Gaze Patterns 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

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. Videos and gaze of both interlocutors were monitored with an experimental setup using coupled cameras and screens equipped with eye trackers. With the aim to understand the functions of gaze in social interaction and to develop a gaze control model for our talking heads we investigate the influence of cognitive state and social role on the observed gaze behaviou*r.

1. Introduction

When interacting with video-realistic ECA (Embodied Conversational Agent) we do have strong expectations concerning its actions that are interpreted and evaluated with reference to an expected human behavior. This is true for the entire interaction and not only when the ECA communicates: the way he pays attention to us, listens to us or takes the turn matters. The comprehension of the dialog and the credibility of the delivered information may be degraded by incorrect control strategies, imprecise interaction loops or impoverished multimodal behavior.

Eyes are very special stimuli in a visual scene and humans are especially sensitive to the orientation of eyes [1, 9]. 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. Together with facial expression and context, it is used to derive mental states of another person [2]. During interaction it is important to the organization of discourse such as beginning and ending of speech, turn taking, or accentuation of utterances [1, 8].

The work described here presents quantitative measurements of mutual gaze patterns recorded during dyadic face-to-face conversations in relation to the course of the dialog. We show that fixations and blinks depend strongly on cognitive state and role of the speaker in the conversation. These results are exploited for a first version of a gaze generation model of our talking head.

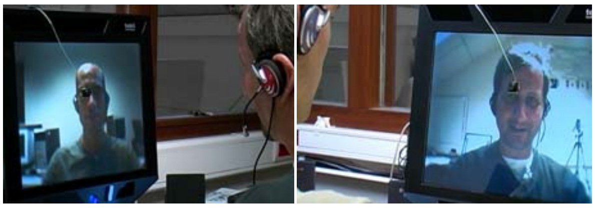

Figure 1: Experimental setup: In contrast to video phones this setup enables real size rendering of video image and eye contact, as the camera is placed on the screen. With additional audio transmission it is therefore very close to a scenario where interlocutors face each other across a table.

2. Eye gaze in face-to-face interaction

Due to limited space, we will only mention some key works related to our modeling framework. This work is original for two main reasons: (a) it provides precise multimodal data (blinks, saccades, etc.) for both interlocutors and (b) it evidences the impact of role of the speaker in the conversation. Bilvi and Pelachaud [4] propose a gaze model for dyadic conversations. Textual input to the system augmented with tags indicating communicative functions drives a statistical model to generate eye movements alternating between direct and averted gaze. Lee et al [10] propose also a similar statistical model based on analysis of video recordings of monologues uttered by one subject. Both models take into account the cognitive activity of the ECA (e.g. speaking vs. listening, etc) but do not integrate any detailed scene analysis: they do not determine exactly at which part of the face their target speakers are looking. Note finally that most models of visual attention (such as developed by Itti et al. [7]) do not include any special treatment of faces.

2.1 Experimental Setup

In order to investigate the patterns 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). It should give them the impression to be facing each other across a table. A small pinhole camera placed at the center of a computer screen films the subject facing the screen. The video image is displayed on the screen facing the interlocutor, which is equipped symmetrically. Prior to each recording session, the screens function as inversed mirrors so that subjects see their own video image in order to adjust their rest position. We determined that eye contact is optimal when the middle of the eyebrows of the video image coincides with the position of the camera on the screen. A camera located above (vs. below) the screen would generate the impression of seeing the interlocutor from above (vs. below). This would make direct eye contact impossible [5].

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. At the beginning of the recording a calibration phase writes a synchronization time stamp to the data streams. This particular setting (mediated interaction, 2D displays, non intrusive eye tracking) limits the working space but is fully compatible with our target application of an interactive ECA displayed on a screen.

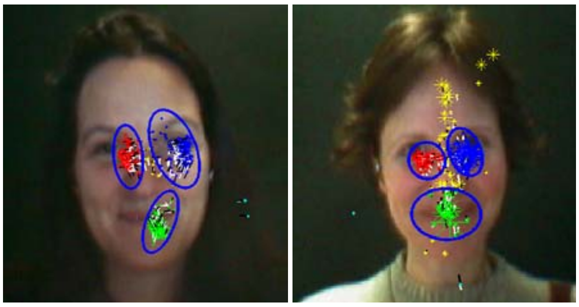

Figure 2: Fixations of a whole session projected onto a reference image. The ellipses define regions of interest that the fixations are assigned to (mouth, eyes, face, else). The size of the asterisks is proportional to the duration of fixations. Left: fixations of our target speaker on one subject. Right: the interlocutor’s data.

2.2 Experiment

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 immediately in a single attempt. The initiator is instructed to face the screen when uttering a sentence. Roles of initiator and respondent are further exchanged. Semantically Unpredictable Sentences (SUS) [3] 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 elicit mutual attention. The scenario imposes a clear chaining of cognitive states and roles that avoids complex negotiation of turn taking and eases state-dependent gaze analysis. We study inter- and intra-subject variability. In each dyad we have a reference target interlocutor (model of our talking head), who interacts with subjects of the same social status, cultural background and sex (French female senior researcher). Each session consists 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. They should not realize that parts of the stimulus are pre-recorded. 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.

2.3 Data processing & labeling

Fixations are identified in the raw gaze data using a dispersion-based algorithm [12]. An affine transform is applied to compensate for head movements determined by a robust feature point tracker. All fixations are projected back on a reference head position (see Figure 2). Elliptic regions are defined by the experimenter on this reference image to assign 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).

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.

We also distinguish role (initiator vs. respondent). Differences are for example expected to occur during listening. When listening to the respondent, the initiator already knows the content of the SUS he just have pronounced and might therefore not need to lip read. Note also that some 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. Most ECA generate blinks with a simple random event generator. We will however show that blinking frequency is highly modulated by CS and role.

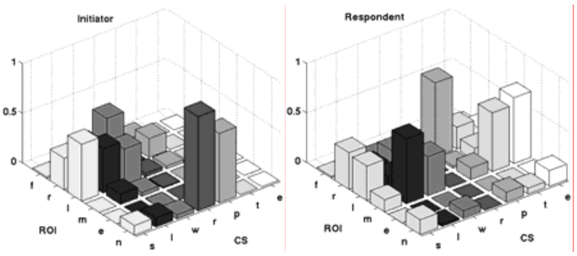

Figure 3**:** Fixation profiles of all interactions of our target speaker over role (initiator, respondent), ROI (face, right eye, left eye, mouth, else) and cognitive state CS (speaking, listening, waiting, reading, pre-phonation, thinking, else). The bars represent the means of the percentage of fixation 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.

3. Results

Up to now we recorded interactive sessions of our target subject with 9 interlocutors. The results clearly confirm the triangular pattern of fixations scanning the eyes and the mouth previously obtained by Vatikiotis-Bateson, Eigsti et al [13] (see Figure 2). They also confirm our choice to distinguish cognitive state and role.

3.1 Fixations and cognitive states

We define fixation profiles as the relative distribution of fixations among the ROI within a given activity. 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 180 measurement samples of our target subject for each CS (90 for each role). Using MANOVA we compare the multivariate means of the fixation profiles of the CS (pre-phonation, speaking, listening, and waiting) that occur in both roles. 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.

We also characterized the duration of fixations over region of interest and role. 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 between duration of fixations of the ROI right eye and left eye. In comparison fixations to the face are significantly shorter and fixations to the mouth significantly longer.

To verify the impact of live feedback, we compared the fixation profiles measured during the online interaction with those of the faked interaction (using the pre-recorded stimulus). MANOVA showed that for each interlocutor mean profiles of online and faked interaction (computed discarding CS) 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).

3.2 Blinks and cognitive states

Our data evidence that blink rate is highly dependent on cognitive state ($p < 0.01$) [see also 11]. 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 even inhibit blinks. Particularly in the role of respondent the CS 'listening' strongly inhibits blinks. Strikingly often blinks occur at the change-over from reading to speaking (pre-phonation). This might be explained by the linkage of blinking and major saccadic gaze shifts proposed by Evinger et al [6].

3.3 Modeling

We built a gaze control model for our talking head 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. 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 [10] to control the gaze of the clone 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 very similar to the distributions observed during live face-to-face interactions.

4. Conclusions and Perspectives

These results confirm the eyes and mouth as dominant target zones [13]. We have shown that role has a significant impact on fixation profiles. When listening, respondents should for instance gaze towards the mouth to benefit from lip reading, while the initiators do not need to benefit from audiovisual speech perception since they already know the content of the message. The segmentation of the interaction into cognitive states explains a large part of the variability of the gaze behavior of our reference subject. The ECA should thus at least be aware of its own cognitive state and its role in the interaction.

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

Based on our findings, we have settled a basis for a state-aware eye-gaze generator of an ECA. 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. Furthermore other cognitive and emotional states as well as other functions of gaze (deictic or iconic gestures) should be implemented.

5. Acknowledgments

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

6. References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

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

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

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

[12] Salvucci, D.D. and J.H. Goldberg. *Identifying fixations and saccades in eye-tracking protocols*. in *Eye Tracking Research and Applications Symposium*. 2000. Palm Beach Gardens, FL.

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