ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Analyzing and Modeling Gaze during Face-to-Face Interaction

Stephan Raidt, Gérard Bailly & Frédéric Elisei

Dept. of Speech & Cognition, GIPSA-Lab, Grenoble Universities – France

Stephan.Raidt/Gerard.Bailly/Frederic.Elisei@gipsa-lab.inpg.fr

1 Abstract

We present here the analysis of multimodal data gathered during realistic face-to-face interaction of a target speaker with a number of interlocutors. During several dyadic dialogs videos and gaze have been monitored with an original experimental setup using coupled cameras and screens equipped with eye tracking capabilities. For a detailed analysis of gaze patterns we distinguish different regions of interest on the face. With the aim to understand the functions of gaze in social interaction and to develop a coherent gaze control model for our talking heads we investigate the influence of cognitive state and social role on the observed gaze behaviour.
Keywords: embodied conversational agents, face-to-face interaction, eye gaze, talking head, gaze model.

2 Introduction

When interacting with an ECA (Embodied Conversational Agent) we do have strong expectations concerning its appearance that we interpret and evaluate according to natural human behavior. The more human-like the appearance of the ECA the higher are the interlocutors' expectations towards human like behavior. This is true for the entire interaction and not only while the ECA actively communicates. We perceive the ECA not only when it speaks but also the way it displays attentiveness, listens to us or its way of managing turn taking. The comprehension of the dialog and the credibility of the delivered information may be degraded by incorrect control strategies, imprecise interaction loops or an impoverished multimodal implementation. In this context we consider gaze as an essential component of face-to-face interaction and proper gaze generation pivotal to maintenance of mutual attention and prosperous interaction.

Eyes are very special stimuli in a visual scene and humans are especially sensitive to the orientation of eyes [1, 16]. The gaze direction carries a huge diversity of information. It reveals the center of interest of a subject and may guide the attention of an interlocutor [17, 27]. Together with facial expression and context, it is interpreted to derive mental states of another person [5]. During interaction it is important to the organization of discourse such as beginning and ending of speech, turn taking, or accentuation of utterances [1, 15].

In previous studies [21, 22], we tested the capability of our animated talking head [2] to direct the attention of a human observer by the means of head and gaze orientation. The work described here presents quantitative measurements of gaze patterns recorded during dyadic face-to-face conversations in relation to the course of the dialog. We show that in live interaction these patterns differ from those obtained when scrutinizing pre-recorded videos. The results are exploited for a first basic version of a gaze generation model of our talking head.

3 Motivation

The aim of our research is to understand the functions of gaze in social interaction and to develop a coherent gaze control model for our talking heads taking into account the results of detailed multimodal scene analysis. Such a coupling between multimodal scene analysis and multimodal scene synthesis is rather usual for anthropoid robots at least for the planning of movement. But there are very few proposals of that kind in the field of speech communication that comprise speech and gesture generation as well as facial animation. Human-computer interaction is mainly focused on the dialog component and the exchange of symbolic information between information retrieval (speech recognition, etc.) and rendering (concept-to-speech synthesis, etc.). Signals gathered from a rich scene analysis are nevertheless crucial to provide a convincing sense of presence by linking virtual actions to the real world [7, 23].

Figure 1: *Experimental setup: In contrast to video phones this setup enables real size rendering of video image and eye contact, as the camera is placed on the screen. With additional audio transmission it is therefore very close to a scenario where interlocutors face each other across a table.*

Since the pioneer work of Argyle and Cook [1] and Kendon [14], few research has been conducted on gaze during face-to-face interaction. Vertegaal et al. [26] analyzed gaze behaviour of a subject involved in multi-party conversation. Gullberg [11] found differences in gaze patterns when involved in live interactions versus off-line video presentation. She concluded that these differences are partly due to social norms. Both studies only survey the gaze of one of the interacting subjects and thus are not able to take into account the gaze of the interlocutors. Furthermore the measurements are not as fine grained as an analysis distinguishing different parts of the face would afford. Bateson et al [24] however showed that the gaze of listeners alternates between eyes and mouth when scrutinizing off-line videos. The percentage of gaze towards eyes vs. mouth is influenced by the perception task. In return visual attention is also known to influence speech comprehension (see for example [19], for sensitivity of McGurk effect to visual attention).

Different approaches have been proposed to model and generate gaze patterns. Itti et al. [13] developed a gaze generation model coupled with a visual attention system that detects salient and pertinent points of interest in a natural scene and triggers exogenous saccades and fixations handled by a biological model of eye motion. Note that there is no special treatment or even detection of faces in this system. Bilvi and Pelachaud [8] implemented a model for gaze generation in dialogue. It takes text tagged with labels of communicative functions as input combined with a statistical model to generate eye movements alternating between direct and averted gaze. Note that no saccadic model is included and that alternation is paced by phone boundaries. Lee et al [18] uses a purely statistical approach. The statistics are based on analysis of a video recording of one subject during face-to-face interaction. Bilvi and Pelachaud and Lee et al both take into account the cognitive activity of the ECA (e.g. speaking vs. listening, etc).

4 Eye gaze in face-to-face interaction

4.1 Experimental Setup

In order to investigate the functions of eye gaze during close dyadic face-to-face interaction, we developed an experimental platform where two subjects can interact via a crossed camera–screen setup (see Figure 1). The experimental set-up should give interlocutors the impression to be facing each other across a table.

A pinhole camera (1.5 cm^3) placed at the center of a computer screen films the subject facing the screen and displays the video image on a second screen. The second screen is symmetrically equipped. During the recording period the video signals are crossed. Prior to the beginning of each recording session, the screens function as inversed mirrors and subjects see their own video image to be able to adjust their position. In order to optimize gaze contact, subjects adjust their rest position in such a way that the middle of their eyebrows coincides with the position of the camera on the screen. A camera located on top of (or below) the screen would generate the impression of seeing the other subject from above (resp. below). This would make direct eye contact impossible [9].

The audio signals are exchanged via microphones and earphones. Video and audio signals as well as gaze directions are recorded during the interaction. For this purpose we use computer screens by Tobii Technology ® with embedded eye trackers. Before the recording a calibration phase writes a synchronization time stamp to the data streams. Gaze patterns can thus be aligned with audiovisual data for off-line analysis. This particular setting (mediated interaction, 2D displays, non intrusive eye tracking) limits the working space but is fully compatible with our target application i.e. a virtual ECA displayed on a screen and able to interact face-to-face with a human interlocutor. We will make our ECA imitate the interaction strategies emerging from the analysis of human dyadic conversations. Comparative evaluation can then be performed.

According to our knowledge this is the first experimental setup that monitors both subjects during such a mediated face-to-face interaction.

4.2 Experiment

Scenario: The experiment involves two subjects into a sentence-repeating game. One subject (initiator) reads and utters a sentence that the other subject (respondent) should repeat subsequently in a single attempt. The initiator is advised to face the screen when uttering a sentence. Roles (initiator and respondent) are further exchanged. The subjects are told that their performance (number of correctly repeated sentences) will be evaluated at the end. Semantically Unpredictable Sentences (SUS) [6] are used to force the respondent to be highly attentive to the audiovisual signal.

With this rather restricted scenario of interaction we try to isolate the main elements of face-to-face interaction and to enhance the aspect of mutual attention. It imposes a clear chaining of turns and roles (reading, speaking, listening and repeating) that avoids complex negotiation of turn taking and eases state dependent gaze analysis.

Subjects: We study inter- and intra-subject variability. In each dyad, we have a permanent target interlocutor, i.e. the female researcher that served as the target speaker for our talking head. She interacts with female subjects of the same social status and cultural background (French, European, and researcher) in order to control social relationship.

Sessions: The advocated purpose of the experiment is the study of the importance of visual feedback for telecommunication. Each session consists in fact of an on-line interaction using the full experimental setup followed by a faked interaction where the subjects are confronted to a previously recorded stimulus. It is taken from an interaction of our target speaker with the main experimenter and is the same for all subjects. The subjects should not realize that parts of the stimulus are pre-recorded. The noticeable impoverished feedback is justified by an absence of video feedback.

Each subject faces thus three tasks of ten sentences each: (1) repeating SUS given on-line by the target speaker; (2) uttering SUS and checking the correct repetition by the target speaker; (3) repeating SUS given off-line by the target speaker.

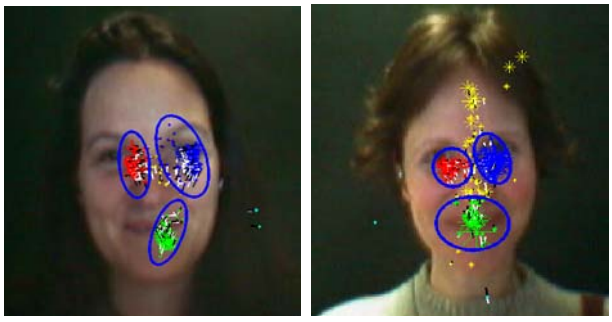

Figure 2: *Fixations of a whole session projected onto a reference image. With the ellipses the experimenter defines regions of interest that the fixations are assigned to. We distinguish between mouth, eyes, face and else. The size of the asterisks is proportional to the duration of fixations. Left: fixations of our target speaker on one subject. Right: the interlocutor's data.*

4.3 Data processing & labeling

Thanks to an original synchronization procedure we developed, the multimodal data (audio, video and gaze data of both subjects) are time-aligned. Multimodal scores are then computed, automatically labelled and downloaded into the ELAN® system [12] for subsequent additions and corrections.

The raw gaze data are processed to characterize fixations. Fixations are identified using a dispersion-based algorithm. An affine transform is applied to compensate for head movements determined by a robust feature point tracker. This permits to represent all fixations on a single reference image (see Figure 2). Elliptic regions are then defined by the experimenter on this reference image assigning the fixations to different regions of interest (ROI): left or right eye, mouth, face (other parts than the three preceding ones such as nose) or else (when a fixation hits other parts of the screen). Successive saccades can be labeled with the same ROI. Corrective saccades are very characteristic of attention (for example during reading [25]). The accumulations of fixations assigned to the ROI may be displaced relative to the targets as they can be identified on the reference image and only overlap partly (see Figure 2). This discrepancy between measured fixation target and obviously intended ROI might be explained by imprecision of eye tracking. It might also be sufficient for the subjects to focalize close to the ROI and exploit peripheral view.

The speech data is aligned with the phonetic transcriptions of SUS sentences and sessions are further segmented into sequences assigned to six different cognitive states (CS): pre-phonation, speaking, listening, reading, waiting and thinking, also distinguishing role.

The role indicates which subject dominates the interaction. Differences are for example expected to occur during listening. The SUS that the initiator utters are unknown to the respondent whereas the initiator already knows the content of the SUS and might therefore be less attentive when the respondent repeats them.

Some of the states depend on role: waiting is the CS of the respondent while the initiator is reading or the CS of the initiator after having uttered a sentence while waiting until the respondent begins to repeat the sentence. There are also syntactic dependencies between CS: pre-phonation preceding speaking is triggered by pre-phonatory gestures such as lip opening, speaking state triggers listening state for the interlocutor, etc. Some CS appear only in one of the two roles. The CS reading only occurs while a subject is initiator (reading next sentence to utter) and the CS thinking only occurs while a subject is respondent (preparing in mind the sentence to repeat).

We also label blinks. Our automatic detection of blinks is triggered by short gaps of invalid gaze data between fixations lasting between 20 and 240 ms.

After the manual correction of the labels of CS and blinks with ELAN® they are exported into Matlab® for statistical analysis.

5 Results

Up to now we recorded interactive sessions of our target subject with 4 interlocutor subjects. The results clearly confirm the triangular pattern of fixations scanning the eyes and the mouth previously obtained by Vatikiotis-Bateson, Eigsti et al [24] for perception of prerecorded audiovisual speech (see Figure 2). They also confirm our choice to distinguish cognitive state and role.

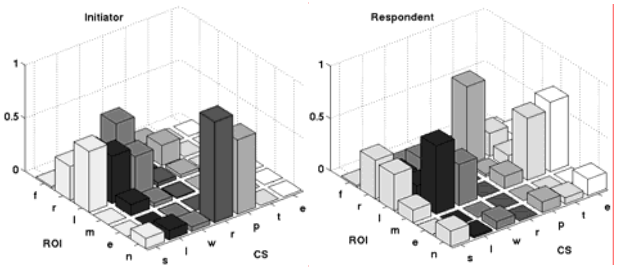

Figure 3: *Fixation profiles of all interactions of our target speaker over role, ROI (face, right eye, left eye, mouth, else) and cognitive state CS (speaking, listening, waiting, reading, pre-phonation, thinking, else). The bars represent the means of the percentage of fixation time on ROI during an instance of a cognitive state. The diagram is completed by bars (ROI named "n") representing the means of percentage of time when no fixations are detected.*

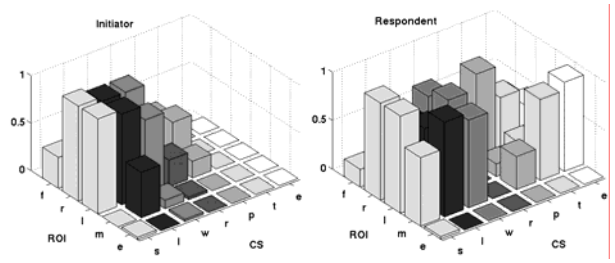

Figure 4: *Probabilities that fixations to ROI appear at least once during a cognitive state, calculated for all interactions of our target speaker. Bars of the same gray scale value constitute the fixation profiles of the different CS.*

5.1 Fixations and cognitive states

We define fixation profiles on a number of time intervals as the percentual distribution of fixations among the ROIs within this time span. For statistical analysis we only consider 4 ROIs: left eye, right eye, mouth and face. The ROI 'else' is disregarded since it almost never occurs during the analyzed sessions. We investigate the influence of the two factors role and cognitive state on the mean fixation profiles calculated for our target subject during the four interactions (see Figure 3). This means about 80 measurement samples of our target subject for each CS, of which 40 for each role. We compare the multivariate means of the fixation profiles of the CS (pre-phonation, speaking, listening, and waiting) that occur in both roles using MANOVA.

MANOVA returns an estimate of the dimension 'd' of the space containing the multivariate group means and p-values to indicate the significance of each dimension. We found that independent of role CS-specific profiles are significantly

different from each other (d=3, p=0; 0; 0.03). Separating the data for role this is even more significant for the role initiator: d=3, p=0; 0; 0.0007 but less significant for the role respondent: d=2, p=0; 0; 0.12). All pair wise comparisons of CS are also significant.

To better be able to interpret the mean fixation profile we also computed the probability of occurrence of at least one fixation in a given ROI for each instance of a given cognitive state as displayed in Figure 4. A fixation to a specific ROI lasting the whole duration of every second instance of a specific CS would for example result in the same value of 50% in the fixation profile as a fixation lasting half of the duration of each instance of that CS. Considering a model for gaze control this means however an important difference that can be identified via the probability of occurrence.

To verify the impact of live feedback, we compared the fixation profiles measured during the online interaction (when the interlocutors of our target subject were acting as respondents) with the measurements during the faked interaction (using the pre-recorded stimulus). MANOVA showed for each interlocutor that independent of CS mean profiles of online and faked interaction are significantly different. When comparing live versus faked interaction separately by cognitive state, we found inter-subject differences. While one subject shows no difference in the direct comparison of CS at all, another subject has different gaze patterns for both listening and speaking (p=0.01; p=0.02) while the two others have only one significantly different CS, respectively speaking (p=0.02) and waiting (p=0.03) (see Figure 5).

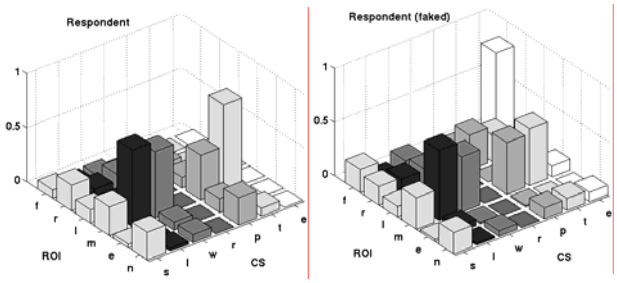

Figure 5: *Example of fixation profiles of the same subject acting as respondent during on-line interaction (left) vs. faked interaction (right).*

5.2 Comments

First of all, these results confirm the eyes and mouth as dominant target zones [24]. We have shown that role has a significant impact on fixation profiles. When listening, respondents should for instance gaze towards the mouth to be able to decode articulation, while the initiators do not need to benefit from audiovisual speech perception since to them the content of the message is already known.

The segmentation of the discourse into cognitive states explains a large part of the variability of the gaze behavior of our reference subject. The high significance of these predictors justifies *a posteriori* our choice of cognitive states.

The comparison of the fixation profiles of online versus faked interaction indicates that faked interaction has an impact on gaze behavior even if gaze patterns of the interlocutor are natural (since videos are extracted from the recording of a real online interaction). This is largely subconscious: note that only one of the subjects actually realized that the second stimulus was prerecorded. We interpret this as an argument that a generic gaze model should not only use rich and pertinent internal states but also profit from scene analysis.

Figure 4 shows interesting variation between the probabilities of occurrence of fixations to the different ROIs. Given a role and a cognitive state, we can be almost certain that our reference subject will fixate certain regions of the face of her interlocutor but that she will never gaze at others. For instance she never fixates the mouth when speaking as initiator but always the left and right eye. Note also the strong tendency to look especially at the right eye of her interlocutor when preparing to speak as respondent and the important rise of probability to fixate the mouth when respondent.

The statistics of the fixation profile given above just consider the time spent on a given ROI on a frame basis. Fixations span of course across boundaries between cognitive states. In order to build an effective statistical model of saccade generation (cf. §5.4), we also characterize the duration of fixations over region of interest and role. Statistical analysis with an ANOVA gives no reason to distinguish for role. The influence of ROI however is highly significant (df=9, F=18.84, p=0). Post hoc analysis shows that there is no difference of fixation duration between the ROI right eye and left eye. In comparison to them fixations to the face are significantly shorter and fixations to the mouth significantly longer.

5.3 Blinks and cognitive states

Blinking is a very important biological movement. Blinking helps the eyelids to spread a tear film over the eye. But blinks are also generated to protect the eye (blink reflex) and have been shown to co-occur with triggering of large gaze shifts and head movements (gaze-evoked blinks described by [10]). Most ECA generate blinks with a simple random event generator. Our data evidence however that blinking rate is highly dependent on cognitive state (see also [20]).

Assuming that blinks occur randomly or at a regular frequency their amount should be proportional to time. To test this hypothesis we used Chi-square goodness-of-fit test to compare the observed amount of blinks to the amount expected. Therefore we considered the added-up duration of the cognitive states and the mean frequency over the entire duration of the interaction. The cognitive states taken into account are pre-phonation, speaking, listening, and waiting. For all subjects this hypothesis was rejected at $p < 0.01$.

A detailed analysis of the influence of CS on blink rate showed that ‘speaking’ accelerates blink rate, whereas ‘reading’ and ‘listening’ slow it down or inhibit blinks. Especial in the role of respondent ‘listening’ seems to have a strong tendency to inhibit blinks. Strikingly often blinks occur at the change-over from reading to speaking (pre-phonation). This might be explained by the fact that the subjects wet the eyes in perspective of a long period of mutual attention. An alternative explanation is the linkage of blinking and major saccadic gaze shifts proposed by Evinger et al [10].

5.4 Modeling

We built a first gaze control model for our talking head (see Figure 8) by training and chaining role- and CS-specific Hidden Markov Models (HMM). Given a succession of CS with associated durations it computes parameters describing the fixations of the ECA towards the various ROI on the face of its interlocutor. HMM states equal to the different ROI and observations equal to the durations of fixations.

The transition probabilities of the HMM are computed from the transition matrix between the different ROI within a given CS and role as observed during the experiment. Figure 6 shows the transition matrices observed during the CS 'speaking' and 'listening' for the two roles instructor and respondent. The colons and lines represent respectively the current and subsequent possible fixation targets. The grey level is proportional to the transition probability between pairs: high probabilities are coded by dark gray. The amount 'n' of accumulated fixations to a target, denoted on bottom of each colon, indicates the number of items on which this estimation is based and thus gives the reliability of the probabilities displayed in the matrix colons. Fixations to the face for instance are very rare and thus the transition probabilities from face to other ROI not very reliable since they are calculated from only few observations.

An initial state in each HMM has been added to cope with the particular distribution of the first fixation. The observation probabilities determine the duration of the fixation emitted by the HMM at each transition. The probability density functions of these durations are computed from fixations gathered from the interactions. Fixations to the mouth are for instance longer than fixations to the eyes.

Based on these parameters we use the same generation process as proposed by Lee [18] to control the gaze of the clone of our target speaker (cf. Figure 8).

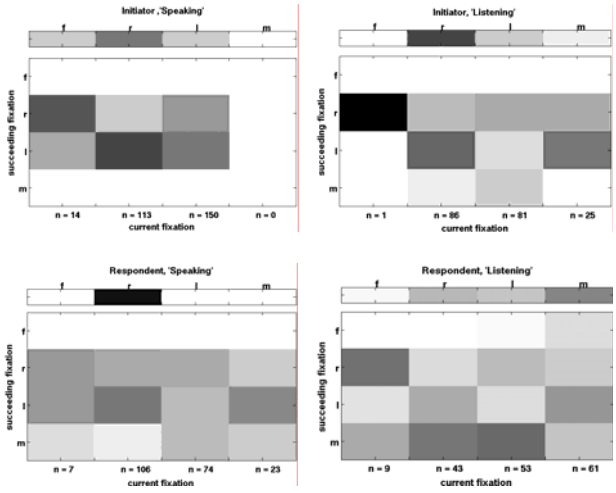

Figure 6: *Distribution of probability over ROI of first fixation of a cognitive state and transition matrix inside the cognitive state for speaking and listening in both roles and over entire interaction for all recording sessions of our target speaker.*

Until now we have not yet evaluated the model experimentally but the distributions of fixations according to ROI and cognitive state obtained with this gaze control model are – as expected – very similar to the distributions observed during live face-to-face interactions (see Figure 7).

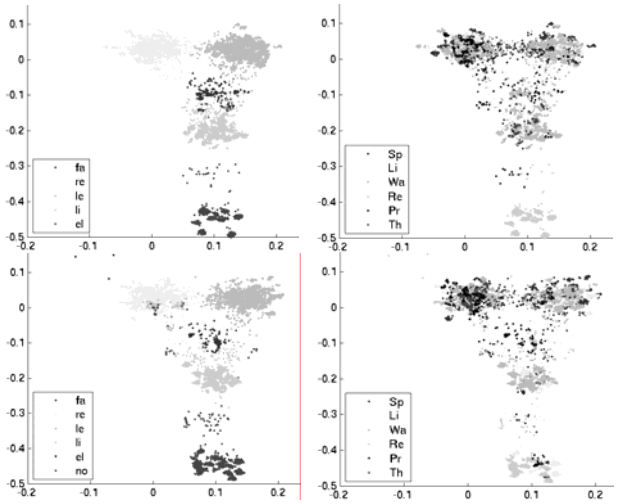

Figure 7: *Comparing original gaze patterns (top) with patterns generated using the statistical model driven by the sequence of cognitive states (bottom). Left: fixations labeled with ROI; right: fixations labeled with cognitive state.*

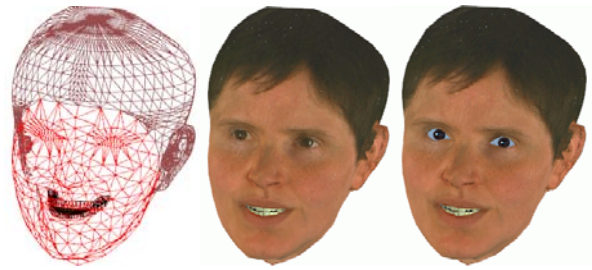

Figure 8: *Our virtual talking face is driven by 12 facial degrees-of-freedom [3]. The eyes and eyelids movements are controlled by 5 degrees-of-freedom that captures the correlations between gaze and eyelids deformations [4].*

6 Conclusions and Perspectives

Gaze and blinking are essential visible cues for sense of presence of ECA (Embodied Conversational Agent). They are important cues for signaling the ECA's awareness of its environment, the cognitive and emotional state of its interlocutors as well as its own states. We have shown that for the generation of realistic gaze direction the control model should at least be aware of its own cognitive states and its role in the interaction. We have settled a basis for a state-aware eye-gaze generator for controlling the eye movements of a virtual ECA based on these findings.

In order to develop an improved gaze generator we should isolate the significant events detected in the multimodal scene that impact the closed-loop control of gaze. We should notably investigate the influence of eye saccades produced by the interlocutor as potential extrinsic driving events of gaze. We expect for example to find patterns of gaze avoidance after periods of eye contact.

Furthermore we should implement other cognitive and emotional states as well as other functions of gaze (deictic or iconic gestures). This supposes a more elaborated and richer multimodal scene analysis as well as a deeper comprehension of the intentions of the interlocutors.

7 Acknowledgments

This work is supported by the project Presence of the cluster ISLE financed by the Rhône-Alpes region.

8 References

[1] Argyle, M. and M. Cook, *Gaze and mutual gaze.* 1976, London: Cambridge University Press.

[2] Badin, P., et al., *Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images.* Journal of Phonetics, 2002. **30**(3): p. 533-553.

[3] Bailly, G., et al. *Degrees of freedom of facial movements in face-to-face conversational speech.* in *International Workshop on Multimodal Corpora*. 2006. Genoa - Italy.

[4] Bailly, G., et al. *Embodied conversational agents : computing and rendering realistic gaze patterns.* in *Pacific Rim Conference on Multimedia Processing*. 2006. Hangzhou.

[5] Baron-Cohen, S., D.A. Baldwin, and M. Crowson, *Do children with autism use the speaker's direction of gaze strategy to crack the code of language?* Child Development, 1997. **68**(1): p. 48–57.

[6] Benoît, C., M. Grice, and V. Hazan, *The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences.* Speech Communication, 1996. **18**: p. 381-392.

[7] Bickmore, T. and J. Cassell, *Social dialogue with embodied conversational agents*, in *Advances in natural, multimodal dialogue systems*, J. van Kuppevelt, L. Dybkjaer, and N. Bernsen, Editors. 2005, Kluwer Academic: New York.

[8] Bilvi, M. and C. Pelachaud. *Communicative and statistical eye gaze predictions.* in *International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 2003. Melbourne, Australia.

[9] Chen, M. *Leveraging the asymmetric sensitivity of eye contact for videoconference*. in *SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves.* 2002. Minneapolis, Minnesota.

[10] Evinger, C., et al., *Not looking while leaping: the linkage of blinking and saccadic gaze shifts.* Experimental Brain Research, 1994. **100**: p. 337-344.

[11] Gullberg, M. and K. Holmqvist. *Visual attention towards gestures in face-to-face interaction vs on screen.* in *International Gesture Workshop*. 2001. London, UK.

[12] Hellwig, B. and D. Uytvanck, *EUDICO Linguistic Annotator (ELAN) Version 2.0.2 manual*. 2004, Max Planck Institute for Psycholinguistics: Nijmegen - NL.

[13] Itti, L., N. Dhavale, and F. Pighin. *Realistic avatar eye and head animation using a neurobiological model of visual attention*. in *SPIE 48th Annual International Symposium on Optical Science and Technology*. 2003. San Diego, CA.

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

[15] Kendon, A., *Does gesture communicate? A Review.* Research on Language and Social Interaction, 1994. **2**(3): p. 175-200.

[16] Langton, S. and V. Bruce, *Reflexive visual orienting in response to the social attention of others.* Visual Cognition, 1999. **6**(5): p. 541-567.

[17] Langton, S., J. Watt, and V. Bruce, *Do the eyes have it ? Cues to the direction of social attention.* Trends in Cognitive Sciences, 2000. **4**(2): p. 50-59.

[18] Lee, S.P., J.B. Badler, and N. Badler, *Eyes alive.* ACM Transaction on Graphics, 2002. **21**(3): p. 637-644.

[19] Paré, M., et al., *Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect.* Perception and Psychophysics, 2003. **65**: p. 553-567.

[20] Peters, C. and C. O'Sullivan. *Attention-driven eye gaze and blinking for virtual humans.* in *Siggraph*. 2003. San Diego , CA.

[21] Raidt, S., G. Bailly, and F. Elisei. *Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points?* in *Language Ressources and Evaluation Conference (LREC)*. 2006. Genova - Italy.

[22] Raidt, S., F. Elisei, and G. Bailly. *Face-to-face interaction with a conversationnal agent: eye-gaze and deixis.* in *International Conference on Autonomous Agents and Multiagent Systems*. 2005. Utrecht University, The Netherlands.

[23] Thórisson, K., *Natural turn-taking needs no manual: computational theory and model from perception to action*, in *Multimodality in language and speech systems*, B. Granström, D. House, and I. Karlsson, Editors. 2002, Kluwer Academic: Dordrecht, The Netherlands. p. 173–207.

[24] Vatikiotis-Bateson, E., et al., *Eye movement of perceivers during audiovisual speech perception.* Perception & Psychophysics, 1998. **60**: p. 926-940.

[25] Vergilino-Perez, D., T. Collins, and K. Dore-Mazars, *Decision and metrics of refixations in reading isolated words.* Vision research, 2004. **44**(17): p. 2009-2017.

[26] Vertegaal, R., et al. *Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes.* in *Conference on Human Factors in Computing Systems*. 2001. Seattle, USA: ACM Press New York, NY, USA.

[27] Yarbus, A.L., *Eye movements during perception of complex objects*, in *Eye Movements and Vision*, L.A. Riggs, Editor. 1967, Plenum Press: New York. p. 171-196.