values are scaled up or down to give a better result in the avatar.

Table 2 Action Unit conversion from Kinect to Elckerlyc [26]

Kinect (Candice)		ElckerlycComponent (FACS)		Values	Scalar
AU0	Upper lip raiser	10	Upper lip raiser	Positive	0.5
AU1	Jaw lowerer	26	Jaw Drop	Positive	1
AU2	Lip stretcher	18	Lip Puckerer	Negative	1.25
		20	Lip Stretcher	Positive	0.75
AU3	Brow lowerer	4	Brow lowerer	Positive	2
AU4	Lip corner depressor	12	Lip Corner Depressor	Positive	1
AU5	Outer brow raiser	1 2	Inner brow raiser Outer brow raiser	Positive	2

In addition eye closing behavior is also send to the ElckerlycComponent using an Action Unit. However data about the eyelids is received from the Eye Ratios topic. When both eyes are closed a close eyes signal is sent to the avatar using the FACS Action Unit 43(Eyes closed). Eyes are only closed when both eyes are signaled as closed to prevent too many false positives from slipping through. When both eyes are not closed an open eyes signal is sent using the same Action Unit.

4.2.2.3 Eye Yaw Pitch

Lastly the eye behavior data itself needs to be converted. In the avatar controlled joints can be controlled with roll pitch and yaw. For the eyes roll is irrelevant. Yaw is the horizontal rotation and pitch is the vertical rotation. To convert the ratio to the pitch and yaw we use the following functions:

```
f1)      yaw = (ratio.x -0.5)*2*35

f2)      pitch = (ratio.y – 0.4)*2*35
```

The ratio is a value between 0 and 1, while the pitch and yaw are in degrees. To convert the ratio to degrees first 0.5 is subtracted and the resulting value is multiplied by 2. This gives us a value between -1 and 1. This is then multiplied by 35 to get the degrees with a value between -35° and +35°. The value is a result of trial and error, but appears to give a nice result. You will also notice that at the pitch only 0.4 is subtracted, this means pitch will be between -28° and 42°. This is to correct a deviation as the center of the eye is low in the eye image as seen in section 4.2.1.3. This data is only sent if it is available, so when at least one eye is open.

4.2.3 ElckerlycComponent

The ElckerlycComponent animates the avatar. It uses components of the AsapRealizer [1] (formerly Elckerlyc) to do this. The component receives the data from the FaceMirrorComponent and applies it to the avatar. The pitch, roll and yaw are applied directly to the avatar while the Action Units are converted to proper actions internally. In Figure 7 an example of a neutral and expressive Elckerlyc face is shown.

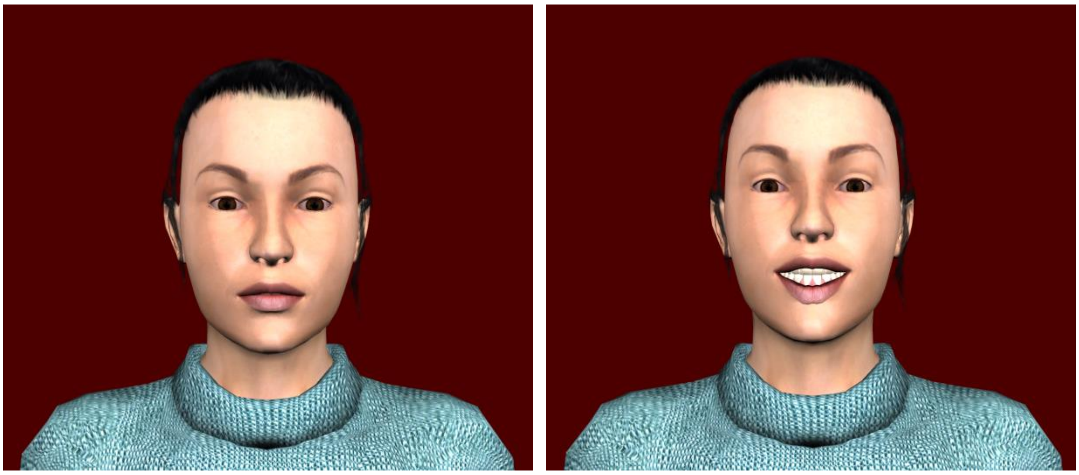


Figure 7 Neutral and expressive Elckerlyc face

This whole process leads to a copying of the user’s head pose, some facial features and eye behaviors onto an avatar. At this time there is only one embodiment which is Armandia shown above.

4.3 Preliminary System Evaluation

Though no extensive evaluation is done before the experiment there are a series of pilots. These pilots show that the system mimics head behavior well, and also mimics eye behavior ok most of the time. The Kinect does occasionally lose track of the head, and the blink module is also a bit overactive at times. This was however not disruptive enough to cancel the experiment.

In relation to the requirements the system did work real-time. Except for those instances where the Kinect lost the head. It seems robust and was fairly unobtrusive, especially because of the lack of calibration. A more extensive system evaluations is done in section 7.2 using data gathered during the experiment.

5 Experiment Setup

As mentioned we perform an experiment to test the effect of delayed gaze behavior in avatar mediated conversation. This is a within-subject experiment in which two people will converse with one another through Armandia, an embodiment of the AsapRealizer [1]. This embodiment will mimic the head and face of the participants based on information from the Kinect and image processing of the eyes (Section 4.2.1.3). As we know that gaze has an effect on dialogue, we can test what happens when we transform the captured gaze behavior. See Figure 8 for a schematic view of the experiment setup. This setup is a variation of that used by Poppe, Ter Maat & Heylen. [11]

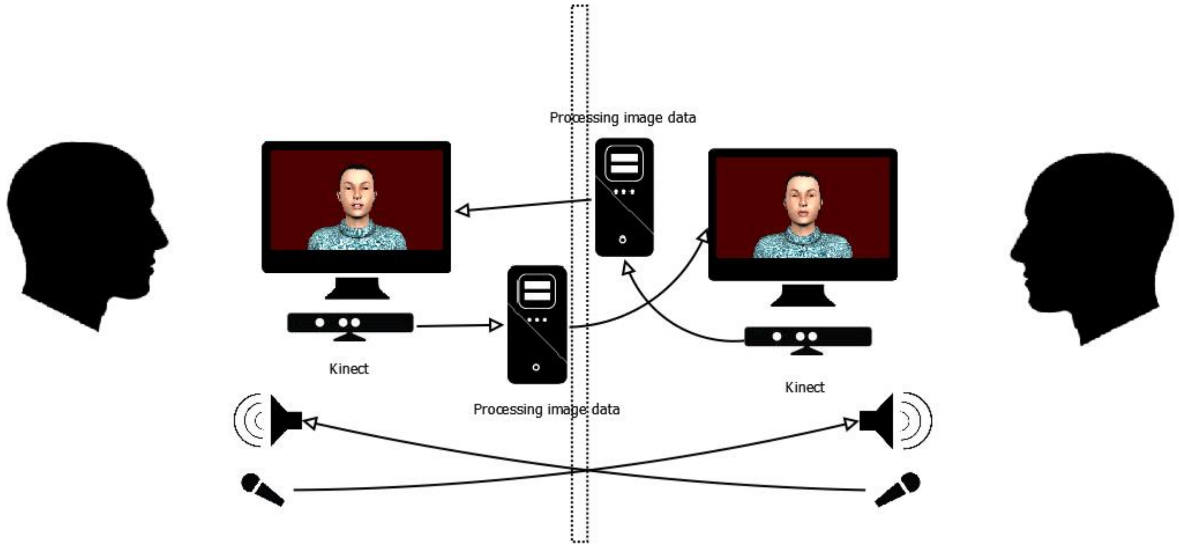


Figure 8 Schematic view of the experiment setup.

5.1 Procedure

Participants first get a minute to get used to the system. After this the experiment begins. There are two sessions of about 7 minutes each. In one the gaze behavior is left untouched, in the other the behavior is modified by delaying head and eye movements by 4s. This modification is chosen because it decouples the audio cues (speech) from the visual cues (gaze behavior). Further the time of the delay is chosen so that it should be enough to see an effect, but not be too obvious. An earlier study into the chameleon effect showed that participants did not notice a delay of 4s in mimicry behavior [5]. Here we shall see whether the same is true for delayed gaze and head behavior. Both participants have the same delay so neither has an advantage.

Participants get a shared task to encourage conversation. This is mainly to provide a topic of conversation, there are no different roles. In each session participants were given a scenario (Appendix C) to encourage conversation. In the first participants were asked to organize a party for 50 people, including location, food & drinks and such. In the other they are tasked with planning a vacation together. Participants are filmed and the avatar is recorded with the program Bandicam [36]. In addition eye and head behaviors are logged by the program.

After each session participants are asked to fill out a questionnaire. (Appendix D) The questionnaire includes items on rapport, dominance and user satisfaction.

6 Measurements & Processing

In this section we discuss the collection and processing of data. We explain how we labeled the data and how we analyze it. We have recorded video and audio of the participants and logged eye and head behaviors of the avatar. We also recorded video of the avatar, however this shall only be used for demonstration purposes.

The experiment is analyzed on both an objective and subjective scale. For the first we analyze the recorded videos and for the second we use the questionnaire data. First we discuss the system data followed by video analysis and lastly the questionnaire.

6.1 System Data

The system keeps a log of both eye and head behaviors. Below we see some examples of some log lines.

```
2013-08-07 14:30:02.786 EYES CLOSED

2013-08-07 14:30:02.79 HEAD ROLL -1.8094524 PITCH 7.5239906 YAW 10.379645

2013-08-07 14:30:02.838 EYES OPEN YAW 24.97531 PITCH 7.864197
```

As you can see the line starts with a timestamp, followed by either HEAD or EYES. For the head this line also includes roll pitch and yaw. For the eyes it includes OPEN/CLOSED with yaw and pitch. These logs are parsed by MATLAB [3]. For the evaluation we look at the time between entries to see when the system is not receiving data. Furthermore we look at how much of the time the eyes are closed and lastly we look at the derivative of the head and eye movements. This gives us an impression of how well the system performed.

6.2 Video Analysis

The videos are analyzed to learn something about the efficiency of turn taking. Turn-taking is considered more efficient when there is less pause and overlap in a conversation. Therefore we annotate for these quantities.

The videos are manually annotated on speaking time using the annotation tool ELAN. [4] ELAN is also capable of automatically annotating based on sound, but the quality of this annotation was not very good. Perhaps because the recorded sound levels are fairly low.

Both videos of a session are loaded into ELAN and speaking time of each participant is annotated on a separate tier. On another tier some session data is annotated, such as an early start or late stop of the video.

Next we have ELAN annotate the gaps of each participant on a new tier. It does this by simply taking the inverse of the speech annotations which were made by an annotator. We can now have ELAN create overlap annotations from both speech tiers, and gap annotations from the overlap of both gap tiers. Any annotations before the session started or after the session ended are removed, except from the session tier

Now we can calculate the session duration by subtracting total session annotation time from the video duration. We get all other statistics from ELAN’s view statistics option under the tiers tab. Only the percentages need to be recalculated to use the session duration rather than the video duration.

This gives us the amount of speech, overlap and gaps in the session. We look at the percentage of conversation time taken up by speech, gap and overlap. Further we look at number of utterances per minute and overage duration of speech, gaps and overlap. It could be that more overlap is simply cause by having more utterances. Turn taking is considered more efficient with less overlap and gaps.

6.3 Questionnaire

We use a questionnaire (Appendix D) to measure participant perception for rapport, dominance and user satisfaction. The dominance and rapport scales are taken from previous research and are validated. There are nineteen items on the questionnaire of which one is an open question and the others are Likert items. All items were rated on a scale anchored on 1 = strongly disagree and 7 = strongly agree. The operational unit of rapport, dominance and user satisfaction is determined by a combination of five to seven items. The internal consistency of this unit will be checked using Cronbach's alpha [37].

Table 3 Rapport Likert Scale from Jap, Robertson & Hamilton [38].

	Rapport Likert Scale	Expected Correlation
1.	I felt aware of and interested in the other participant.	+
2.	I liked and felt warm toward the other participant.	+
3.	I felt like the other participant and I were "on the same wavelength."	+
4.	I felt a comfortable rhythm with and felt coordinated with the other participant.	+
5.	I felt rapport with the other participant.	+
6.	I felt that the other participant understood the feelings that I expressed.	+
7.	I felt that the other participant shared my feelings of rapport.	+
Unit	Average of 1-7	

Table 3 shows the questionnaire items on rapport. These are taken from Jap, Robertson & Hamilton. [38] The measure includes one item each of mutual attention (1), positivity (2) and five items of coordination (3-7). All seven items are positive measures of aspects of rapport and are thus expected to have a positive correlation with feelings of rapport and thus the average of these items is used as the operational unit.

Table 4 Dominance Likert Scale from McGrae & Costa [39]

	Dominance Likert Scale	Expected Correlation
8.	I felt that the other participant was dominant.	+
9.	I felt that the other participant was submissive	-
10.	I felt that the other participant was passive.	-
11.	I felt that the other participant was active.	+
12.	I felt that the other participant was quiet.	-
13.	I felt that the other participant was talkative.	+
Unit	Average of 8-13 with inverse of 9, 10 and 12	

Table 4 shows the questionnaire items on dominance. The scale consists of six items taken from McGrae & Costa [39]. There are two opposite items on dominance (8,9), two on activity (10,11) and two on talkativity (12,13). As these are opposite items half are expected to have a positive correlation, and the other a negative correlation with perceived dominance. Therefore the operational unit is the average of these six items where the results of 9, 10 and 12 are inverted. Inversion is done by subtracting the result from eight.

Table 5 User Satisfaction Likert Scale

	User Satisfaction Likert Scale	**Expected Correlation**
14.	I focused a lot on the avatar.	+
15.	I focused a lot on the voice.	-
16.	The avatar was credible.	+
17.	The avatar helped the conversation.	+
18.	The avatar hindered the conversation.	-
19.	What did you think of the avatar in general?	n/a
Unit	Average of 14-18 with inverse of 15 and 18	

Table 5 shows the questionnaire items on user satisfaction of the avatar. There are five Likert items and one open question. There were two opposite items on focus on the avatar or voice (14,15), one item on credibility(16) and two opposite items on helpfulness for the conversation (17,18). Items 14, 16 and 17 are expected to have a positive correlation with user satisfaction while the opposite items 15 and 18 are expected to have a negative correlation. Thus the operational unit is the average of these five items where the results of 15 and 18 are inverted. Lastly the open question gives us some general comments on the performance of the avatar which might also indicate a difference between the two conditions. These comments can also help us evaluate performance of the avatar.

7 Results & Discussion

Now that the experiment has been performed we can look at the results. The evaluation is split into a system and user evaluation. Though it would have been better to do an extensive system evaluation beforehand, pilots showed that the system was usable and it was decided to use the data provided by the experiment to do a more extensive system evaluation.

The user evaluation is further split into questionnaire and video analysis results. First we evaluate the system, and then we discuss the user evaluation.

7.1 Participants & Sessions

We performed the experiment with 20 people in pairs. 17 of these people were men and 3 were women. The participants were aged between 18 and 34, with an average age of 28.

Furthermore the sessions had an average duration of 6:18 minutes where the shortest lasted 3:34 minutes and the longest 10:23 minutes

7.2 System evaluation

Before we can evaluate the experiment we must first know how the avatar performed. For this we analyzed the generated log files using MATLAB [3] (See 6.1). First we must say that the avatar crashed a few times and required a restart. This occurred six times and was divided over conditions and pairs. Next we look at the latency between the data points for the head and eye modules.

7.2.1 Latency

The Kinect color camera gives data at roughly 30 fps or every 30ms in the best case. Processing of this data creates a delay. Also not all frames are usable by the head and eye modules. Sometimes the head or eyes, or its properties, are not properly identified. In this case nothing is done with the frame. If this happens a lot this can create big gaps between data given to the avatar. In turn these big gaps cause less fluent motion of the avatar.

When all goes well the avatar should receive data roughly every 30-80ms. This is however not always the case. Therefore we have grouped the gaps into five categories. These are 0-100ms, 100-200ms, 200-500ms, 500-1000ms and over 1000ms. Anything above 200ms is already quite high, but if the gaps go over 1000ms it becomes very noticeable to the participants.

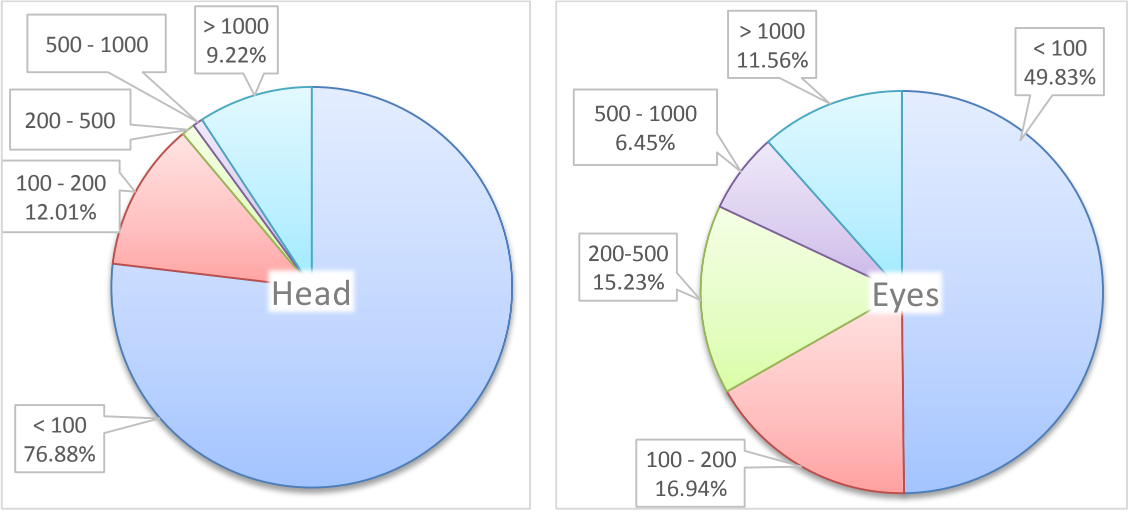


Figure 9 Percentage of time taken up by gaps of given latency in ms for the head and eye modules.

Figure 9 shows the amount of time these gaps take up for both the head and eye modules. As you can see the head performs quite well in this regard, except for the 9% of the time where the head is not detected by the Kinect. In most other cases gaps stay under 100 and 200ms.

The eyes do not perform as well. The latency is under 100ms only half the time. For a third of the time the gaps are over 200ms. This is partly logical because the detection of the eyes builds on the detection of the head. This means that only 2% of the gaps of over 1000 are caused by the eye module itself. However the eye module generates more gaps between 200 and 1000ms than the head module. The eye modules still has gaps of less than 200ms 66% of the time, and less than 500ms 81% of the time.

7.2.2 Open Eyes

We also used these gaps to calculate how much of the time the eyes were open. Unfortunately the eyes were open only 42.61% of the time. Humans have their eyes open more than that, so this indicates that the blink module did not function as well as expected. The cause of this discrepancy lies in that the module only works well when the subject is looking directly at the Kinect. When the subject is looking down, e.g. at a piece of paper, the blink module will detect closed eyes. This is not that illogical as a person does partly close his eyes when looking down. The movement data gathered on the eyes in the next section only applies to the time that the eyes are open.

7.2.3 Rotation Derivative

Next we look at the rate of change of the head and eye movements. This tells us how smooth the movements were. First we can see the range of this derivative in Figure 10. The rate of change is given in degrees per millisecond. As we can see the range of the yaw’ especially is quite large. This does make some sense as yaw is used for looking up and down. People looked up and down regularly at their paper. However the lower end of the yaw’ range for the head is so much larger than the upper end that it can be suspected that something went wrong here.

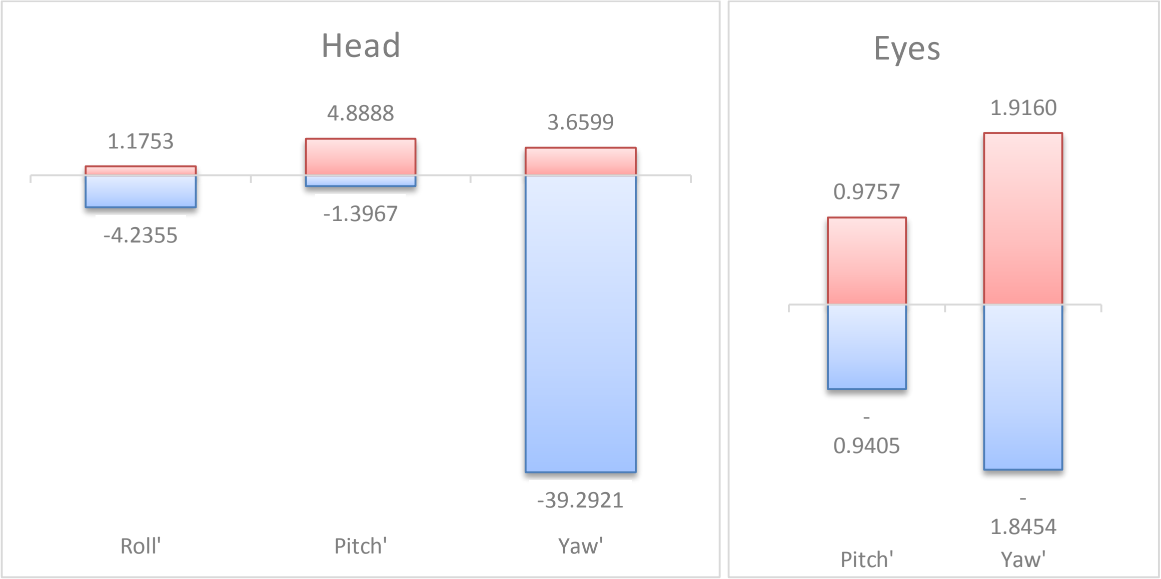


Figure 10 Range of Roll’, Pitch’ and Yaw’ for the Head and Eye modules in °/ms.

Of course only the range does not tell us that much, we also need to investigate the spread. First we calculated the mean and median, but of course both were zero or near zero. Further we calculated the standard deviation and the interquartile range to tell us something about the spread. Both of these are shown in Figure 11.

The interquartile range is the difference between the first quartile and the third quartile of an ordered range of data. It contains the middle 50 percent of the distribution and is unaffected by extreme values. [40]

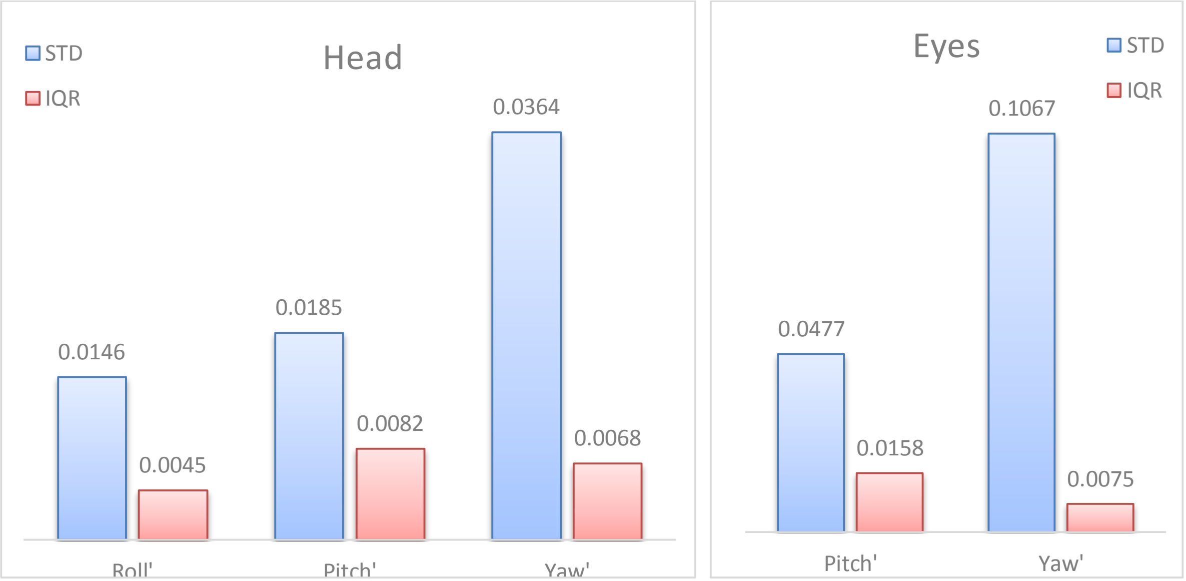


Figure 11 Standard deviation and interquartile range of Roll’, Pitch’ and Yaw’ for the Head and Eye modules in °/ms.

The interquartile range is quite small for roll’, pitch’ and yaw’ for the head and yaw’ for the eyes. For eye pitch’ it is somewhat larger. We can also see that the standard deviation is a lot larger, especially for the yaw’. This tells us that most motions are rather small but there is also a lot of variation outside of this range, especially for yaw. Compared to these values the limits of the range in Figure 10 seem especially large and might indicate some jerky movements.

In conclusion we can say that the avatar does not perform as well as expected. The head module does copy user behavior within 200ms 88% of the time. Standard deviations are also fairly low indicating little variation. The eye module however did not perform as well. It copied user behavior within 200ms only 66% of the time, though 81% within 500ms. The eyes are closed too often, especially when participants looks down at their paper.

In relation to the requirements the system has most trouble in the area of robustness. It was mostly real-time and unobtrusive in the sense that participants were not hindered by measurement equipment. In fact perhaps the avatar was a bit too unobtrusive as some participants did not pay much attention to it. The limited capabilities of the avatar could influence the experiment.

7.3 User Evaluation

Even though the capabilities of the avatar are limited, and there were some problems. The performance of the avatar was roughly the same for the two conditions and it did mimic head and gaze behavior most of the time. Therefore the results of the experiment, though diminished in worth, should still be able to tell us something.

The user evaluation consists of an analysis of questionnaires filled out by the participants and of the recorded videos. In the previous section we discussed the performance of the system. We first look at the questionnaire data.

7.3.1 Questionnaire

The questionnaire consisted of 19 questions which were part of three scales. These were rapport, dominance and user satisfaction. The questionnaire can be found in Appendix D and a more elaborate explanation of the Likert scales is given in 6.3.

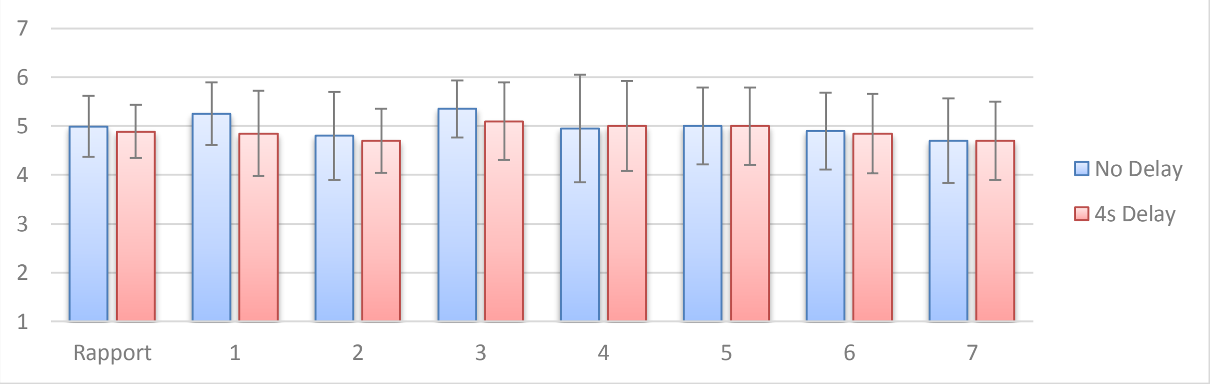


Figure 12 Rapport scale results with individual measures.

In Figure 12 we see the result of the rapport scale and its individual items. Cronbach's alpha for this scale is 0.84 which indicates a good correlation between the individual measures. As you can see the results with and without delay in gaze behavior is minimal. Only measures 1 and 3 show a drop from no delay to a 4s delay. Measure 1 asks whether participants feel aware and interested in the other participants and measure 3 asks whether participants are on the same wavelength. A pair-wise t-test on the scale gives us the following:

```
t(19)= 0.316, p = 0.755
```

This tells us that there is no significant difference on the rapport scale between the delay conditions.

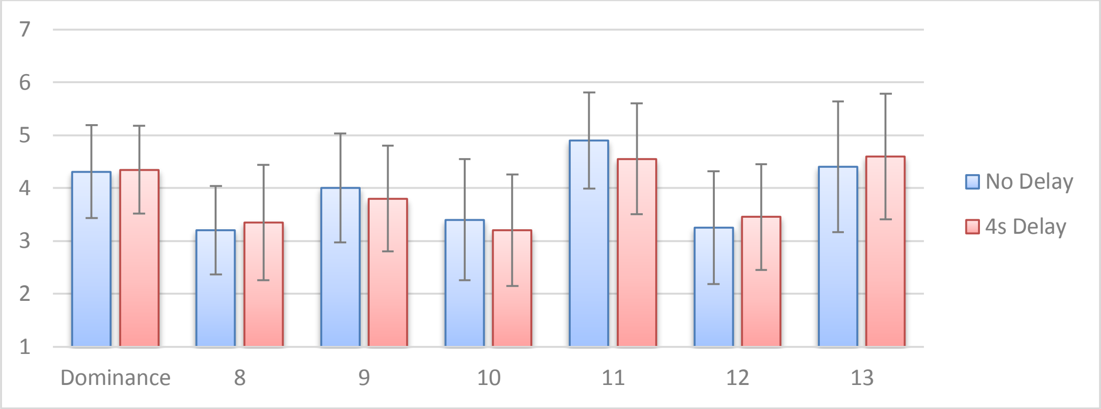


Figure 13 Dominance scale results with unaltered individual measures.

Figure 13 shows the dominance scale and the average unmodified value of each of the items. At first sight this scale might not seem to correlate well, but this is because measures 9, 10 and 12 are supposed to be inversely correlated. Taking this into account the scale has a Cronbach's alpha of 0.89 which indicates a good correlation between the individual measures. As with rapport it appears that there is little difference between the results for the no delay or 4s delay conditions. Individual measures do have some minor differences. A pair-wise t-test gives us:

```
t(19)= 0.770, p = 0.451
```

This tells us that there is no significant difference on the dominance scale between the delay conditions.

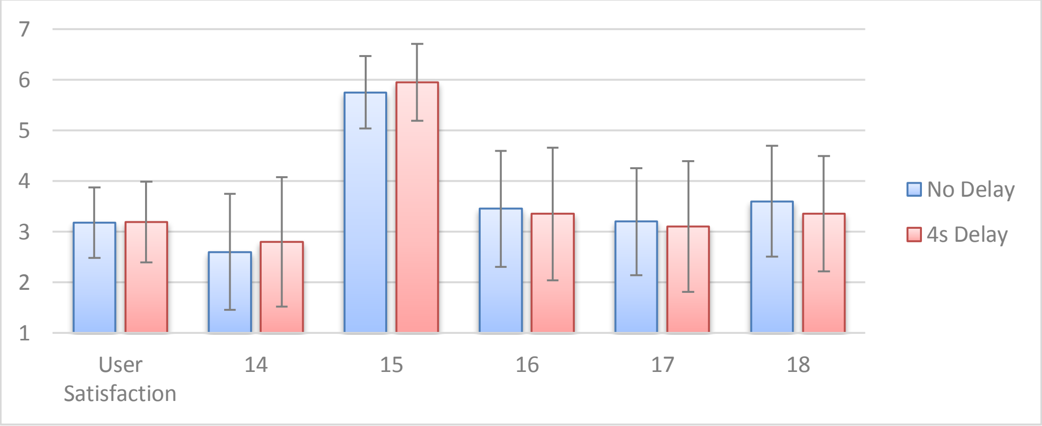


Figure 14 User satisfaction results with unaltered individual measures.

The last scale, shown in Figure 14, relates to user satisfaction. Once again this scale includes reversely correlated measures, these are 15 and 18. The figure displays the unmodified average value of each of the items. With a Cronbach's alpha of 0.77 the individual measures also seem to have a reasonably good correlation. There is almost no difference between the delay conditions. A pair-wise t-test gives us:

```
t(19)= 0.931, p = 0.363
```

This tells us that there was no significant difference on the user satisfaction scale between the delay conditions.

The only measure that scored high was 15 which asked whether participants focused on the voice of the other participant. This indicates that the participants didn’t pay much attention to the avatar at all.

The responses to the open question (19) also support this. For instance some participants remarked:

> *I did not feel the need to look at the avatar.*
>
> *I looked more at the avatar but I did not expect some expressions, so I focused more on the voice again. I looked at the avatar, but did not really see what happened. Could have been a photo instead.*
>
> *Unnatural, was not spontaneous to look at, looked more like a video.*

Which does not shine a bright light on our avatar. However there is a spark of hope as some other participants remarked:

> *I think the avatar was okay, managed to show if the other participant was looking at me or not.*
>
> *Especially the "looking down when writing" was differentiable.*
>
> *It is stimulating and transforming face expressions in a good way.*

The overall user satisfaction is quite low, this could very well be because of the limited capabilities of the avatar as discussed in section 7.2. This in turn could have caused participants to pay less attention to the avatar.

In closing, there are no significant difference between the delay conditions in the questionnaire results.

7.3.2 Video Analysis

Aside from the questionnaire we analyzed the recorded videos using annotations. We annotated for speech and let the annotation tool determine gaps and overlaps. Here we look at the frequency, average duration, and percentage of the session of speech, gaps and overlaps. For a more detailed look at what we annotated see section 6.2. These results should tell us something about the efficiency of turn taking where less gaps and overlap us more efficiency.

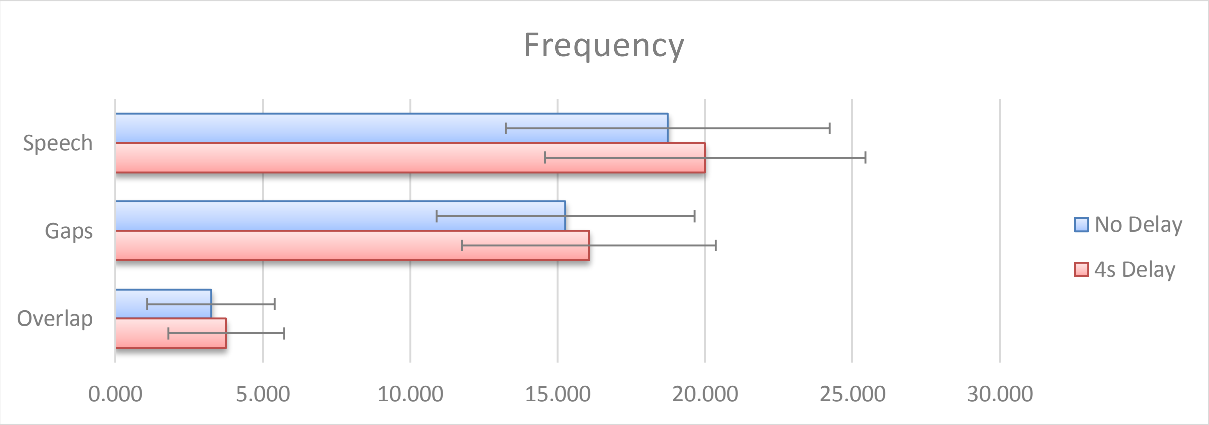


Figure 15 Number of utterances, gaps and overlap per minute under both delay conditions.

First we look at the frequency of utterances, gaps and overlap as seen in Figure 15. There are slightly more utterances, gaps and overlap under the 4s delay condition. A pair-wise t-test gives us:

```
Speech:    t(9) = 0.022, p = 0.983
Gaps:      t(9) = 0.148, p = 0.886
Overlap:   t(9) = 0.146, p = 0.887
```

There is no significant difference between the delay conditions for the frequency of speech, gaps and overlap.

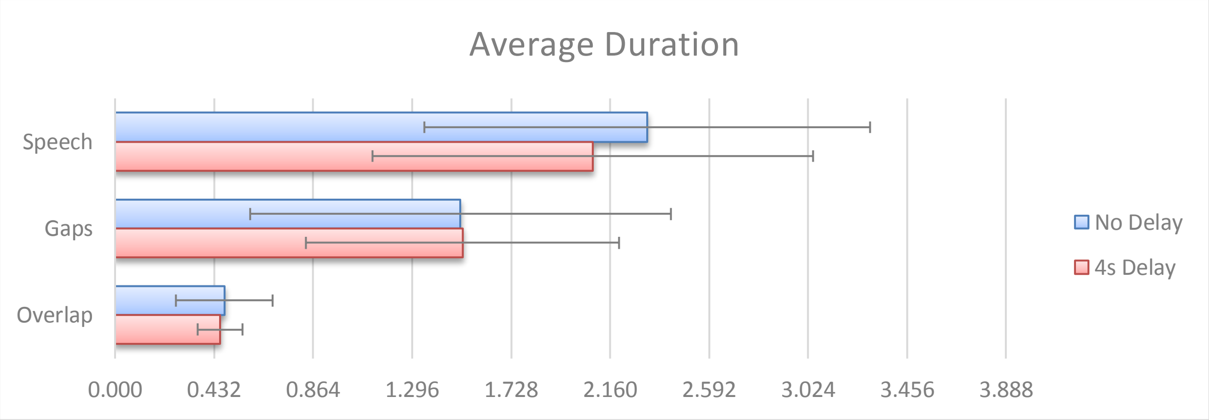


Figure 16 Average duration in seconds of utterances, gaps and overlap under both delay conditions.

In Figure 16 we see the average duration of each utterance, gap and overlap under both delay conditions. Gaps and overlaps appear to be of similar duration under both conditions while utterances are slightly shorter. However, let us look at some t-tests again:

```
Speech:    t(9) = 0.011, p = 0.991
Gaps:      t(9) = 0.938, p = 0.373
Overlap:   t(9) = 0.778, p = 0.457
```

There is also no significant difference between the delay conditions for the duration of speech, gaps and overlap.

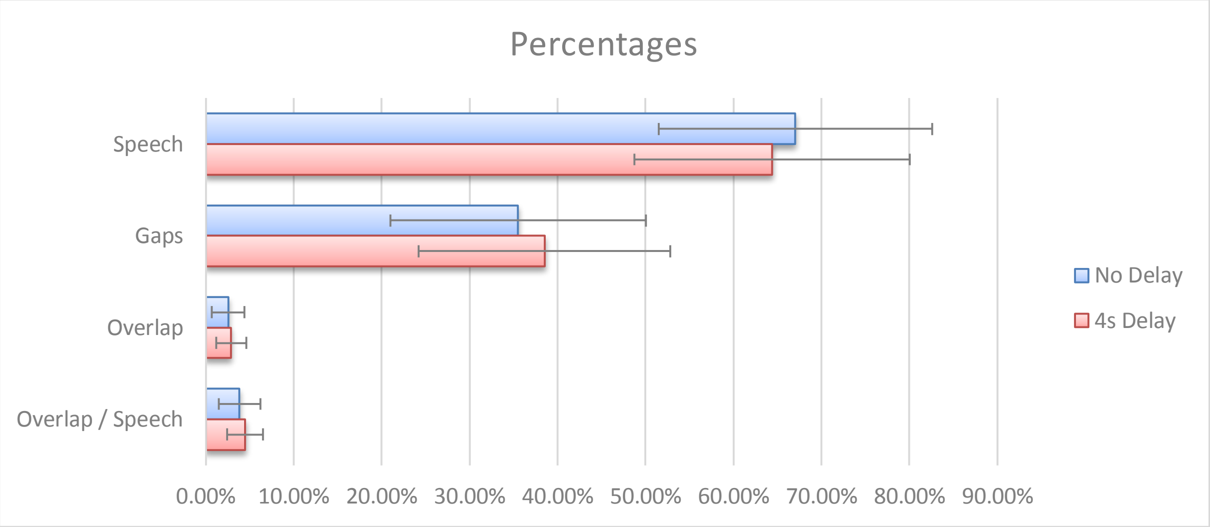


Figure 17 Percentage of the session taken up by speech, gaps and overlap. And overlap in relation to speech.

Lastly we look at the percentage of the session speech, gaps and utterances take up. In Figure 17 we see the total percentages of speech, gaps and overlap. We also see overlap as a percentage of speech. Note that the percentages of speech, gaps and overlap add up to more than 100 because of the overlap. We can see that there is slightly less speech, more gaps and more overlap under the 4s delay condition. There Is however also quite a bit of variation. As it is difficult to see exact differences and this is our main interest, we have a pie chart of speech, gaps and overlap in a session below.

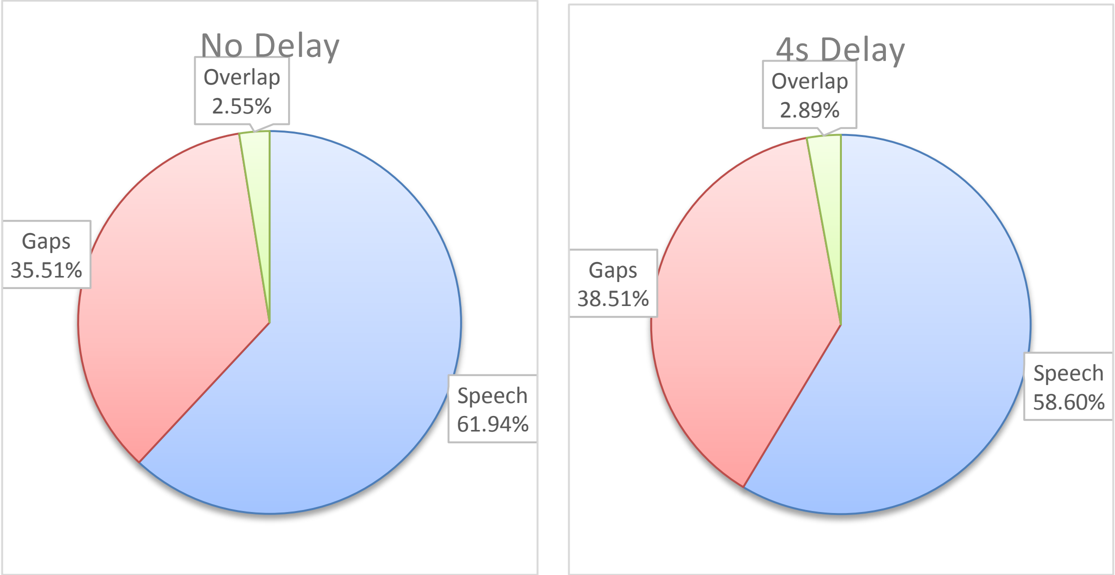


Figure 18 Percentage of the session taken up by speech, gaps and overlap under both delay conditions.

Figure 18 shows us the average distribution of speech, gaps and overlap in a session under each delay condition. As you can see both gaps and overlap are slightly increased while speech is slightly decreased in percentage. To determine statistical significance we do some t-tests again:

```
Speech:         t(9) = 0.148, p = 0.885
Gaps:           t(9) = 0.063, p = 0.951
Overlap:        t(9) = 0.440, p = 0.670
Overlap/Speech: t(9) = 0.266, p = 0.796
```

There is no significant difference between the delay conditions for the percentage of the session taken up by speech, gaps and overlap. Interestingly when taking speech level into account the difference in overlap is even less significant.

This analysis tells us there is no significant difference between the delay conditions for turn taking efficiency.

8 Conclusion

We have reached the end of our study on the effect of gaze on avatar mediated conversations. Even though the system could have performed better, we still think the results hold some value. First we will go over our hypotheses.

H1: A delay of gaze behavior decreases efficiency of turn-taking.

We hypothesized that a modification of gaze behavior decreases the efficiency of turn-taking. The video analysis done to test this hypothesis can be seen in chapter 7.3.2. There are no significant differences between the delay conditions. The most relevant difference, where gaps lasted slightly longer in the delay condition, had a chance of 37% of randomly occurring. We have failed to reject the null hypothesis.

Next we looked at rapport, dominance and user satisfaction using questionnaires (Section 7.3.1). After each session in our experiment, either with or without gaze delay, participants were given a questionnaire with 7 questions on rapport, 6 on dominance and 6 on user satisfaction. All questions were on a 7-point Likert scale except for the last which asked for a general impression of the avatar.

H2: A delay of gaze behavior reduces feelings of rapport.

We hypothesized that a modification of gaze behavior reduces feelings of report. We compared the questionnaire results on rapport between the delay conditions using pairwise t-tests. This showed no significant differences for rapport between the delay conditions. We have failed to reject the null hypothesis.

H3: A delay of gaze behavior reduces perceived dominance.

We also hypothesized that a modification of gaze behavior reduces feelings of dominance. We also compared the questionnaire results on dominance between the delay conditions using pairwise t-tests. We have found no significant differences between the delay conditions. We have failed to reject the null hypothesis.

H4: Perceived realism and user satisfaction are higher without a delay of gaze behavior.

Lastly we hypothesized that delaying gaze behavior would reduce perceived realism and user satisfaction. Here we once again used pairwise t-tests to compare the questionnaire results on user satisfaction between the delay conditions. We found no significant difference and have failed to reject the null hypothesis.

Now we have failed to reject the null hypothesis on all four hypotheses it appears that delaying gaze behavior by 4 seconds does not have an effect on our hypotheses with the used avatar. We have also observed that some participants don't look at the avatar at all, while others appreciate the novelty.

In theory gaze should play a role in all four of our hypotheses. There are several reasons why we could have found no differences for these hypotheses. First, there simply is no difference. Second, limited capabilities (Section 7.2) of the avatar caused participants to pay less attention to the avatar.

Third, poor likeness of the avatar caused participants to have trouble identifying it with the other participant.

For future work an effect of gaze behavior should try to avoid these possible issues. Perhaps a technically better avatar, or one that copies more modalities such as more facial expressions, can be developed. It could also help if the avatars were to look more like the participants, at least by using a correct gender avatar in relation to the avatar. On the other hand people already communicate via avatars that do not look like them on the internet and in games.

Alternatively it is possible to use video manipulation of webcam images. It is very noticeable when mouth movements are out of sync with sound however. This should be avoided, for example by covering up the mouth with a mouth cap. This last option would at least solve the likeness issue, and any issues in behavior detection.

In closing, though our avatar had somewhat limited capabilities and would sometimes lose track of the participant, it appears that delaying mimicking of just head and eye motion has no effect on turn taking efficiency, rapport, dominance, user satisfaction. At the very least we can say that delay in head and eye behavior mimicking is not the most deciding factor in perception of an avatar in these fields.

9 References

[1] D. Reidsma and H. v. Welbergen, "AsapRealizer in Practice - A Modular and Extensible Architecture for a BML Realizer," *Entertainment Computing,* 2012.

[2] Microsoft, "Kinect for Windows," 2013. [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/. [Accessed 7 2013].

[3] The MathWorks, Inc., *MATLAB and Statistics Toolbox Release 2012a,* Natick, Massachusetts, United States., 2012.

[4] Max Planck Institute for Psycholinguistics, "ELAN," The Language Archive, Nijmegen, The Netherlands, 2008. [Online]. Available: http://tla.mpi.nl/tools/tla-tools/elan/. [Accessed 20 8 2013].

[5] J. N. Bailenson and N. Lee, "Digital chameleons, Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments," *Pstchological Science,* vol. 16, no. 10, pp. 814-820, 2004.

[6] D. Heylen, "Head Gestures, Gaze and the Principles of Conversational Structure," *International Journal of Humanoid Robotics,* vol. 3, no. 3, pp. 1-27, 2006.

[7] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychological Inquiry,* vol. 1, pp. 285-293, 1990.

[8] N. Wang and J. Gratch, "Don't Just Stare at Me!," in *Proceedings of ACM CHI 2010 Conference on Human Factors in Computing Systems*, Atlanta, GA, USA, 2010.

[9] M. Cook and M. Smith, "The Role of Gaze in Impression formation," *Br. J. Clin. Psychol.,* vol. 14, pp. 19-25, 1975.

[10] D. Carney, "Beliefs About the Nonverbal Expression of Social Power," *Journal of Nonverbial Behavior,* vol. 29, no. 2, pp. 105-123, 2005.

[11] R. Poppe, M. t. Maat and D. Heylen, "Online Behavior Evaluation with the Switching Wizard of Oz," in *Proceedings of Intelligent Virtual Agents*, Santa Barbara, CA, 2012.

[12] O. Sandgren, R. Andersson, J. van de Weijer, K. Hansson and B. Sahlen, "Coordination of gaze and speech in communication between children with hearing impairment and normal-hearing peers.," *Journal of Speech, Language and Hearing Research.,* 2013.

[13] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica,* vol. 32, pp. 1-25, 1967.

[14] J. Duncan, "Selective Attention and the Organization of Visual Information," *J. Exp. Psych,* vol. 113, pp. 501-517, 1984.

[15] M. Argyle and M. Cook, "Gaze and Mutual Gaze," Cambridge University Press, Oxford, England, 1976.

[16] J. Edlund and J. Beskow, "MushyPeek: A Framework for Online Investigation of Audiovisual Dialogue Phenomena," *Language and Speech,* vol. 52, no. 2/3, pp. 351-367, 2009.

[17] Y. Nakano, G. Reinstein, T. Stocky and J. Cassell, "Towards a model of face-to-face grounding.," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Morristown, NJ, 2003.

[18] T. L. Chartrand and J. L. Lakin, "Mimicry, The Antecedents and Consequences of Human Behavioral," *Annual Review of Psychology,* vol. 64, pp. 285-308, 2013.

[19] D. Reidsma, R. op den Akker, R. Rienks, R. Poppe, A. Nijholt, D. Heylen and J. Zwiers, "Virtual meeting rooms: from observation to simulation," *Ai & Society,* vol. 22, no. 2, pp. 133-144, 2007.

[20] J. N. Bailenson, N. Yee and J. Blascovich, "Transformed social interaction in mediated interpersonal communication," *Mediated interpersonal communication,* pp. 77-99, 2008.

[21] J. Bailenson, A. Beall, J. Loomis, J. Blascovich and M. Turk, "Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments," *PRESENCE: Teleoperators and Virtual Environments,* vol. 13, pp. 428-441, 2004.

[22] L. Huang, L. Morency and J. Gratch, "Virtual rapport 2.0.," in *Proceedings of the International Conference on Interactive Virtual Agents (IVA)*, Reykjavik, Iceland, 2011.

[23] D. Gatica-Perez, A. Vinciarelli and J.-M. Odobez, "Nonverbal Behavior Analysi," in *Interactive Multimodal Information Management*, H. Bourlard and A. Popescu-Belis, Eds., EPFL Press, 2013.

[24] G. J. G. a. L. V. G. Fanelli, "Real time 3d head pose estimation: Recent achievements and future challenges," *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on. IEEE,* pp. 1-4, 2012.

[25] D. Witzner Hansen and Q. Ji, "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze," *Pattern Analysis and Machine Intelligence, IEEE Transactions,* pp. 478-500, 2010.

[26] R. Poppe, M. t. Maat and D. Heylen, "The Effect of Multiple Modalities on the Perception of a Listening Agent," in *Proceedings Intelligent Virtual Agents*, 2013.

[27] J. San Agustin, H. Skovsgaard, E. Mollenbach, M. Barret, M. Tall, D. W. Hansen and J. P. Hansen, "Evaluation of a low-cost open-source gaze tracker.," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, Austin, Texas, 2010.

[28] M. Schröder, "The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems," *Advances in Human-Computer Interaction,* 2010.

[29] Microsoft, "Kinect Face Tracking," 2013. [Online]. Available: http://msdn.microsoft.com/en-us/library/jj130970.aspx. [Accessed 7 2013].

[30] L. Universitet, "Candice 3 Model," 4 5 2012. [Online]. Available: http://www.icg.isy.liu.se/candide/ . [Accessed 8 2013].

[31] OpenCV, "OpenCV," 2012. [Online]. Available: http://opencv.org/. [Accessed 7 2013].

[32] N. Amin, "pupil-detection-from-an-eye-image," 2012. [Online]. Available: http://opencv-code.com/tutorials/pupil-detection-from-an-eye-image/. [Accessed 7 2013].

[33] OpenCV, "Histogram Equalization," 2013. [Online]. Available: http://docs.opencv.org/doc/tutorials/imgproc/histograms/histogram_equalization/histogram_equalization.html. [Accessed 8 8 2013].

[34] OpenCV, "Structural Analysis and Shape Descriptors," 2013. [Online]. Available: http://docs.opencv.org/modules/imgproc/doc/structural_analysis_and_shape_descriptors.html . [Accessed 8 8 2013].

[35] P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement., Palo Alto: Consulting Psychologists Press, 1978.

[36] Bandisoft, "Bandicam," 2013. [Online]. Available: http://www.bandicam.com/. [Accessed 8 8 2013].

[37] L. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika ,* vol. 16, no. 3, pp. 297-334, 1951.

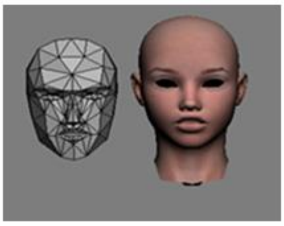
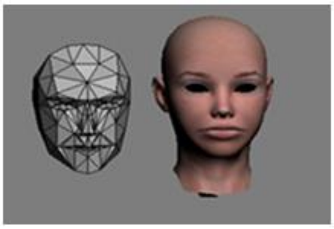
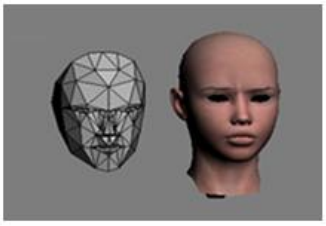
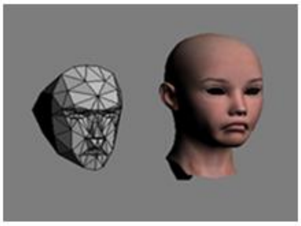
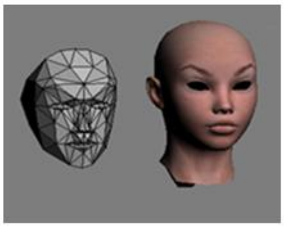
[38] S. Jap, D. Robertson and R. and Hamilton, "The dark side of rapport: Agent misbehavior face-to-face and online.," *Management Science,* vol. 57, no. 9, pp. 1610-1622, 2011.

[39] R. R. McGrae and P. T. Costa, "Validation of the Five-Factor Model of Personality Across Instruments and Observers," *Journal of Personality and Social Psychology,* vol. 52, no. 1, pp. 81-90, 1987.

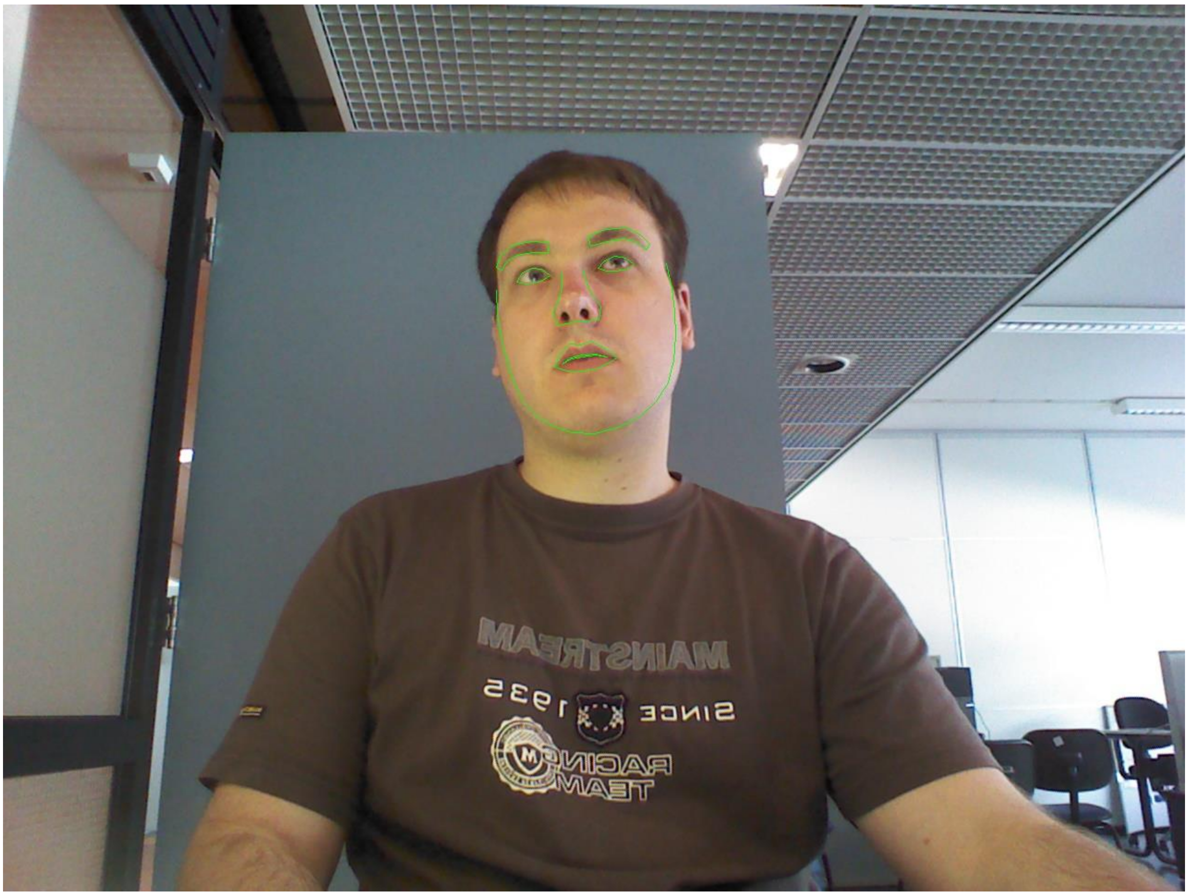
[40] G. Upton and I. Cook, Understanding Statistics, Oxford: Oxford University Press, 1996, p. 55.

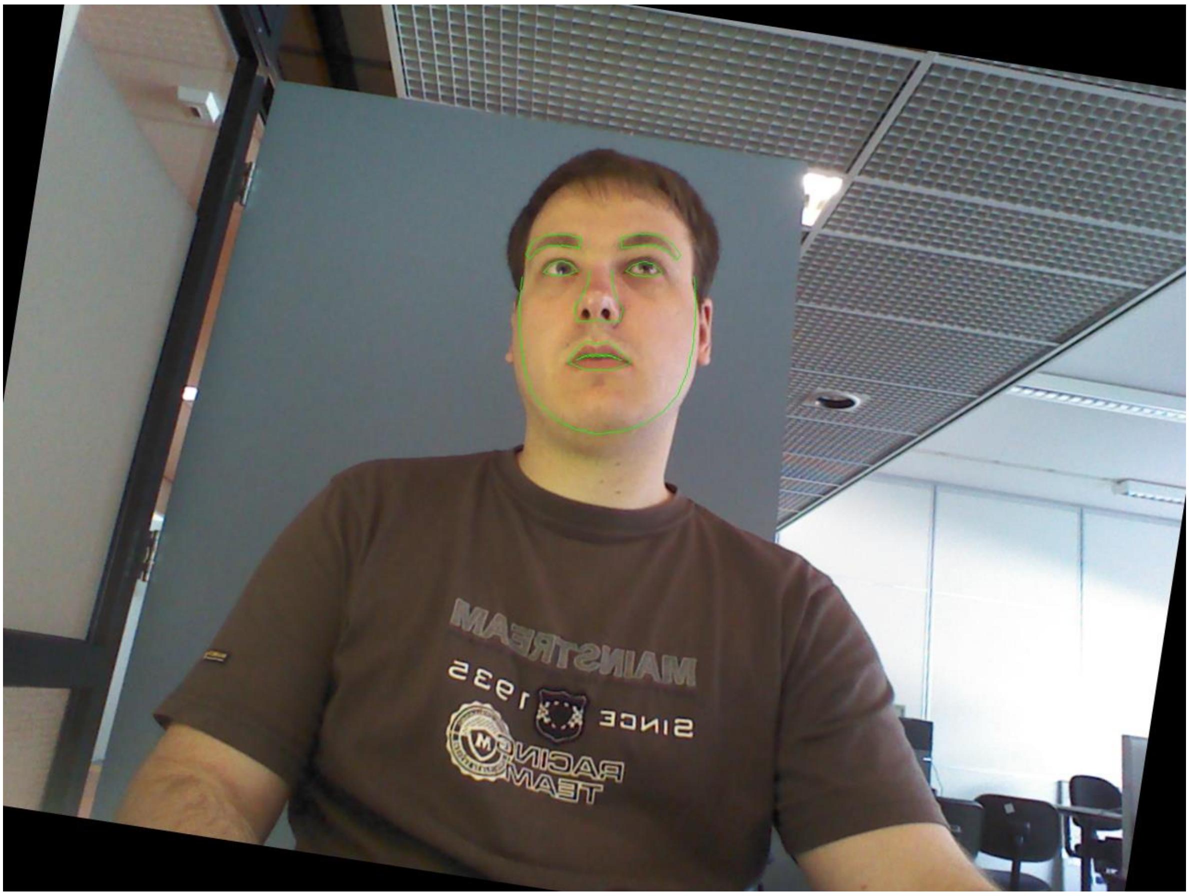
Appendix A Kinect Action Units

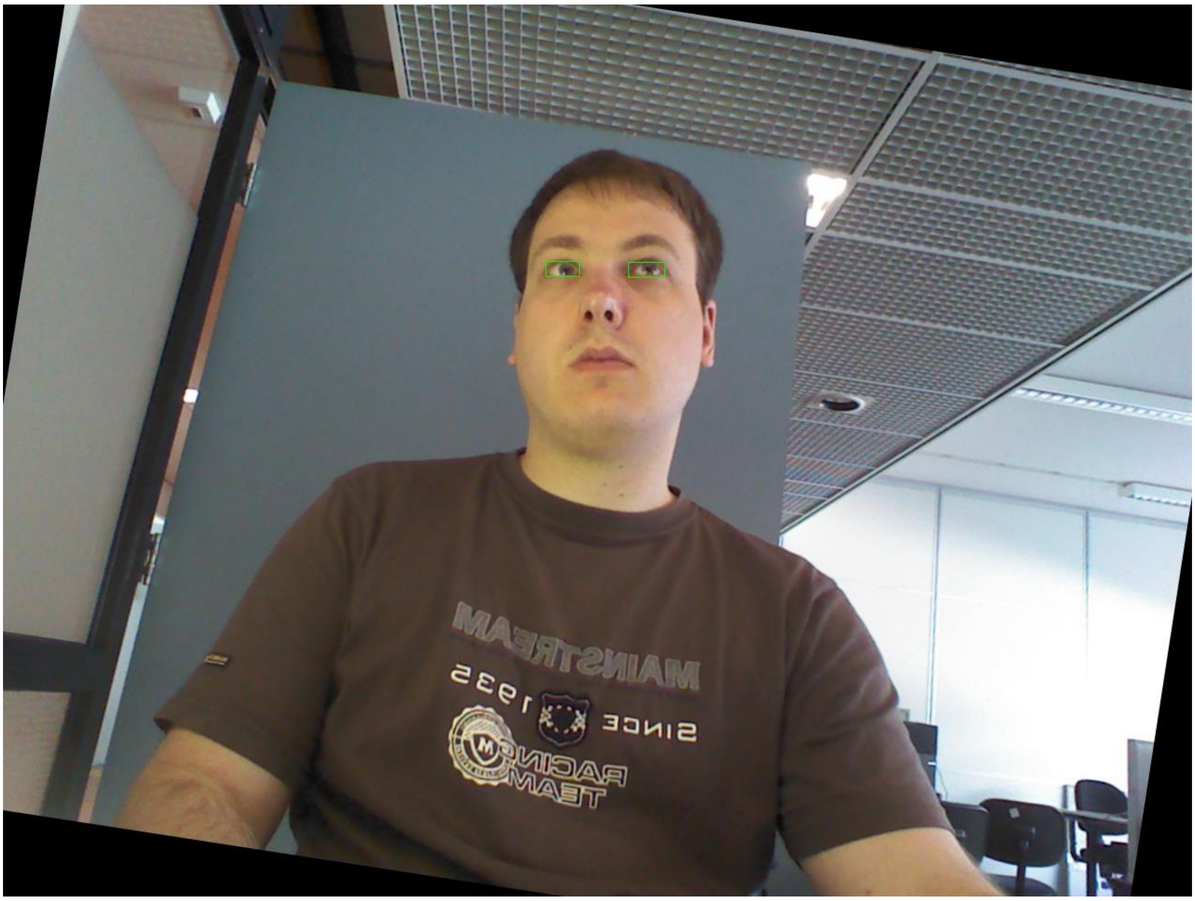
Table 6 The kinect action units. Source: [29]

AU Name and Value	Avatar Illustration	AU Value Interpretation
Neutral Face (all AUs 0)		
AU0 – Upper Lip Raiser (In Candid3 this is AU10)		0=neutral, covering teeth 1=showing teeth fully -1=maximal possible pushed down lip
AU1 – Jaw Lowerer (In Candid3 this is AU26/27)		0=closed 1=fully open -1= closed, like 0
AU2 – Lip Stretcher (In Candid3 this is AU20)		0=neutral 1=fully stretched (joker's smile) -0.5=rounded (pout) -1=fully rounded (kissing mouth)
AU3 – Brow Lowerer (In Candid3 this is AU4)		0=neutral -1=raised almost all the way +1=fully lowered (to the limit of the eyes)
AU4 – Lip Corner Depressor (In Candid3 this is AU13/15)		0=neutral -1=very happy smile +1=very sad frown
AU5 – Outer Brow Raiser (In Candid3 this is AU2)		0=neutral -1=fully lowered as a very sad face +1=raised as in an expression of deep surprise

Appendix B Full-size Images Kinect







Appendix C Scenarios

These were the scenarios used to inspire conversation.

SCENARIO 1 – PARTY TIME

Organize a party together, think of the following things:

People	50
Budget	€ 1000
Subjects of interest	Location Time Food & drink

SCENARIO 2 – VACATION FUN

Plan a vacation together, think of the following things:

People	2
Budget	€ 2000
Subjects of interest	Destination Season Duration Travel / sleeping arrangements Activities

Appendix D Questionnaire

QUESTIONNAIRE – IN THE EYE OF THE WIZARD

1. I felt aware of and interested in the other participant.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

2. I liked and felt warm toward the other participant.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

3. I felt like the other participant and I were “on the same wavelength.”

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

4. I felt a comfortable rhythm with and felt coordinated with the other participant.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

5. I felt rapport with the other participant.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

6. I felt that the other participant understood the feelings that I expressed.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

7. I felt that the other participant shared my feelings of rapport.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

8. I felt that the other participant was dominant.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

9. I felt that the other participant was submissive

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

10. I felt that the other participant was passive.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

11. I felt that the other participant was active.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

12. I felt that the other participant was quiet.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

13. I felt that the other participant was talkative.

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

14. I focused a lot on the avatar

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

15. I focused a lot on the voice

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

16. The avatar was credible

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

17. The avatar helped the conversation

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

18. The avatar hindered the conversation

○	○	○	○	○	○	○
Very Strongly Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Very Strongly Agree

19. What did you think of the avatar in general?