Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application

Nikolaus Bee, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, Marc Cavazza

Angaben zur Veröffentlichung / Publication details:

Bee, Nikolaus, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, and Marc Cavazza. 2010. "Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application." In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction - ICMI-MLMI '10, Beijing, China, November 08 - 10, 2010*, edited by Wen Gao, Chin-Hui Lee, Jie Yang, Xilin Chen, Maxine Eskenazi, and Zhengyou Zhang, 9. New York, NY: ACM Press. https://doi.org/10.1145/1891903.1891915.

Discovering Eye Gaze Behavior during Human-Agent Conversation in an Interactive Storytelling Application

Nikolaus Bee, Johannes Wagner,
Elisabeth André, Thurid Vogt
Department of Human-Centered Multimedia
Faculty of Applied Informatics, Augsburg University
Universitätsstr. 6a, 86159 Augsburg, Germany
{bee,andre}@informatik.uni-augsburg.de

Fred Charles, David Pizzi, Marc Cavazza
University of Teesside, School of Computing
Borough Road, Middlesbrough
TS13BA, United Kingdom
{f.charles,d.pizzi,m.o.cavazza}@tees.ac.uk

ABSTRACT

In this paper, we investigate the user's eye gaze behavior during the conversation with an interactive storytelling application. We present an interactive eye gaze model for embodied conversational agents in order to improve the experience of users participating in Interactive Storytelling. The underlying narrative in which the approach was tested is based on a classical XIXth century psychological novel: Madame Bovary, by Flaubert. At various stages of the narrative, the user can address the main character or respond to her using free-style spoken natural language input, impersonating her lover. An eye tracker was connected to enable the interactive gaze model to respond to user's current gaze (i.e. looking into the virtual character's eyes or not). We conducted a study with 19 students where we compared our interactive eye gaze model with a non-interactive eye gaze model that was informed by studies of human gaze behaviors, but had no information on where the user was looking. The interactive model achieved a higher score for user ratings than the non-interactive model. In addition we analyzed the users' gaze behavior during the conversation with the virtual character.

Categories and Subject Descriptors

I.4 [**Methodology and Techniques**]: Interaction techniques; H.1 [**User/Machine Systems**]: Human factors

Keywords

human-agent conversation, eye gaze, virtual agent, interactive storytelling

1. INTRODUCTION

Implementing the interactive storytelling (IS) concept involves many computing technologies: virtual or mixed reality for creating the artificial world, and artificial intelligence techniques and formalisms for generating the narrative and characters in real time. As a character in the narrative, the user communicates with virtual characters much like an actor communicates with other actors. This requirement introduces a novel context for multimodal communication as well as several technical challenges. Acting involves attitudes and body gestures that are highly significant for both dramatic presentation and communication. At the same time, spoken communication is essential to realistic interactive narratives [7].

A large variety of interfaces have been proposed for interactive storytelling including desktop-based interfaces as well as novel forms of interaction based on the use of electronic toys, conversation with virtual characters or instrumented story environments. For example, the eCIRCUS project investigates natural language conversation with virtual characters in FearNot! [3] as well as various forms of bodily and tangible interactions including interaction with a pressure sensitive dancing pad, gesture-based interaction with Nintendo's WiiMote and tangible interaction using mobile phones in ORIENT [4]. Apart from our earlier work [7] where we developed a story character that responds to the user's emotive tone, there is, however, hardly any conversational interface to interactive storytelling that emphasizes the socio-emotive aspects of interaction and integrates sophisticated technologies to recognize the user's emotive state. Furthermore, hardly any attempt has been made to study the role of eye gaze in interactive storytelling.

In our earlier work, we investigated emotional speech recognition as a novel interaction technique enabling unconstrained and natural speech interaction, by mapping a limited set of recognized emotional categories to narrative situations and virtual characters feelings. The background narrative for this system was an adaptation of three chapters of the XIXth century classic Madame Bovary by Gustave Flaubert [13]. Emma Bovary is married to a country doctor, Charles Bovary, but boredom in her married life has drawn her towards Rodolphe Boulanger. The user plays the role of Rodolphe who may address Emma or respond to her complaints and love declarations by using free-style spoken natural language input.

Eye gaze plays an important role in face to face conversations. Eye gaze can give feedback to an interlocutor and regulate and synchronize the flow of an conversation [1],

ICMI-MLMI'10, November 8-10, 2010, Beijing, China.

[17], [19]. According to Kendon [17], we can distinguish between at least four functions of seeking or avoiding to look at the partner in dyadic interactions: (1) to provide visual feedback, (2) to regulate the flow of conversation, (3) to communicate emotions and relationships, (4) to improve concentration by restriction of visual input. Even though all four functions are of relevance in the context of the interactive storytelling system, we will focus as a first step on the role of eye gaze as a means to provide visual feedback. In order to come across as believable, the virtual agent in our system should show that she is aware of the user and notices where he or she is looking. For example, when users start to stare at the agent, which is often the case in systems where users interact with virtual characters (see, for example, [25]), the virtual agent should naturally avert his gaze as humans would do in social interactions. The interactive storytelling system provides a good testbed for a gaze-aware agent since it allows the user to freely interact with the agent without any constraints on style or expressivity which might break the illusion.

In the following, we briefly review related work on techniques and studies focusing on eye gaze in human-agent communication. We then describe how the user's speech and gaze behaviors is analyzed using a framework for the synchronized analysis of multimodal input. After that, we present two eye gaze models that are both informed by studies of human eye gaze behaviors: an *interactive* eye gaze model that is sensitive to the user's eye gaze and a *non-interactive* eye gaze model that does not have the information on where the user is looking. Next, we report on a study we conducted within the this interactive storytelling system in order to compare the two eye gaze models focusing on the users' experience and their attitude towards the agent. And finally, we analyze the users' eye gaze during the speech dialogs with the virtual character.

2. RELATED WORK

A number of studies informed by human-human conversation that investigate the role of eye gaze in human-agent communication provide evidence that natural eye gaze behaviors of an agent are not only more positively perceived, but elicit also more natural responses in human users (see, for example, [8, 15, 18, 20, 24, 33]).

Colburn and colleagues [8] investigated whether natural eye gaze behaviors of an avatar elicit more natural eye gaze behaviors in users communicating with it. When an avatar was present, subjects spent more time looking at the screen. Even more attention was directed to the avatar when the agent relied on an eye gaze model that was informed by psychological studies on human-human conversation. Colburn and colleagues hypothesize that humans feel less shy when talking to a monitor than when talking to a real human. The effect occurred, however, only in the user-as-speaker condition which Colburn and colleagues attribute to the bad quality of the employed lip-synch mechanism. While Colburn and colleagues concentrate on the behavioral response to avatars employing an informed eye gaze model, Garau and colleagues [15] as well as Lee and colleagues [20] investigate the effect of informed gaze models on the perceived quality of communication by means of questionnaires. Both research teams observed a superiority of informed eye gaze behaviors over randomized eye gaze behaviors. A follow-up study by Vinayagamoorthy and colleagues [33] focused on the correlation between visual realism and behavioral realism. They found that the model-based eye gaze model improved the quality of communication when a realistic avator was used. For cartoonish avatars, no such effect was observed. While all these approaches modify parameters, such as the timing of changes in eye gaze, depending on whether the agent is speaking or listening, they do not track the users' eye gaze behaviors.

Steptoe and colleages [30] used mobile eye trackers in order to drive the eye gaze behaviors of a user's avatar in a multiparty CAVE-based system. They found that eye gaze behaviors known from human-human communication also occurred in their 3D environment. For example, participants looked at the speaker when being asked a question or looked away when thinking of an appropriate response. The avatars in their 3D environment just mimicked, however, the eye gaze behavior of the human users and did not generate eye gaze behaviors autonomously.

Rehm and André [25] described an experiment where they investigated the user's level of attention in a multi-party scenario consisting of two human and one synthetic interlocutors. Their agent was not able to perceive the users. However, since the conversation followed a pre-defined sequence of turns, the agent knew whether the user to her left or to her right was speaking and could move her head into that direction.

Similar to Steptoe and colleagues, they found that certain eye gaze practices known from human-human conversation were followed. However, the users looked significantly more often to the agent when she was talking to them than when a human user was talking to them. The experiment left open whether this difference was caused by the novelty effect of the agent or by difficulties of the users to understand the agent.

Many systems investigating interactive models of visual attention make use of head trackers. They are able to roughly assess in which direction the user is looking, but do not have more detailed information on the user's eye gaze direction. Another application using an virtual agent is the MACK system [22]. The authors use a head tracker to determine a user's gaze in a direction giving task. The animated agent explains directions on a map and monitors the user's head. In this application, lack of negative feedback indicates successful grounding. If grounding fails, the agent will perform a repair action to help the user. Based on an analysis of human-human conversation, Sidner and colleagues [28] developed a model of engagement for a conversational robot that was able to track the user's face and adjusted its gaze accordingly. Even though the set of communicative behaviors of the robot was strongly limited, an empirical study revealed that users indeed seem to be sensitive to a robot's conversational gestures and establish mutual gaze with it.

One of the earliest work that uses eye trackers for agent-based human interaction comes from Starker and Bolt [29]. They adapt "The Little Prince" to the users' current interest in a virtual scene that shows one planet from the story by

Antoine de Saint-Exupéry. Dependent on the duration and focus of the user's gaze further details of the scene are described via a text-to-speech system. Another example is the FRED system by Vertegaal and colleagues [32] that makes use of 3D animated facial agents in a multi-agent setting that are controlled by a conversational gaze model. The agents have the capability of noticing whether the user (or another agent) is looking at them. Together with the speech data they can determine if they have to listen to someone else or if they can talk. The focus of this work is the regulation of conversational flow in a multi-agent environment. That is the users' eye gaze in combination with their speech is used by the agents to determine whether to speak or to listen. Unlike Vertegaal and colleagues, we concentrate on mechanisms to establish mutual eye gaze and to respond to obtrusive staring behaviors in combination with turn taking. In our earlier work [10], we made use of an eye tracker to detect interest and attentiveness in a presentation. By means of an experiment, we showed that agents that adapted the content of their presentation to a user's eye gaze were perceived as more natural and responsive than agents that did not have this capability. The role of eye gaze as an important indicator for user attention and interest was also confirmed in a recent experiment by Nakano and Yamaoka [23].

3. ANALYSIS OF CONVERSATIONAL AND SOCIAL BEHAVIORS

Unlike earlier systems [9], our focus is not on the analysis of the semantics, but on the socio-emotional aspects of such a conversation. To analyze the user's behaviors when interacting with the virtual character, we developed a framework for multimodal signal processing in real-time [35].

3.1 Architecture

As depicted in Figure 1 the framework mediates between the sensors, which capture the user interaction, and the system, which generates in real-time the response according to the input. The information provided by our framework ranges from raw sensor data, such as eye coordinates or skin conductivity level, over low level features, such as voice pitch or heart rate, to high level description, such as the level of interest or emotional states. Exchange of information to the character control system happens continuously based on a regular update interval, or discrete, either driven by the signals, for example based on activity detection, or on request, for example when a decision has to be made. For the work presented here, we do not make use of all channels the framework supports. Rather, we focus on the acoustic properties of speech and on the user's eye gaze behaviors.

Emotional categories extracted from the user's utterance are analyzed in terms of the current narrative context to produce a specific influence on the target character, which will become visible through a change in its behavior, achieving a high level of realism for the interaction. The character's behavior is driven by an emotional planner, which determines the actions a character may undertake based on its feelings. In addition to analyzing the acoustic of speech as input to the emotional planner, we track the user's eye gaze. So far, we do not make use of eye gaze to drive the narrative. Rather, we focus on eye gaze as a means to make users feel that the character is aware of them. That is the user's eye gaze has a direct impact on the character's behavior who would, for example, avert her gaze if the user continuously stares at her, see Section 4.3.

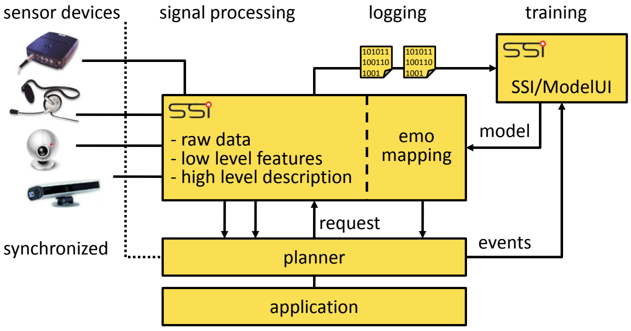

Figure 1: We measure user interaction with different sensor devices, which are synchronized and preprocessed through the framework.

3.2 Emotional Speech

Affective input from the voice is analyzed by EmoVoice [34], which has been integrated as a tool box into our framework. Real-time recognition of vocal emotions is a three-step process. First, the acoustic input signal coming continuously from the microphone is segmented into chunks by Voice Activity Detection (VAD), which segments the signal into speech frames with no pauses within longer than about 0.5 seconds. Next, from this speech frame, a number of features relevant to affect are extracted. The features are based on pitch, energy, Mel Frequency Cepstral Coefficients (MFCC), the frequency spectrum, the harmonics-to-noise ratio, duration and pauses. The actual feature vector is then obtained by calculating statistics (mean, maximum, minimum, etc.) over the speech frame ending up with around 1300 features. A full account of the feature extraction strategy can be found in [34].

In the last step, the feature vector is classified into an affective state. Integrated classifiers are Support Vector Machines (SVM) and Naïve Bayes (NB), while the latter one is used more often because it is faster and thus responds better to real-time demands. The NB classifier is very fast, even for high-dimensional feature vectors, and therefore especially suitable for real-time processing. However, it has slightly lower classification rates than the SVM classifier which is a very common algorithm used in offline emotion recognition. In combination with feature selection and thereby a reduction of the number of features to less than 100, SVM is also feasible in real-time.

3.3 Eye Gaze

Many systems investigating interactive models of visual attention make use of head trackers [22, 28]. They are able to roughly assess in which direction the user is looking, but do not have more detailed information on the user's eye gaze direction. In our work, we make use of the SMI iView X RED eye tracker. It operates with a sampling rate of 50 Hz and the tracking accuracy is less than 0.5°. The distance between the eye tracker and the user should be about 60 - 80 cm. The advantages of an unobtrusive, contact-less eye tracker include that users do not have to wear a sometimes bulky apparatus and thus are not steadily reminded that their gaze is tracked. Further, the SMI iView X RED eye tracker allows head movements horizontally and vertically up to 20 cm in each direction.

To find fixations we use the I-DT algorithm described by Salvucci et al. [27]. According to I-DT, a fixation is detected when the eye coordinates of a frame lie within the distribution *disp*. For each frame *disp* is calculated with the following formula: $disp = (max_x - min_x) + (max_y - min_y)$ where min_x, max_x, min_y and max_y are the minimum and maximum coordinate values of all points inside the frame. If *disp* is beyond a certain threshold the current frame is detected as the beginning of a fixation and then expanded by following points until the threshold is exceeded. This marks the end of a fixation. The samples in the final window are averaged to a single fixation point. For our purpose a minimum length of 120 ms and threshold of 15 pixels have been found to give reasonable results.

4. VIRTUAL CHARACTER

Our virtual character is a full body 3D character that synchronizes speech, facial displays, head and eye movements to converse with the human user. For rendering the character and its animations the Horde3D GameEngine [2] is used.

4.1 Facial Expression

Ekman and Friesen developed the Facial Action Coding System (FACS) to classify human facial expressions [11]. FACS divides the face into action units (AU) to describe the different expressions a face can display (e.g. inner brow raiser, nose wrinkler, or cheek puffer). Although FACS was originally designed to analyze natural facial expressions, it turned out to be usable as a standard for production purposes too. That is why FACS based coding systems are used with the generation of facial expressions displayed by virtual characters, such as Kong in Peter Jackson's King Kong [26]. But the usage of FACS is not limited to virtual characters in movies. The gaming industry with Half-Life 2 by Valve, also utilizes the FACS system to produce the facial expressions of their characters [31].

Emma, our virtual character (see Fig. 2), was enhanced to use the FACS to synthesize a huge set of different facial expressions. The action units were designed using morph targets and thus give the designer the full power in defining the facial expression outlook. The system includes a tool to control the single action units [6].

The FACS-based approach for a facial animation system provides the opportunity to use the Facial Expression Repertoire (FER) [12], which maps over 150 emotional expressions to the action units of FACS. Not only does it explain in detail, which action unit must be activated for certain facial expressions, it further provides a rich dataset of videos which show how the action units ought to be designed. The morph targets for the action units are modeled using the actor's templates from the FER.

4.2 Speech

The system interfaces the Microsoft Speech API to synchronize the audio output with the lip movements. This allows us to use any text-to-speech that supports SAPI 5. As the quality of common TTS systems may not be satisfactory, we integrated a module to synchronize pre-recorded audio speech files with the lip movements of the virtual character. This allows us to use highly emotional sentences or affect bursts to be spoken through a virtual character. As FACS defines several action units involving mouth muscles (e.g. lip funneler, lip tightener, mouth stretch), we utilize the FACS system for lip movements. The approach is similar to displaying facial expressions. The output from the editor to modify the single action units is stored in an XML file. Reusing the FACS approach for visemes enables Emma to display facial expressions and lip movements for speech in parallel.

4.3 Gaze Model

A number of studies that investigate the role of eye gaze in human-agent communication provide evidence that natural eye gaze behaviors of an agent that is informed by studies of human-human conversation are not only more positively perceived, but elicit more natural responses in human users (see, for example, [8, 15, 20, 33]). In our work, we start from the gaze model developed by Fukayama and colleagues [14] which allows us to specify a number of gaze parameters that influence the impression a character conveys. Their model includes two states: looking at the user and averting the gaze from the user. Three parameters define how often, how long (500 to 2000 ms) and where the virtual agent looks. The gaze targets consist of a set of random points from either all over the scene, above, below or close to the user. The probabilities of changing from one state to the other or staying in the same state depend on the amount and the mean duration of the gaze parameters. Fukayama and colleagues rated the impression particular gaze patterns conveyed that were produced by modifying the gaze parameters. They found that a medium amount of gaze and a mean duration between 500 to 1000 ms conveys a *friendly* gaze behavior. The orientation of the gaze direction did not play a decisive role in distinguishing between friendly and dominant gaze behavior, except a downward gaze was considered as less dominant. Fukayama and colleagues evaluated their gaze behavior model by only displaying eyes to the users. Thus, we evaluated their model with a full virtual head that in addition moves his head and eyes. Basically, we followed their settings, but distinguished whether the agent is speaking or listening.

Our gaze model was extended with further parameters as our virtual agent is capable of reacting to the user's current gaze using an eye tracker. The maximal and minimal duration of mutual gaze can now be set as well. Furthermore, we may indicate the maximal duration the virtual agent gazes around. We modeled two different gaze modes for our agent. In the *interactive* mode, the character looks for about 2 s (between 1 and 3 s) at the user before she averts her gaze again for about 4 s (between 2 and 6 s). Whenever the user is looking at Emma, she is tries to establish mutual gaze and to hold it for about 1 s (between 0.75 and 1.25 s). In the *non-interactive* mode, the agent's gaze model is parameterized in such a way that the agent seems to randomly look at the user or avert its gaze, and the virtual character gazes on average for a period of 1 s (0-2 s) in any state. For both modes, the duration of gaze to and away from the user is slightly adapted depending on whether the agent is talking or listening to account for the fact people look more at the interlocutor when listening than when talking, see [1].

5. EVALUATION OF THE GAZE MODELS

In the following, we present the results of a study we conducted using the interactive system as a test bed in order to find out how users perceive a character that reacts to their eye gaze. In particular, we wanted to know whether the integration of an eye gaze model had any impact on the user's perception of social presence (P), their level of rapport with the character (R), their engagement (E), the social attraction of the character (A) and the subjective perception of the story (S).

5.1 Experimental Setting

We prepared an experimental setting to compare the two eye gaze models introduced in Section 4.3 *interactive* and *non-interactive* while users are interacting with a virtual character.

The user is placed in front of a table on which the eye tracker was placed. The eye tracker with an incline of 23° was installed 80 cm above ground and 140 cm away from the projection surface. The user is seated 60 - 80 cm in front of the eye tracker. In total the user is about 2 m away from the virtual agent, which is within the *social space* according to [16]. The projection surface sizes 120 × 90 cm, which displays the virtual agent in life-size (see Fig. 2). To avoid that the user automatically stares at the virtual agent (which would happen if it was placed in the center of the visual display), we placed it on the left side. To offer an enriched scene where the user has the choice to look away from the virtual agent, Emma was placed in the dining room of her house, which includes chairs and tables (see Fig. 2).

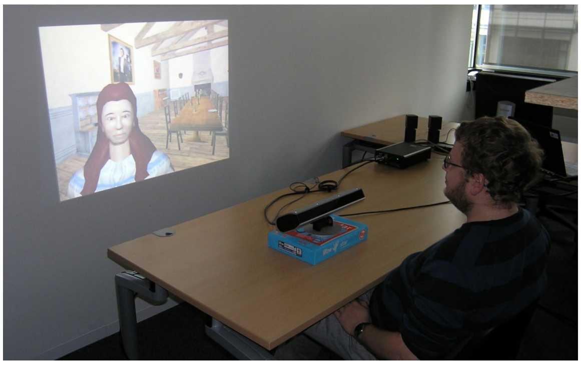

Figure 2: Set-up for the interaction with Emma.

The procedure was as follows: First, the subjects were placed in front of the projection screen. Then the eye tracker was calibrated, which took less than 2 minutes.[1] The subjects were first informed about the background of the story. Then, they were told that they would enter the story in the role of Rodolphe who finds Emma alone in the salon and should try to engage her in a conversation. To exclude any side effects resulting from dynamically evolving stories of varying quality, we decided to use fixed story lines for the experiment. Thus, for the experiment, just EmoEmma's eye gaze behavior was automated, but see [7] for an experiment with EmoEmma which included automated emotion recognition from speech. We do not consider fixed story lines as a major problem in this particular case since Emma's verbal utterances were carefully chosen so that the users could in general make sense of them. In addition, the scenario chosen - the user in the role of Rodolphe is expected to approach Emma to start an affair with her - left the user with enough space for interpretation. In the experiment, Emma produced 12 turns pausing briefly (5-10 s) after each of them to give the user a chance to respond. Emma started with 'Hello Rodolphe, I am so delighted!' and the user could for example answer with 'Hello Emma, I feel just the same way!'. The whole process for each subject took about 20 minutes including the introduction to the story sequence whereby one interaction sequence took about 3 minutes. The order of the two gaze models (i.e. *interactive* and *non-interactive*) was randomized for each subject to avoid any bias due to ordering effects. Overall, we recruited 19 subjects (2 females and 17 males) with a mean age 25.3 (SD = 3.1) for the experiment. All subjects were native speakers of German.

5.2 Social Presence, Engagement and Interactional Rapport

The objective of the study was to find out whether the different modes had any impact on the subjects' experience ratings. In particular, we used a post-questionnaire with a 9-point rating scale (from strongly disagree to strongly agree) to assess the subjects' sense of social presence (P), their level of rapport with the character (R), their engagement (E), the social attraction of the character (A) and the subjective perception of the story (S).

Measures. *Social Presence (P).* We assessed the subjects' sense of social presence using the items "I had the feeling that Emma was aware of me.", "I had the feeling of personal contact to Emma.", "Emma was impersonal." (reverse coded), and "Emma was reserved." (reverse coded).

Rapport with the Character (R). The level of rapport with the virtual character was measured using the items "I would have liked to continue the interaction with Emma.", "Emma's behavior was natural.", "I had the feeling that Emma reacted on me.", and "Emma's behavior was synchronous to mine.".

Engagement (E). We indexed the user's level of engagement with the following two items: "I enjoyed the first meeting with Emma." and "I found it easy to flirt with Emma.".

Social Attraction of the Character (A). The users' social attraction of the character was measured using "I had the feeling, that Emma was interested in me." and "Emma was sympathetic.".

Perception of the Story (S). The subjective perception of the story was assessed using the items "I would like to know how the episode with Emma continues.", "I had no problems to

[1] To measure user engagement, we also connected users with skin conductance and blood volume pressure sensors and recorded their upper body. These data have, however, not yet been analyzed.

empathize with the part of Rodolphe.", and "I had the feeling to influence the story with my eye gaze.".

Results. The significance analyses between the interactive gaze mode and the non-interactive mode were conducted using a paired two-tailed t-test. A look at Figure 3 reveals that all groups received more positive ratings for the *interactive* gaze model than for the *non-interactive* gaze model.

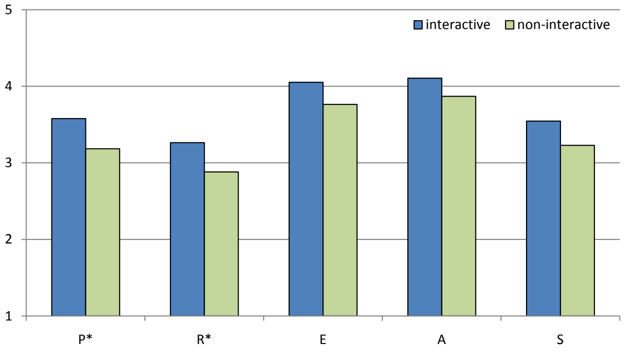

Figure 3: Results for the questions compared with the *interactive* and *non-interactive* gaze model while interacting with Emma ($^*p < 0.05$).

The significance test reveals that the presence measure differs significantly between the *interactive* and *non-interactive* gaze mode (P: $t(75) = 2.6, p = 0.01, r = 0.29$). Also the rapport measure reveals significant differences between these two modes (R: $t(75) = 2.3, p = 0.02, r = 0.26$). However, the other measures did not reveal any significant differences (E: $t(37) = 1.6, p = 0.11$; A: $t(37) = 1.2, p = 0.25$; S: $t(56) = 1.5, p = 0.15$).

5.3 Analysis of the Subjects' Eye Gaze Behaviors

First of all, we investigated to what extent the subjects were looking at Emma while she was speaking or silent. This gives us evidence whether the user interacts with Emma in a similar way as they would do in human-human interaction. We calculated the fixation points from the raw eye gaze data using the algorithm presented in Section 3.3. Furthermore, we divided the scene into two areas. The first area covers the eyes of the virtual character and the second area the rest of the scene.

We found that independent from the gaze mode, the users were looking at Emma around 76% of the time in contrast to Kendon [17] who found that in human-human interaction a human is looking on average 50% of the time at an interlocutor. Further, Kendon reports that this quote varies from 28% to 70% whereas we found a variation of 46% to 98%.

Argyle and Cook [1] found that humans look about 75% at interlocutors while listening and 41% while speaking. Independent from the gaze mode, we found that users interacting with a virtual agent look about 81% of the time at the agent while listening and about 71% of the time at Emma while speaking. Although the users were in total much more looking at Emma, the relationship between listening and speaking remains comparable (i.e. the user looks at the interlocutor considerably longer when listening than when speaking) to human-human interaction. These findings are in line with an study conducted by [25] and they ascribe them to the novelty effect of the agent.

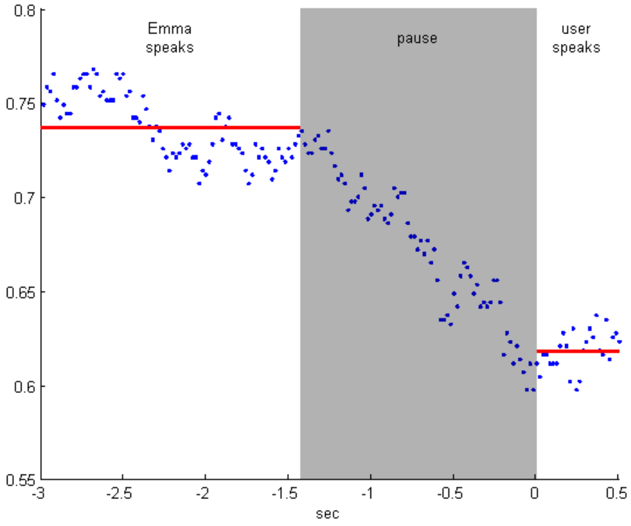

Figure 4: Eye gaze pattern before the users start speaking. The vertical axis indicates the gaze target(0 = looking away, 1 = looking at Emma, *red line* = average) during conversation. The user starts speaking at $t = 0$.

Considering a multimodal gaze model that takes the user's eye gaze and speech into account, we analyze where the users are looking when they start and stop speaking. We expect findings that can be integrated in a multimodal interactive gaze model for a virtual character that enables the agent to detect whether a user plans to say or answers something or is expecting further advices from the system. In this way an attentive system could recognize whether the current stimulus already suffices to expect an answer or feedback from the user or if the system needs to elaborate the current dialog part.

Figure 4 shows the gaze pattern when the users start speaking. We chose to analyze an 3.5 seconds interval, where we looked at the 3 seconds before the users started to speak and 0.5 seconds after the users started to speak. The users started to speak at $t = 0$ and we collected overall 430 utterances for this analysis. In Figure 4, three phases are shown: *Emma speaks*, *pause* and the *user starts speaking.* The pause after Emma speaks and the user answers is on average 1.43 seconds (SD = 1.05). The vertical axis indicates the users' current gaze target, where 0 means the user looks away and 1 that the user looks at Emma's face. On average, the users looked significantly more at Emma while she was speaking than when the users started to answer, where they averted their gaze ($t(102) = 32.8, p = 0, r = 0.96$). The finding is not only statistically significant, but has also a large effect (r) and so indicates a substantive finding. Morency and colleagues [21] also found that users avert their gaze while thinking or answering.

In Figure 5 we plot the users' gaze pattern at the end of their utterance. We analyzed a 2.5 seconds interval, where we looked at 0.5 seconds before the users stop speaking and 2 seconds afterwards. The users stopped speaking at $t = 0$ and we collected overall 378 utterances from the users. The users started to looked significantly more often at Emma face after they stopped speaking ($t(123) = 6.2, p = 0, r = 0.49$). Looking at the gaze pattern in Fig. 5 reveals that after the users end their utterance, their gaze behavior looks like an increasing sawtooth pattern. Which means that they are rhythmically alternating their gaze between Emma's face and the rest of the virtual scene.

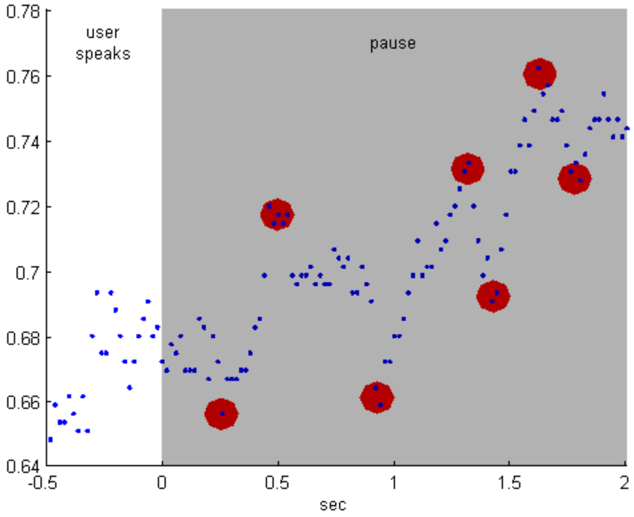

Figure 5: Eye gaze pattern after the users stopped speaking. The vertical axis indicates the gaze target(0 = looking away, 1 = looking at Emma). The user stops speaking at $t = 0$.

6. CONCLUSION

In this paper, we presented an interactive eye gaze model for a virtual character. The eye gaze model was integrated and tested within an existing story telling system in which the user could freely interact with the main character impersonating one of the story characters. An evaluation provided interesting results regarding the users' perception of the interaction and their attitude towards the character. We found that the interactive gaze mode led to a better user experience compared to the non-interactive gaze mode. Indeed, the interactive gaze mode achieved a higher score for all items of a questionnaire measuring the user's sense of social presence, their level of rapport with the agent, their engagement, the social attraction of the character and the subjective perception of the story. These results are in line with our previous work [5] where we analyzed the interaction with a virtual character based on eye gaze only. Additionally, we found that users adhere to patterns of gaze behaviors for speaker and addressee that are also characteristic of dyadic human-human interactions. However, they looked significantly more often to the virtual interlocutor than is typical of human-human interactions.

Our future work will concentrate on a more thorough analysis of the users' eye gaze behaviors in the two different modes to investigate whether the different eye gaze behaviors of the agent have any impact on the users' eye gaze behaviors. So far, we implemented an interactive eye gaze model to improve the users' overall experience when interacting with a story character. Eye gaze, could, however, also be used as an implicit or explicit input channel to drive a narrative. For example, a story might develop differently depending on the user's level of engagement when conversing with the story characters. We therefore plan to focus on the role of eye gaze tracking along with pause and speech detection as an important indicator of engagement in a conversation to influence the virtual characters' behavior. In combination with the emotional tone of the user's voice, the user's level of engagement could be used as an additional factor to determine scene evolution.

7. ACKNOWLEDGMENTS

The work reported in this paper is partially supported by the European Commission under grant agreement IRIS (FP7-ICT-231824).

8. REFERENCES

[1] Argyle and Cook. *Gaze & Mutual Gaze*. Cambridge University Press, 1976.

[2] Augsburg University. Horde3D GameEngine. `http://mm-werkstatt.informatik.uni-augsburg.de/projects/GameEngine/`.

[3] R. Aylett, S. Louchart, J. Dias, A. Paiva, M. Vala, S. Woods, and L. Hall. Unscripted narrative for affectively driven characters. *IEEE Comput. Graph. Appl.*, 26(3):42–52, 2006.

[4] R. Aylett, N. Vannini, E. André, A. Paiva, S. Enz, and L. Hall. But that was in another country: agents and intercultural empathy. In *AAMAS '09: Proc. of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 329–336, Richland, SC, 2009.

[5] N. Bee, E. André, and S. Tober. Breaking the ice in human-agent communication: Eye-gaze based initiation of contact with an embodied conversational agent. In *9th International Conference on Intelligent Virtual Agents (IVA)*, pages 229–242. Springer, 2009.

[6] N. Bee, B. Falk, and E. André. Simplified facial animation control utilizing novel input devices: A comparative study. In *International Conference on Intelligent User Interfaces (IUI '09)*, pages 197–206, 2009.

[7] M. Cavazza, D. Pizzi, F. Charles, T. Vogt, and E. André. Emotional input for character-based interactive storytelling. In *AAMAS '09: Proc. of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 313–320, 2009.

[8] A. Colburn, M. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces. Technical report, Microsoft Research, 2000.

[9] S. Dow, M. Mehta, E. Harmon, B. MacIntyre, and M. Mateas. Presence and engagement in an interactive drama. In *CHI '07: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 1475–1484, New York, NY, USA, 2007. ACM.

[10] T. Eichner, H. Prendinger, E. André, and M. Ishizuka. Attentive presentation agents. In *Intelligent Virtual Agents (IVA 2007)*, pages 283–295, 2007.

[11] P. Ekman and W. Friesen. *Unmasking the Face.* Prentice Hall, 1975.

[12] Facial Expression Repertoire. Filmakademie Baden-Württemberg.

[13] G. Flaubert. *La revue de Paris.* France, 1856.

[14] A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita. Messages embedded in gaze of interface agents — impression management with agent's gaze. In *CHI '02: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 41–48, New York, NY, USA, 2002. ACM Press.

[15] M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *CHI '01: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 309–316. ACM Press, 2001.

[16] E. T. Hall. A system for notation of proxemic behavior. *American Anthropologist*, 65:1003–1026, 1963.

[17] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[18] M. Kipp and P. Gebhard. IGaze: Studying Reactive Gaze Behavior in Semi-immersive Human-Avatar Interactions. In *Intelligent Virtual Agents (IVA '08)*, pages 191–199, 2008.

[19] C. L. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986.

[20] S. P. Lee, J. B. Badler, and N. I. Badler. Eyes alive. *ACM Transactions on Graphics*, 21(3):637–644, 2002.

[21] L. P. Morency, M. C. Christoudias, and T. Darrell. Recognizing gaze aversion gestures in embodied conversational discourse. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 287–294, New York, NY, USA, 2006. ACM Press.

[22] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *ACL '03: Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, pages 553–561. Association for Computational Linguistics, 2003.

[23] Y. I. Nakano and Y. Yamaoka. Information state based multimodal dialogue management: Estimating conversational engagement from gaze information. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, editors, *Intelligent Virtual Agents, 9th International Conference, IVA 2009, Amsterdam, The Netherlands, September 14-16, 2009, Proceedings*, volume 5773 of *Lecture Notes in Computer Science*, pages 531–532. Springer, 2009.

[24] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. In *Intelligent Virtual Agents (IVA' 05)*, pages 229–240, London, UK, 2005. Springer-Verlag.

[25] M. Rehm and E. André. From chatterbots to natural interaction - face to face communication with embodied conversational agents. *IEICE Transactions on Information and Systems, Special Issue on Life-Like Agents and Communication*, 88-D(11):2445–2452, 2005.

[26] M. Sagar. Facial performance capture and expressive translation for king kong. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Sketches*, page 26. ACM, 2006.

[27] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *ETRA '00: Proc. of the symposium on Eye tracking research & applications*, pages 71–78. ACM Press, 2000.

[28] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. Where to look: a study of human-robot engagement. In *IUI '04: Proc. of the 9th international conference on Intelligent user interfaces*, pages 78–84. ACM Press, 2004.

[29] I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *CHI '90: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 3–10. ACM, 1990.

[30] W. Steptoe, R. Wolff, A. Murgia, E. Guimaraes, J. Rae, P. Sharkey, D. Roberts, and A. Steed. Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *CSCW '08: Proc. of the ACM 2008 conference on Computer supported cooperative work*, pages 197–200. ACM, 2008.

[31] Valve. Facial Expressions Primer from Half-Life 2. `http://developer.valvesoftware.com/wiki/Facial_Expressions_Primer`.

[32] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *CHI '01: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 301–308. ACM Press, 2001.

[33] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater. An Eye Gaze Model for Dyadic Interaction in an Immersive Virtual Environment: Practice and Experience. *Computer Graphics Forum*, 23(1):1–11, 2004.

[34] T. Vogt, E. André, and N. Bee. Emovoice - a framework for online recognition of emotions from voice. In *Proc. of Workshop on Perception and Interactive Technologies for Speech-Based Systems.* Springer, 2008.

[35] J. Wagner, E. André, and F. Jung. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction (ACII 2009)*, 2009.