Gaze, conversational agents and face-to-face communication

Gérard Bailly, Stephan Raidt, Frédéric Elisei

Gaze, conversational agents and face-to-face communication

Gérard Bailly, Stephan Raidt[1] & Frédéric Elisei

Speech & Cognition department, GIPSA-lab, UMR 5216 CNRS – Université de Grenoble, France

Contact: gerard.bailly@gipsa-lab.grenoble-inp.fr,

Abstract

In this paper, we describe two series of experiments that examine audiovisual face-to-face interaction between naive human viewers and either a human interlocutor or a virtual conversational agent. The main objective is to analyze the interplay between speech activity and mutual gaze patterns during mediated face-to-face interactions. We first quantify the impact of deictic gaze patterns of our agent. We further aim at refining our experimental knowledge on mutual gaze patterns during human face-to-face interaction by using new technological devices such as non invasive eye trackers and pinhole cameras, and at quantifying the impact of a selection of cognitive states and communicative functions on recorded gaze patterns.

Keywords

Conversational agents; face-to-face communication; gaze; eye fixations; blinks; deixis; mutual attention; games

1. INTRODUCTION

Building Embodied Conversational Agents (ECA) able to engage in convincing face-to-face conversation with a human partner is certainly one of the most challenging Turing tests one can imagine (Cassell, Sullivan et al. 2000). The challenge is far more complex than the experimental conditions of previous Loebner Prize competitions[2] where dialog is conducted via textual information: the ECA should not only convince the human partner that the linguistic and paralinguistic contents of the answers that are generated in response to human inquiries have been produced by a human intelligence, but also generate the proper multimodal signals that should fool human perception. We are however very close to being able to conduct such experiments (see for example the joint Interspeech/Loebner Speech & Intelligence Competition 2009). Automatic learning techniques that model perception/action loops at various levels of human-human interaction are surely key technologies for building convincing conversational agents. George, the talkative bot that won the Loebner Prize 2005, learned its conversation skills from the interactions it had with visitors to the Jabberwacky website, and through chats with its creator, Mr Carpenter. Similarly the first Turing test involving a non interactive virtual speaker (Geiger, Ezzat et al. 2003) has demonstrated that image-based facial animation techniques are able to generate and render convincing face and head movements.

Combining a pertinent dialog management capability with convincing videorealistic animation is still not sufficient to produce a real sense of presence (Riva, Davide et al. 2003). The sense of "being there" requires basic components of situated face-to-face communication such as mixed initiative, back channeling, turn taking management, etc. The interaction requires a detailed scene analysis and a control

1 Now with Carmeq GmbH, Carnotstr. 4 D-10587 Berlin

2 The Loebner Prize for artificial intelligence awards each year the computer program that delivers the most human-like responses to questions provided by a panel of judges over a computer terminal.

loop that knows about the rules of social interaction: the analysis and comprehension of an embodied interaction is deeply grounded in our senses and actuators and we have strong expectations about how action-perception loops are encoded by multimodal signals.
We describe here part of our efforts for designing virtual ECAs that are sensitive to the environment (virtual and real) in which they interact with human partners. We focus on the control of eye gaze and blinks. We describe the multiple scientific and technological challenges we face, the solutions that have been proposed in the literature and the ones we have implemented and tested.

2. EYE GAZE AND HUMAN-COMPUTER INTERACTION

2.1 Face-to-face interaction, attention and deixis

Eye gaze is an essential component of face-to-face interaction. Eyes constitute a very special stimulus in a visual scene. Gaze and eye-contact are important cues for the development of social activity and speech acquisition (Carpenter and Tomasello 2000). In conversation, gaze is involved in the regulation of turn taking, accentuation and organization of discourse (Kendon 1967; Argyle and Cook 1976). We are also very sensitive to the gaze of others when directed towards objects of interest within or even outside our field of view, notably when interpreting facial expressions (Pourtois, Sander et al. 2004). In the Posner cueing paradigm (Posner 1980; Posner and Peterson 1990), observers' performance in detecting a target is typically better in trials in which the target is presented at the location indicated by a former visual cue than in trials in which the target appears at an uncued location (Driver, Davis et al. 1999). Langton et al. (1999; 2000) have also shown that observers react more quickly when the cue is an oriented face than when it is an arrow or when the target itself changes.

Eye gaze is thus capable of attracting visual attention whereas visual features associated with the objects themselves such as highlighting or blinking attract less attention, unless they convey important information for the recognition of a scene. Perceptual salience is thus not the only determinant of interest. The cognitive demand of a task has a striking impact on the human audiovisual analysis of scenes and their interpretation. Yarbus (1967) showed notably that eye gaze patterns are influenced by the instructions given to the observer during the examination of pictures. Similarly Vatikiotis-Bateson et al. (1998) showed that perceivers' eye gaze patterns during audiovisual speech perception are influenced both by environmental conditions (audio signal-to-noise ratio) and by the recognition task (identification of phonetic segments vs. the sentence's modality). Buchan et al. (Buchan, Paré et al. 2007) replicated this experiment comparing judgments on speech and emotion.

2.2 Interacting with humanoids and avatars

The faculty of interpreting others' eye gaze patterns is thus crucial for humans and machines interacting with humans. For the "theory of mind" (TOM) as described by Baron-Cohen (Premack and Woodruff 1978), the perception of gaze direction is an important element of the set of abilities that allow an individual, based on the observation of the actions and behavior, to infer the hidden mental states of another. Several TOM characterizations have been proposed (Baron-Cohen, Leslie et al. 1985; Leslie 1994). Baron-Cohen proposes an Eye Direction Detector (EDD) and an Intentionality Detector (ID) as basic components of a Shared Attention Mechanism (SAM) that is essential to bootstrapping the TOM. The actual implementation of these modules requires the coordination of a large number of perceptual, sensorimotor, attentional, and cognitive processes.

Scassellati (2001) notably applied the "theory of mind" concept to humanoid robots developing an "embodied theory of mind" to link high-level cognitive skills to the low-level motor and perceptual

abilities of such a robot. The low-level motor abilities included for example coordinated eye, head and arm movements for pointing. The low-level perceptual abilities consisted in essentially detection of salient textures and motion for monitoring pointing and visual attention. This work still inspires much research on humanoid robots where complex behaviors emerge from interaction with the environment and users despite the simple tasks achieved by the robot such as expressing empathy for Kismet (Breazeal 2000) or following turn-taking for Robita (Matsusaka, Tojo et al. 2003; Fujie, Fukushima et al. 2005).

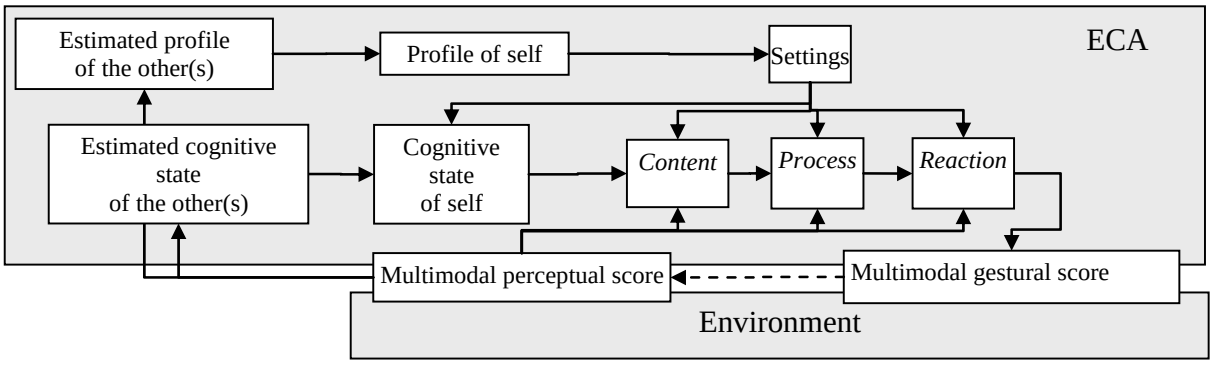


Figure 1: ECA-Human interaction scheme. The environment hosts human partner(s). The perception-action loops are regulated by three main layers: reactive, process and content (Thórisson 2002). The parameters of these modules are set according to the computed ECA profile that may vary according to social needs as put forward in the Communication Accommodation Theory (CAT, Giles and Clair 1979). Note that a mutual belief space is built by converging representations of estimated and actual profiles and cognitive states of the interlocutors. The profile can for example instruct the content module to approach or recede from the phonological space of the interlocutor.

3. INTERACTING WITH AN ECA

Most ECAs derive their "theory of mind" from high-level linguistic information gathered during the dialog. These virtual agents are generally not equipped with the means to derive meaning from the implicitly and explicitly communicational gestures of a human interlocutor and are also not generally equipped to generate such gestures for communication purposes. For example, in the study by Os et al (Os, Boves et al. 2005) an ECA is used as a virtual advisor. It gives advices, instructions and some back channel information but is not equipped with sophisticated scene analysis and computes his gesturing only on the basis of dialogic information. The failure of the ECA to gain users' satisfaction was attributed to the poor performance of individual components of the system (speech recognition, audiovisual synthesis, etc). Another possible way of improving users' commitment is via grounding.

Eye gaze of ECAs can be generated without grounding these gestures in the scene by simply reproducing statistical properties of saccadic eye movements (Lee, Badler et al. 2002; Peters, Pelachaud et al. 2005). Visual attention models have however been developed aiming at reproducing eye movement patterns of subjects viewing natural scenes. Notably, Itti et al. (2003) propose a model that couples physical scene analysis and control of eye gaze of a virtual ECA while preserving cognitive permeability of the analysis strategy thanks to the use of a so-called pertinence map. Our visual attention model (Picot, Bailly et al. 2007) incorporates the identification of potential objects of interest around

each fixation point: in particular a face recognition module has been added to elicit multimodal behaviour adapted to face-to-face communication.

In situations where context-aware face-to-face interaction is possible, an ECA should be able to give direct and indirect signs that it actually knows about *where* the interaction is taking place, *who* is its interlocutor and *what* service it may provide to the user considering the given environment. By signalling its ability to interpret human behavior, the system encourages the interlocutor to exhibit the appropriate natural activity. Such a complex face-to-face interaction requires intensive collaboration between an elaborate scene analysis and the specification of the task to be performed in order to generate appropriate and convincing actions of the ECA (see Figure 1).

Our perspective is to develop an embodied TOM for an ECA that will link high-level cognitive skills to the low-level motor and perceptual abilities and to demonstrate that such a TOM will provide the information system with enhanced user satisfaction and efficient interaction. The motor abilities are principally extended towards speech communication i.e. adapting content and speech style to pragmatic needs (e.g. confidentiality), speaker (notably age and possible communication handicaps) and environmental conditions (e.g. noise). If the use of a virtual talking head instead of a humanoid robot limits physical actions, it extends the domain of interaction to the virtual world. The user and the ECA can thus employ both physical and virtual objects – such as icons surrounding the virtual talking head – in their interaction.

The two experiments described here analyze gaze behaviour during live face-to-face dialogs between naive subjects and one reference interlocutor. Our research strategy is in fact to analyse and model the adaptive behaviour of a few reference subjects when interacting with various subjects and in various conditions: ECA control and embodiment are data-driven and aim at reproducing behaviour of their human alter egos. In experiment I, the reference interlocutor is an ECA: we evaluate here its ability to elicit mutual visual attention with its human interlocutors. This experiment focuses on deictic gestures attracting attention to common objects in the mutual shared visual field. Despite the rough control strategy of the gaze patterns, we show that subjects benefit significantly from these deictic gestures and – as for the case of priming by human gaze - can not avoid using them. Experiment II focuses on eye contact and mutual attention. The reference interlocutor here is the human speaker who donated his voice and appearance to the ECA so that control strategies can be fed back into the ECA, notably to better handle the interplay between deixis and mutual attention in future replications of experiment I.

Both experiments build on previous experimental work performed mainly with non interactive videos. They should be considered as a further step towards the automatic measurement and comprehensive modelling of multimodal behaviour of human and virtual agents during live face-to-face interactions.

4. EXPERIMENT I. ECA THAT ATTRACTS ATTENTION

The first experiment (Bailly, Elisei et al. 2005; Raidt, Bailly et al. 2006) involves a face-to-face interaction between an embodied conversational agent (ECA) and several other naïve human interlocutors. The ECA is embodied as a virtual talking head obtained by cloning one target human speaker named HL in the following. This female speaker is also the target speaker of Experiment II. We test here the impact of assistance given by the ECA on task-oriented interactions. A card game was designed during which the multimodal deictic gestures of the ECA aims at attracting attention of subjects towards regions of interest that are either appropriate or inappropriate for the given task. Despite the fact that subjects were explicitly instructed not to take ECA behaviors into account (as in Langton, Watt et al. 2000), we show that most of the subjects are drastically influenced by ECA hints and that the benefit in case of appropriate gestures – compared with inappropriate or no ECA gestures –

is quite large: 10% of the task duration in case of deictic head and gaze movements, up to 20% when these deictic gestures are accompanied with a spoken instruction.

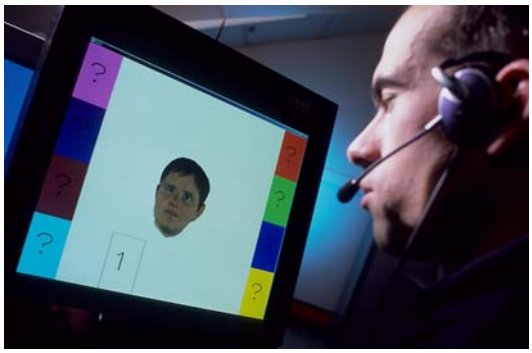


Figure 2 : Face-to-face interaction platform with an ECA (left) that provides deictic hints (see Figure 3) to a subject whose gaze is monitored. The hints help - or disturb - the subject in playing a virtual card game consisting of playing as fast as possible a card (white card here at the bottom of the screen) on one of eight possible positions (coloured boxes) according to different conditions.

4.1 Experimental Setup

4.1.1 Hardware

Sensors. The core element for the experiments described here is a Tobii 1750 eye-tracker[3] realized as a standard-looking flat screen that discretely embeds infrared lights and a camera. It allows us to detect, at up to 60Hz[4], the eye gaze of the user whose head can move and rotate freely in a fairly unrestricted 3D volume having the shape of a 40cm edge cube centered at 50cm away from the screen. To ensure the required accuracy of tracking each user must follow a short calibration procedure. During interaction with the system, the user sits in front of the eye-tracker where our 3D talking head faces him, as shown in Figure 2. Hardware and software allow the user to interact with the system using speech, eye gaze, mouse and keyboard. Additional data inputs (video camera, speech recognition, etc.) are available for other experimental setups.

Actuators. The visual representation of our ECA is implemented as the cloned 3D appearance and articulation gestures of a real human (Revéret, Bailly et al. 2000; Bailly, Bérar et al. 2003), (see Figure 2). The eye gaze can be controlled independently to look at 2D objects on the screen or spots in the real world outside the screen. Ocular vergence is controlled and provides a crucial cue for inferring spatial cognition. The virtual neck is also articulated and head movements accompany the eye-gaze movements. Standard graphic hardware with 3D acceleration allows real-time rendering of the talking head on the screen. The ECA also uses speech: audiovisual utterances can either be synthesized from text input or mimic pre-recorded human stimuli. We expect that the proper control of these capabilities will enable the ECA to maintain mutual attention - by appropriate eye saccades towards the user or his/her points of interest – and actively draw the user's attention.

[3] Please consult http://www.tobii.se/ for technical details.

[4] The eye-tracker delivers in fact asynchronous gaze data with time stamps at a maximum rate of 60Hz. Gaze data are only sent when estimated gaze displacements exceed a fixed threshold. We post-process date off-line for further analysis by eliminating outliners and up-sampling at a fixed frequency of 60Hz.

4.1.2 Software: scripting multimedia scenarios

The interaction during the virtual card game is described by an event-based language. This language allows the simple description and modification of multimedia scenarios. Compiled scenarios allow accurate recording of the multimodal input/output events and their timing.

In our event-based language a finite state machine (FSM) describes each scenario as a series of states with pre-conditions and post-actions. Each input device emits events according to user action and an internal model of the interaction space. Triggerable areas on the screen, such as selectable icons or areas of the talking head are defined and surveyed by the eye tracker. Each time the user looks at such a zone, the system posts new events, such as “entering zone” and “leaving zone” and may emit additional “zone fixation duration” events. The FSM is called each time an event is generated or updated.

As the user progresses in the scenario, the FSM specifies the events that can trigger entry into each state. Pre-conditions consist of conjunctions or successions of expected multimodal events as for instance recognized keywords, mouse clicks or displacements, eye movements or gaze directed to active objects. Each event is time-stamped. Pre-conditions can include tests on intervals between time-stamps of events. This allows, for example, association of speech items in terms of words that are identified as a sub product of speech recognition with a certain focus of attention. Post-actions typically consist of the generation of multimodal events. Time-stamps of these events can be used to delay their actual instantiation in the future (watchdogs for example). Post-actions can also generate phantom events, to simulate multimodal input, to share information or to trigger pre-conditions for following states of the FSM.

4.2 The interaction scenario

4.2.1 Design of the card game

To follow up on the findings of Langton and Driver about the special ability of human faces and eyes to direct attention, we designed an interaction scenario where an ECA should direct the user’s attention in a complex virtual scene. Our aim was to investigate the effect of deictic gestures on the user’s performance during a search and retrieval task. We chose an on-screen pair matching game, where the user is asked to locate the correct target position of a played card.

The card game consists in pairing cards. Eight target cards placed on the sides of the screen are turned over (with visible face up) once the played card initially positioned at the lower middle of the screen is selected with a mouse click (see Figure 3). The target cards are numbered from 1 to 8 and the play card has a random number between 1 and 8. The play card has then to be laid over one of the target cards with the same digit. To avoid memory effects, the target cards are shuffled before each turn. The target position is thus chosen randomly but uniformly distributed amongst the eight possibilities provided that the number of cycles is a multiple of eight. This should compensate for possible influences of the different positions on the users’ performance. The background color of the cards is fixed for each position.

4.2.2 Experimental conditions

In absence of assistance from the ECA, users have to explore most of the possible target cards to find the good match. A condition with no ECA is thus used as reference. When the ECA is displayed, we

instruct the subjects that the deictic behavior of the ECA is unpredictable and that they should not pay attention to it as in Langton et al (2000).

We tested different experimental conditions corresponding to different levels of assistance and help by the ECA that is displayed in the center of the screen when present. Screenshots of the game interface are given in Figure 3. The ECA can utter spoken commands and indicate directions with a rapid eye saccade combined with a slower head turn.

General information explaining the task is given as text on the screen at the beginning of the experiment. The user is instructed to play the chosen card on the target position as fast as possible but no strategy is suggested.

Each experimental condition has 24 measurement cycles. It is preceded by three training cycles for the subjects to become accustomed with the task and experimental conditions. The characteristics of the upcoming condition are described as text informing the user about the expected gaze behavior of the clone.

4.2.3 Data acquisition

For the evaluation of the experiments, the time-to-complete (TC) and the gaze behavior have been monitored. TC was measured as the time span between the first mouse click on the played card and the click on the correct target position. As the card game was displayed on the monitor with embedded eye-tracking, the visual focus of the user on the screen was recorded. We thus computed which objects on the screen were looked at and how much time users spent on them. Possible objects were the eight cards on the sides and the face of the ECA. Eye gaze towards the played card was not monitored.

At the end of the experiment, which lasted about 15 minutes, participants were asked to answer a questionnaire. They notably had to indicate which condition they considered as the easiest to perform and which condition seemed the fastest.

4.2.4 Data processing

Time-to-complete. Before evaluating TC, extreme outliers (distances from median > 5 times inter quartile range) were detected and deleted from the remaining valid data. Such outliers may be due to the fact that a two screen setup was chosen for the experiment. The mouse pointer may leave the screen on one side to appear on the other screen. This happened occasionally when users overshot the target card which made them loose time while moving the mouse pointer back into view. The distribution of TC is log-normal. We thus analyse the logarithms of TC within each experiment.

Number of cards inspected. The number of possible target positions looked at while searching for the correct target position was determined to analyse the search strategy of the subjects. In order to verify the reliability of the data collected by the eye tracker, the percentage of time reliably monitored by the eye tracker during the duration of the search of a respective cycle was calculated. A threshold of 95% was chosen to consider blinks and classify measured data as valid. If this requirement was not fulfilled, the data of this cycle was just discarded. If less than 60% of all cycles of a condition were not valid, the data of this condition were entirely discarded. We characterized the log-normal distribution of the number of cards inspected during a game. ANOVA and T-test analysis was then performed on valid data and significant differences between pairs of distributions are indicated in figures with stars.

4.3 Run A: does our clone have cues to direct social attention?

This first run aimed at evaluating the capacity of our ECA to attract users' attention using facial cues and quantifying the impact of good and bad hints on the users' performance. This work builds on the psychophysical experiments on visual priming of Langton et al. (Langton and Bruce 1999; Langton, Watt et al. 2000) using photographs of faces looking at different parts of the screen. In Langton's experiments, the impact on reaction time was more effective for the vertical dimension and remained small (10ms). We show here that these findings are confirmed by more realistic conditions and that much larger benefits can be expected.

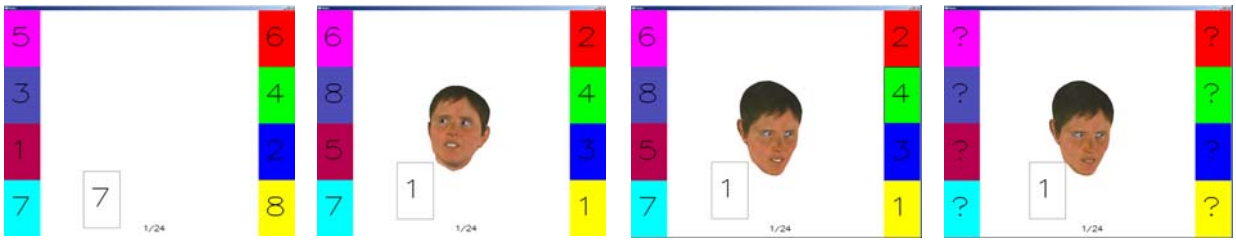
4.3.1 Conditions

The first series of experiments consisted of four different conditions, screenshots of which are displayed in Figure 3. For condition 1, no ECA was displayed. For condition 2, the ECA was visible and provided bad hints: it indicated randomly one of the non-matching positions with a facial gesture as soon as the user selected the played card. In condition 3, it indicated the correct target position. In condition 4, cards remained upside down and the correct visual cues provided by the ECA were the only ones to find the correct target position without a try-and-error strategy.

In all conditions where the ECA is displayed it encouraged the user with randomly chosen utterances alternating between motivation and congratulation after each turn. The utterances were generated off-line to avoid computation delays.

We had strong predictions about the data to be collected. Corresponding to the design of the experiment we expected a negative influence on subjects' performance when the clone provided misleading cues and a positive influence when it provided good hints. The condition where no clone was displayed served as a reference. From the fourth condition, we expected to measure the precision with which the gaze direction of the ECA could be perceived.

Ten users (six male and four female) participated in the first series of experiments. Participants ranged from 23 to 33 years of age and most were students. All regularly used a computer mouse and none reported vision problems. The dominant eye was the right eye for all but one subject. Each user had to play the game with the four experimental conditions as described above. Order of conditions was counterbalanced among subjects.



Cond. 1: no clone Cond. 2: wrong cues Cond. 3: correct cues Cond. 4: numbers hidden

Figure 3: Experimental conditions: The experiment was divided into four conditions with different levels of help and guidance by the clone.

* ($p<10^{-3}$)

* ($p<10^{-4}$)

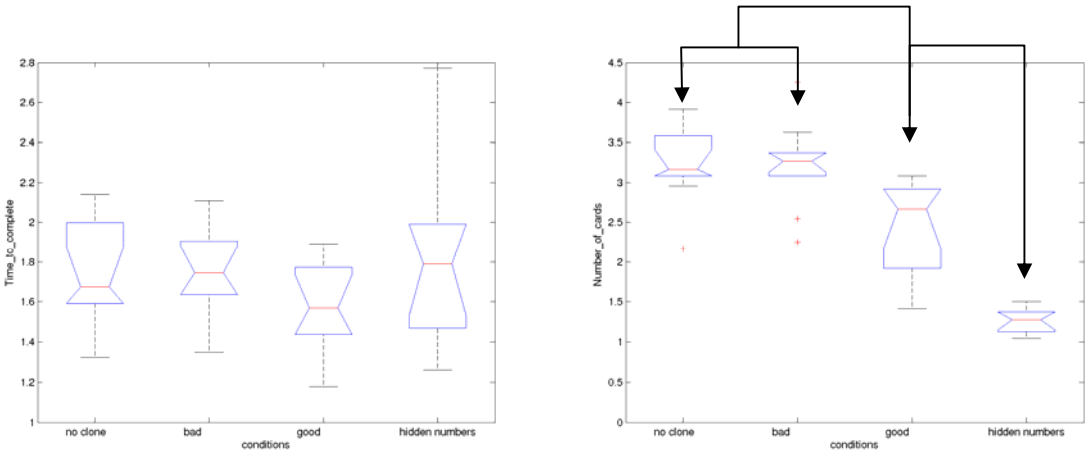


Figure 4: *Mean TC (left) and number of cards inspected (right) for our four conditions and all subjects pooled. Each bar represents 10 x 24 = 240 card pairings. An ANOVA analysis distinguishes between 3 groups of conditions with highly significant differences for number of cards inspected ($p<0.05$).*

4.3.2 Results

Errors. Most errors occur during condition 4 where users could only rely on the gestures of the ECA. In total there are 34 cases in which subjects clicked on a wrong card before finding the correct target position (15% error), although one subject accomplished the task without any errors. This indicates that users have difficulties to precisely interpret the gaze direction of the ECA. Nevertheless, all of these errors occurred between neighbouring cards. Since the average number of inspected cards is close to 1 (see Figure 4), we consider the assistance given by the facial gestures as sufficient as long as the user has additional information to localize the target position as it is the case in the other conditions.

Time-to-Complete. There is no significant effect for conditions on mean TC for this run ($F(3,36)=1$; $p>0.39$). Conditions 1 and 2 lead in fact to similar results: giving no cues has almost the same effects as giving bad cues (see Figure 4): subjects just ignored consciously the behaviour of the ECA.

Number of cards inspected. Conditions have a strong effect on this parameter ($F(3,36)=36.96$; $p<0.001$). Pair-wise T-tests reveal a clear advantage for condition 3 over conditions 1 and 2. On average users indeed inspected 0.5 fewer cards with a correct gaze than with a wrong or absent deictic gaze (see Figure 4). We interpret this as a clear decrease of cognitive load since less cognitive resources are used for matching cards.

Questionnaire. 4 of the 10 subjects think they were faster with the helpful assistance of the ECA and preferred this condition to play the card game.

4.4 Run B : impact of concomitant speech on deixis?

This experiment aims at evaluating the benefit of multimodal deixis in drawing user's attention using facial cues together with a spoken instruction.

4.4.1 Conditions

This second series of experiments consists likewise of four different conditions. As a major difference with Run A, the head and gaze movements of the clone are accompanied by the uttering of the demonstrative adverb "là!" (engl.: "there!"). The signal duration is 150ms. Condition 1 with no clone was replicated for reference. In conditions 2 (wrong cues) and 3 (good cues) speech onset is initiated 150ms after the onset of the deictic gestures: this delay corresponds to the average duration of the eye

saccade towards the target position implemented in our ECA. All other rewarding utterances are now omitted. Condition 4 of experiment I is replaced by a condition with correct hints, where an additional delay of 300 ms was introduced between the gestural and the following acoustic deictic gestures in order to comply with data on speech and gesture coordination (Castiello, Paulignan et al. 1991). We expect this natural coordination to enhance the ability of the ECA to attract user attention.

Fourteen users (ten male and four female) participated in this experiment. They range from 21 to 48 years of age and most are students. All regularly use a computer mouse and none reported vision problems. The dominant eye is the right eye for 8 subjects and the left eye for the other 6 subjects. As in Run A, order of presentations is counterbalanced.

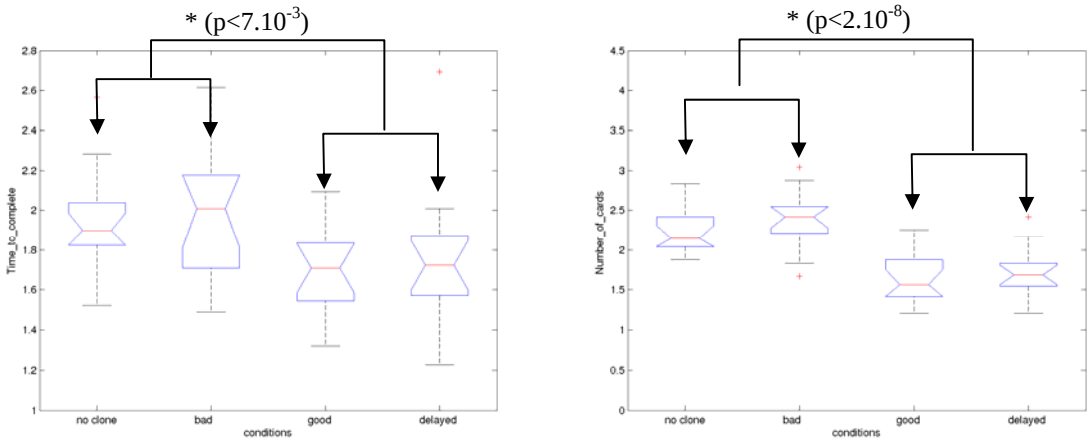


*Figure 5. Statistical analysis of TC and Number of cards inspected for each condition. Each bar represents 15 x 24 = 360 card pairings. * denotes significant differences.*

4.4.2 Results

The average number of cards was greatly reduced for conditions 1, 2 and 3 in comparison with Run A while the average TC for the reference condition 1 was increased by 230 ms: a turn lasts on average 1.72 s for Run A vs. 1.95 s for Run B. We hypothesize that the spoken instruction forced subjects to take a first decision rapidly while the perception of the verbal prompt increased the response time.

Time-to-Complete. Conditions have now a strong effect on this parameter ($F(3,52)=3.76$; $p<0.02$). The difference between TC for conditions 1 and 2 vs. 3 and 4 (see Figure 5) is now significant ($p<0.007$). The detailed analysis of TC reveals a clear advantage for 7 of 14 subjects for condition 3 (with correct cues) vs. condition 2 (with misleading cues), and for 8 of 14 subjects compared to condition 1 (without the ECA). These users now gain a substantial amount of almost 400 milliseconds (~20% of the mean duration of a game) at each turn when the target position was correctly cued.

Number of cards inspected. Conditions have a clear effect on this parameter ($F(3,52)=17.8$; $p<0.001$). Note that all subjects inspect on average 2 cards. Pair-wise T-test analysis reveals however a clear advantage ($p<0.002$) for condition 3 and 4 (correct cues) over conditions 1 and 2 (none or bad cues). On average users indeed inspect 0.7 fewer cards with a correct gaze than with a wrong or absent deictic gaze (see Figure 5). No clear tendency can be reported when considering influence of delay on performance.

Questionnaire. 11 of the 14 subjects estimated that they were quickest when correct hints were given by the ECA. Most of the subjects declared that they glance a lot at the ECA giving correct hints and discard

gestures in condition 2 but that these cues have poor influence on their speed. The movements of the ECA were judged realistic.

4.5 Discussion

When considering the number of cards inspected and the number of wrong selections in condition 4 of Run A, the current control and rendering of deictic gestures of the ECA are sufficient for subjects to localize targets, as long as there is information available at the target position to make the final decision. Without such additional information the gestures of the ECA are sometimes not precise enough to discriminate between close neighboring objects. Apart from the limitations of 3D rendering on a 2D screen, this may be due to the synchronization between gaze and head orientation that are not yet derived from empirical data. An additional limitation is the poor rendering of the facial deformations around the eyes of the ECA when eye gaze deviates from head direction, e.g. eyelids lower when gaze direction is down. We have developed an eyelid model (Elisei, Bailly et al. 2007) that may resolve this problem. Further runs of this experiment will be used to evaluate the perceptual pertinence of these objective enhancements of the ECA control and embodiment.

Several subjects complained of being disturbed by rewarding utterances in Run A. Therefore these utterances do not contribute to maintain attention and increase naturalness of the interaction. A more appropriate feedback should be short and clear according to the instruction given to subjects to react quickly. The use of speech to complement gaze and head deixis in Run B demonstrates that the pertinence of the ECA behavior has a significant impact on subjects' performance.

The constraints on speech-gaze synchronization seem quite lax. Kaur et al. (Kaur, Tremaine et al. 2003) have shown that "the fixation that best identifies the object to be moved occurs, on average, 630 ms before speech onset with a range of 150 to 1200 ms for individual subjects". Note however that these large time lags concern fixations triggering natural planning of deictic gestures… not the final intended communicative gestures. Multimodal perceptual binding is known to be flexible and binding window is known to narrow with age (Lewkowicz 1996). Apparent performance – such as measured by reaction times – may be nevertheless quite insensitive to multimodal asynchrony while regions that mediates this binding may be more or less activated depending on time shifts (Miller and D'Esposito 2005). Additional experiments should be conducted to have a better insight on the impact of multimodal synchrony on cognitive load.

The results of Run B show that the pairing performance could still be improved using multimodal deixis: imperative deixis using hand-finger pointing (Rochet-Capellan, Laboissière et al. 2008) is expected to further augment this benefit. A more important finding is the reduced number of cards inspected. The majority of participants managed to complete the task looking at significantly less cards when the ECA was providing helpful assistance. This means that even if they do not improve their speed, the search process is more efficient and probably more relaxed. We conclude that ECA assistance diminishes the cognitive load of the subjects. The answers to the questionnaire confirm this finding as the positive ratings for naturalness of the ECA and the preference for the condition where it is providing correct hints are revealed more clearly for Run B compared with Run A.

The experimental scenario presented here could likely be further improved by displaying more objects on the screen and using smaller digits. This should enhance the benefit of ECA assistance since this would require a closer examination of the objects and increase the number of objects to check in order to find the correct one without the assistance of the ECA. However, we consider the results with the current implementation as sufficient confirmation of our assumptions and encouraging motivation to study further possibilities to enhance the capabilities of the ECA.

5. EXPERIMENT II. MEDIATED FACE-TO-FACE COMMUNICATION

In Experiment I, subjects triggered saccades towards the ECA mainly to register gaze direction. Spoken instructions were very short and did not provide complementary information to gaze. During more complex interactions involving - or not - common objects of interest, interlocutors spend most of the time looking at each other. Gaze patterns during conversation are complex and depend on multiple factors such as cognitive load, cognitive and emotional state, turn, topic, role in the conversation as well as social factors (Argyle and Cook 1976; Goodwin 1980; Haddington 2002). Experiment II was designed to collect precise data on mutual gaze in order to build and parameterize a model that could supplement our ECA with effective visual attention.

The second experiment thus involves a mediated face-to-face conversation between our target human speaker and several naïve human interlocutors. The face-to-face conversation is performed through a set of two screens, pinhole cameras and microphone/headphone pairs (see Figure 6). Videos, audio signals and gaze estimates are synchronized and recorded on both sides. Thanks to a very precise measurement of the gaze direction with eye trackers and image-based estimation of the position of eyes and mouth of the respective scrutinized video, we are able to distinguish between four regions of interest on the face (see Figure 7) as well as to detect saccades, fixations and blinks.

In order to be able to build an empirical model of visual attention, the number of parameters has to be minimized and number of observations maximized. We thus focused on simple pre-scripted dialogs where most of the factors influencing gaze behavior listed above are either kept constant or implicitly controlled: interlocutors are involved in a sentence repeating game where each speaker either instructs and listens or listens and repeats. Roles (initiator vs. respondent) are swapped at the middle of the game.

We show that gaze patterns and blinking rate are influenced by the respective roles and cognitive states of the interlocutors during mediated audiovisual face-to-face conversations.

Figure 6. Mediated face-to-face conversation (from Raidt 2008). People sit in two different rooms and dialog through couples of cameras, screens, microphones and loudspeakers. Gaze of both interlocutors are monitored by two eye-trackers embedded in the TFT screens. Note that pinhole cameras and seats are positioned at the beginning of the interaction so that the cameras coincide with the top of the nose of each partner's face.

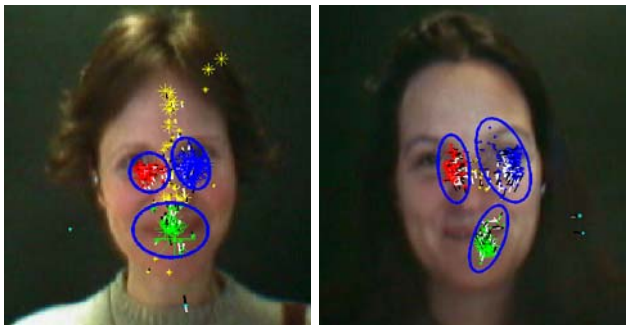


Figure 7 : Analyzing gaze patterns after compensating for head movements. The four regions of fixation on the face: left and right eye, mouth and face. The later label is assigned to fixations towards the mid-sagittal plane above the nose. Regions are determined a posteriori by the experimenters by positioning dispersion ellipsis on fixation points gathered for each experiment after compensating for head movements.

5.1 Experimental Setup

The experimental platform is displayed in Figure 6. It gives interlocutors the impression of 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 (Chen 2002).

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.

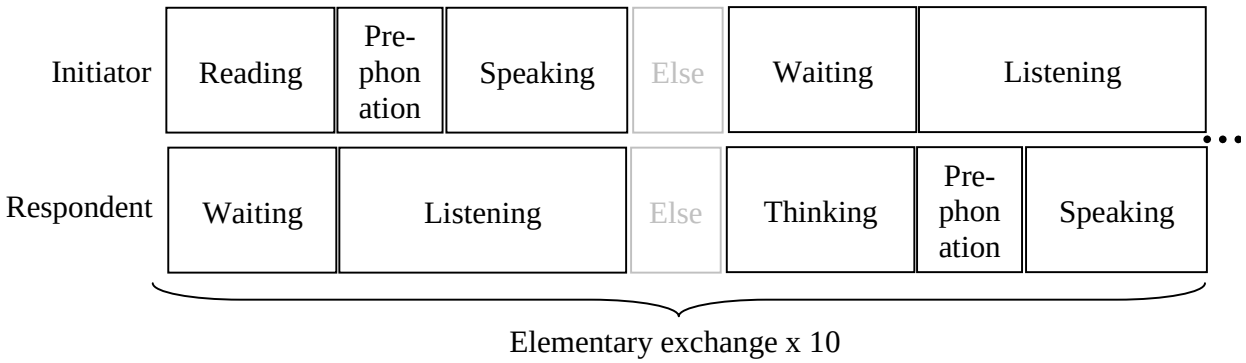


Figure 8: Time chart of the "predefined" cognitive states for one elementary exchange of the one-line interaction scenario. Prephonation starts at the first lip opening gesture preceding the utterance (usually 300ms before any speech is produced). This cognitive state has been added for analyzing the particular

gaze patterns at speech onsets. “Reading” starts at the first fixation downwards. “Else” states are added between each turn if necessary.

5.2 Scenario

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 subsequently swapped. Semantically Unpredictable Sentences (SUS) (Benoît, Grice et al. 1996) are used in this sentence-repeating game 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 (see Figure 8) that avoids complex negotiation of turn taking and eases state-dependent gaze analysis. We study inter- and intra-subject variability. In each dyad the reference target interlocutor HL interacts with subjects of the same social status, cultural background and sex (French female senior researcher). We recall here that HL is the human model of the talking head used in experiment I.

Each session consists of an on-line interaction using the full experimental setup (Run A) followed by a faked interaction (Run B). Subjects were instructed that the experiment was designed to quantify the impact of visual feedback on visiophony. For Run B, the reference speaker HL was supposed to participate as instruction giver with live audio but without visual feedback. HL was thus expected to have ungrounded gaze behavior. But the subjects are in fact confronted with an audiovisual stimulus videotaped during a previous Run A by HL with one female colleague. Most subjects do not notice that the Run B stimuli are pre-recorded.

Each subject thus experiences 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.

5.3 Data processing & labeling

Fixations were identified in the raw gaze data using a dispersion-based algorithm (Salvucci and Goldberg 2000). An affine transform was applied to compensate for head movements determined by a robust feature point tracker. All fixations were projected back on a reference head position (see Figure 7). Dispersion ellipses for all regions of interest (ROI) were positioned a posteriori by the experimenter on this reference image given all fixation points produced during the interaction. Fixations were then assigned to one to these ROIs: 'left eye’, ‘right eye’, ‘mouth’, ‘face’ (other parts than the three preceding ones such as nose), ‘other’ (when a fixation hits other parts of the screen) or ‘none’.

The speech data were aligned with the phonetic transcriptions of SUS sentences and sessions were further segmented into sequences assigned to six different cognitive states (CS): ‘pre-phonation’, ‘speaking’, ‘listening’, ‘reading’, ‘waiting’ and ‘thinking’ (see Figure 8).

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 linguistic content of the SUS he has just pronounced, and visual benefit provided by lip reading is not as crucial for speech comprehension as it is when the message is not known. 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

sequential dependencies between CS: 'pre-phonation' preceding speaking is triggered by pre-phonatory gestures such as lip opening, 'speaking' triggers 'listening' 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 (mentally preparing the sentence to repeat). We also labeled blinks. Most ECAs generate blinks with a simple random event generator. We will however show that blinking frequency is highly modulated by CS and role.

5.4 Results

We recorded interactive sessions of our target subject with 12 interlocutors but only 9 of these pairs have valid gaze data for all sessions. The data analyzed in the following thus concerns these 9 subjects. The results clearly confirm the triangular pattern of fixations (see Figure 7) scanning the eyes and the mouth previously obtained by Vatikiotis-Bateson, Eigsti et al (1998). They also confirm our choice to distinguish cognitive state and role.

5.4.1 Run A: mediated one-line face-to-face

This run starts immediately after a first calibration procedure of the eye gaze tracker. HL first instructs from a list of 10 SUS then roles are exchanged and a further exchange of 10 SUS is performed.

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 considered 5 ROIs: 'left eye', 'right eye', 'mouth', 'face' and 'none'. The ROI 'else' was disregarded since it almost never occurred during the analyzed sessions. We investigated the influence of the two factors 'role' and 'cognitive state' on the mean fixation profiles calculated for our target subject during the nine interactions (see Figure 9). This means at least 180 fixations of our target subject for each CS (90 for each role). Using MANOVA we compared the multivariate means of the fixation profiles of the CS (pre-phonation, speaking, listening, and waiting) that occur in both roles. We found that fixation profiles are significantly influenced by role ($F(5,871)=297$, $p<0.001$) and CS ($F(30,4375)=50.5$, $p<0.001$).

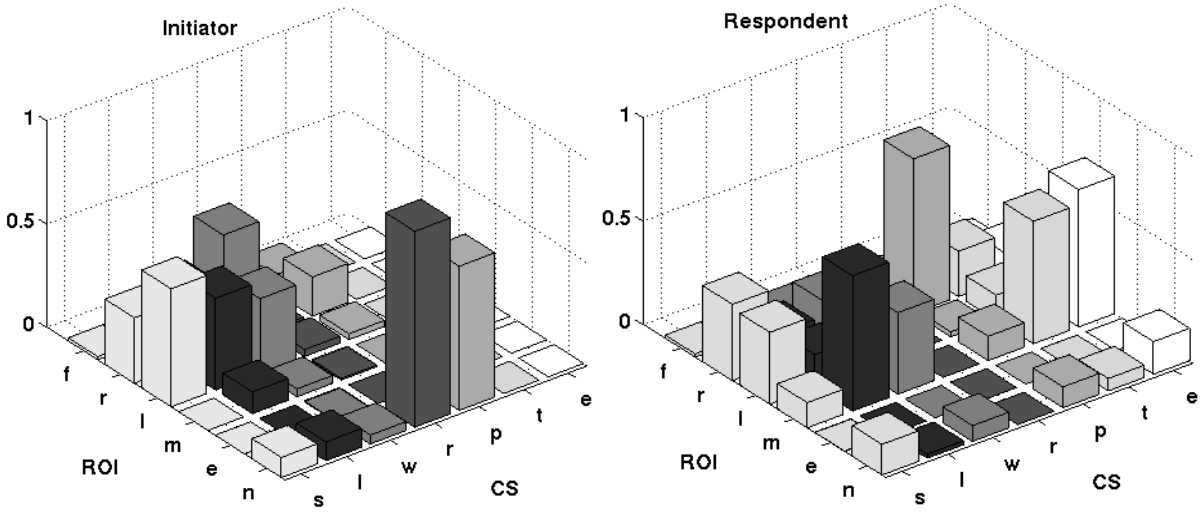


Figure 9: Fixation profiles of all interactions of our target speaker over role (initiator, respondent), ROI (face, right eye, left eye, mouth, else) and cognitive state CS (speaking, listening, waiting, reading, pre-phonation, thinking, else). The bars represent the means of the percentage of fixation time on ROI

during an instance of a cognitive state. The diagram is completed by bars (ROI named "n" for "none") representing the means of percentage of time when no fixations are detected.

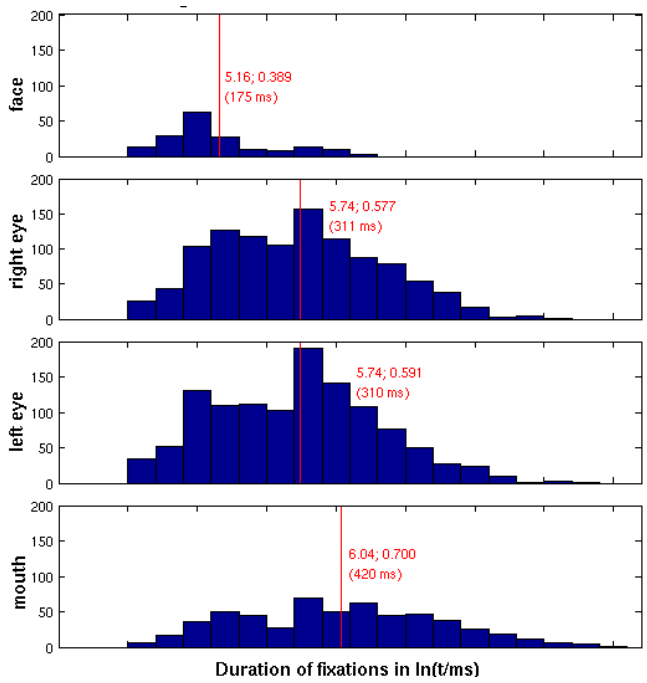


Figure 10. Distribution of fixation durations as a function of ROI. Longer fixations to the mouth could be explained by smooth pursuit of a ROI in motion.

Durations of fixations. We also characterized the duration of fixations over region of interest and role (see Figure 11). The ANOVA provided no reason to distinguish for role ($F(1,2981)=2.22$, $p>0.15$). The influence of ROI however is highly significant ($F(4,2986)=66.5$,$p<0.001$). Post hoc analysis shows that there is no difference between durations of fixations of the ROIs "right eye" and "left eye" ($F(1,2248)=0.13$,$p>0.7$). In comparison, fixations to the face are significantly shorter and fixations to the mouth significantly longer.

Blinks. Increased cognitive load is known to be correlated with fewer blinks (Wallbott and Scherer 1991). Our data reveal that blink rate does not strongly depend on role ($F(1,94)=7$,$p=0.01$) but is highly dependent on cognitive state ($F(6,94)=54$,$p<0.001$) (see also Peters and O'Sullivan 2003). 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 (see Figure 11). Particularly, in the role of respondent, the CS 'listening' strongly inhibits blinks. Strikingly, blinks often 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 (1994).

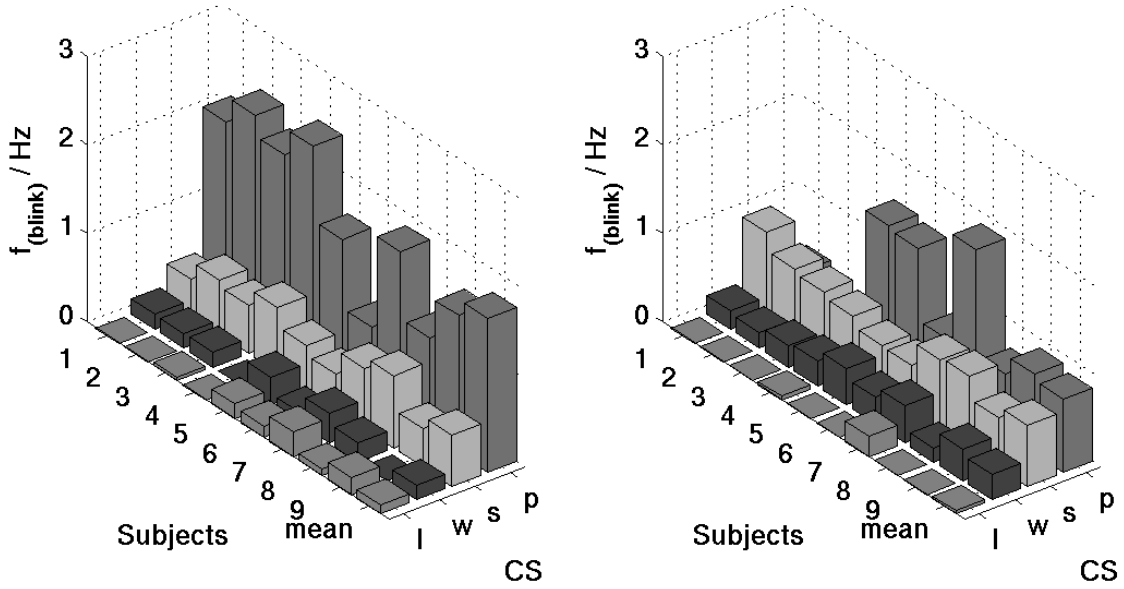


Figure 11. Blinking rate (average number of blinks per second) as a function of cognitive state for each interlocutor over role of our target speaker (left: initiator, right: respondent). Significant contrasts are evidenced such as less blinking when speaking compared to listening. Only data for 4 CS common for both roles are considered for analysis. Note that CS "reading" and "thinking" generate almost no fixations towards the face and that blinks are almost impossible to detect in this case.

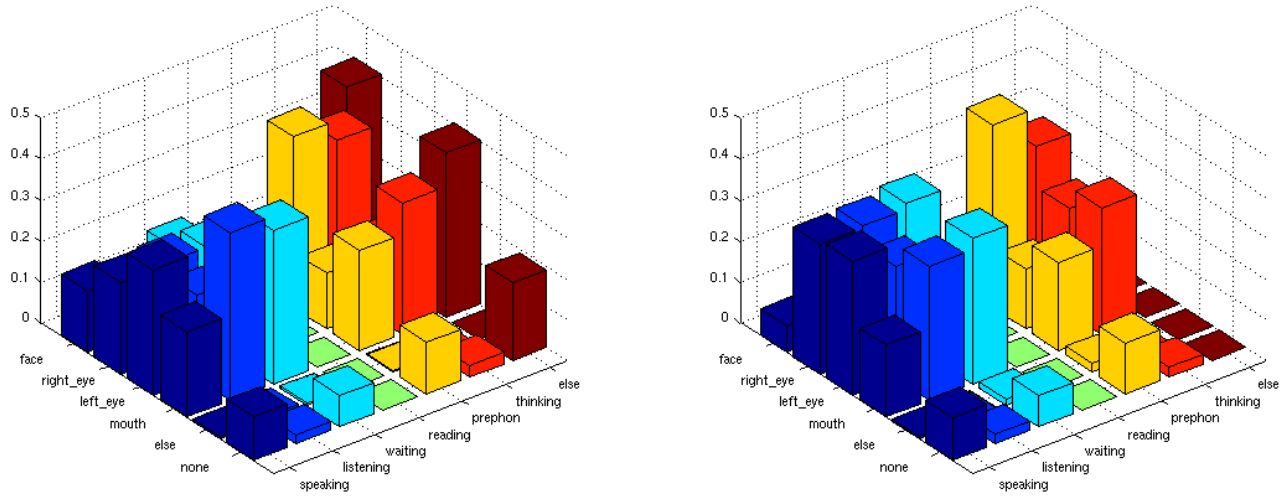


Figure 12. Comparing mean fixation profiles of all subjects acting as respondent in Run B (faked interaction at the left) vs. Run A (one-line interaction at the right). Same conventions as Figure 9. The presence of 'Else' states during faked interactions could be explained by the absence of reactive chaining of instructions.

5.4.2 Run B: faked face-to-face

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). After finishing Run A naïve subject were instructed that HL would not have any visual feedback. In fact a pre-recorded

stimulus was played back hoping that the sentence repeating game would keep the same rhythm. In fact no subject reported any awareness of this manipulation. Figure 12 shows that gaze behaviour confirms this conservative sense of presence: a MANOVA contrasting runs and CS indeed shows that interaction condition is not significant ($F(5,56)=1.55$, $p=0.18$) whereas CS has a significant impact on subjects' behaviour ($F(25,280)=1.65$, $p<0.03$). The only difference between interaction conditions is the spurious presence of 'Else' states during faked interactions due to minor desynchronizations between cognitive activities of the interlocutors that do not impact the global picture.

This control experiment confirms that the cognitive state is the main determinant of the gaze behaviour and this behaviour is sufficiently consistent – at least for HL – among interactions to fool the control of perception-action loops of the interlocutors.

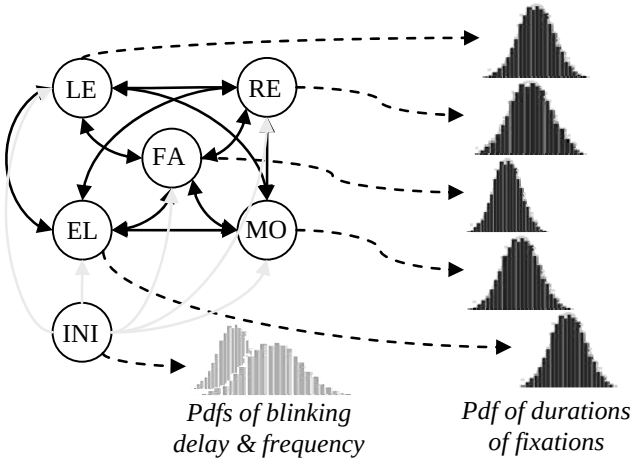


Figure 13. Modeling duration of fixations and blink frequency as a HMM. Hidden states are possible ROI of the gaze (LE vs RE: left and right eyes; MO: mouth; FA: face and EL elsewhere). When entering a state a fixation is emitted (dotted arrows) according to a state-dependent distribution (dark gray). The transition matrix between states (dark arrows) specifies the CS-dependent scan of the interlocutor's face. The initial state (INI) does not emit any fixation but sets the average delay of the first blink and the blinking rate according to HMM-specific distributions (light gray). It also specifies the way the ECA will start to scan the face by transition probabilities (light gray arrows).

5.5 Modeling

A generative model of visual attention during face-to-face interaction is proposed. It extends Lee et al.'s model (Lee, Badler et al. 2002) by taking into account both the cognitive states of the speaker as well as the statistical distribution of gaze fixations on important regions of interest on the face of the interlocutor. The control model for the gaze of our talking head was built by training and chaining role- and CS-specific Hidden Markov Models (HMM). Given a succession of CSs with associated durations it computes parameters describing the fixations of the ECA towards the various ROI on the face of its interlocutor. HMM states correspond to the different ROIs while observations specify the durations of the fixations (see Figure 13).

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 mediated face-to-face experiment. An initial state in each HMM was added to address the particular distribution of the first fixation as emphasized by

the particular distribution of prephonatory gaze patterns as well as the blinking behavior. State-dependent 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 data gathered during 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 (Lee, Badler et al. 2002) to control on-line the gaze of the clone of our target speaker: generators of random numbers with appropriate statistical distributions (uniform for state transitions, Gaussian for observations) are used to trigger state transitions and set durations of fixations and blink parameters.

Parameters of the HMM can then be switched and set according to cognitive state and role (see Figure 13) so as to signal scene comprehension by the appropriate statistical distribution of the ECA's gaze over regions of interest of its human interlocutor.

5.6 Discussion

These results confirm the eyes and mouth as dominant target zones (Vatikiotis-Bateson, Eigsti et al. 1998). Our data also confirm that the nose also attracts a significant number of fixations. Buchan et al (Buchan, Paré et al. 2008) suggest that "subjects might be using the nose as a central vantage point that permits a monitoring of the eyes and the face for social information" (see also Blais, Jack et al. 2008, for culture-specific gaze patterns contrasting holistic and analytical strategies). This momentary "de-zooming" of the elements of the face can in fact be interpreted as a scan reset that checks for other facial information's complementary – or indirect connection – to speech communication such as facial expressions. Another possibility is that the pinhole camera and small wire may also have attracted visual attention. Note however that the seats of the interlocutors were positioned at the start of the experiment so that the camera coincides with the top of their nose ridge when their own video is displayed on their screen. The only way to test this hypothesis is to use high quality teleprompters. Apart from its cost, such an experimental setting also impacts on interaction distance.

Coordination of turntaking by gaze patterns has been mainly documented by human transcriptions of dialogs (Goodwin 1980; Rutter and Durkin 1987; Novick, Hansen et al. 1996): eye contact is claimed to signal beginning as well as end of turns. Using automatic eyetracking devices, our data suggest that eye contact with the right eye of the conversational partner may signal that the speaker is ready to speak. We did not observe such a specific trigger for the end. Prosody and other non verbal cues (Duncan 1972) probably support this function. This eye preference is also no longer observed in the course of speech production. Our analysis is however focused on gaze distribution: with much more experimental data, data mining techniques such as time series analysis (Povinelli and Feng 2003) could be used to search for signatures of state transitions by specific sequences of multimodal events including not only gaze patterns but also head movements, facial gestures or speech events. Such sequences of multimodal events are expected to be much more robust to monitor turns and joint attention than single unimodal events.

We have also shown that role has a significant impact on fixation profiles as well as on blinking rate. When listening, respondents should for instance gaze towards the mouth to benefit from lip reading, while the initiators do not need to rely on 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. In order to act as expected by human

interlocutors, the ECA should thus at least be aware of its own cognitive state and its role in the interaction.

6. CONCLUSIONS AND PERSPECTIVES

Believability of context-sensitive ECAs is conditioned by both physical realism and social intelligence. Both impact the regulation of perception/action loops. As developed by Morgan & Demuth (Morgan and Demuth 1996) for syntax: linguistic structures, scene and language understanding as well as social rules are learnt and established from the multimodal signals exchanged. The quality of signal production and perception is of prime importance for building efficient interactive agents. Virtual agents in turn benefit from a detailed analysis of multimodal input and output patterns observed during human-human interactions and from the interplay with their cognitive interpretation. Monitoring multimodal signals during live interactions and building comprehensive models that link them with information structure, communicative and social functions and that ground them with the environment in which the interaction takes place is a key issue for efficient situated face-to-face interaction involving artificial agents.

Based on our findings, we have established a basis for a context-aware eye-gaze generator for 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 exogenous events that drive gaze. Furthermore other cognitive and emotional states as well as other functions of gaze (deictic or iconic gestures) should be implemented.

We will extend this study to other modalities including audible and visible speech, facial expressions, hands and head movements. We expect multimodal behaviors to be coordinated by extensive multi-layered action-perception loops with different scopes and latencies. The control of the loops, the way we react and adapt to the others, is of course not only motivated by our role in the dialog, as shown in this paper, but more generally by social factors that have been voluntarily minimized in this study.

7. ACKNOWLEDGEMENTS

We thank our colleague and target speaker Hélène Loevenbruck for her incredible patience and complicity. We also thank all of our subjects – the ones whose data have been used here and the others whose data have been corrupted by deficiencies of recording devices. Edouard Gentaz has helped us in statistical processing. This paper benefited from the pertinent suggestions of the two anonymous reviewers. We thank Peter F. Dominey and Marion Dohen for the proofreading. This project has been financed by the project Presence of the cluster ISLE of the Rhone-Alpes region.

8. REFERENCES

Argyle, M. and M. Cook (1976). Gaze and mutual gaze. London, Cambridge University Press.

Bailly, G., M. Bérar, F. Elisei and M. Odisio (2003). "Audiovisual speech synthesis." International Journal of Speech Technology **6**: 331-346.

Bailly, G., F. Elisei and S. Raidt (2005). Multimodal face-to-face interaction with a talking face: mutual attention and deixis. Human-Computer Interaction, Las Vegas

Baron-Cohen, S., A. Leslie and U. Frith (1985). "Does the autistic child have a “theory of mind”?" Cognition **21**: 37-46.

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

Blais, C., R. E. Jack, C. Scheepers, D. Fiset and R. Caldara (2008). "Culture shapes how we look at faces." PLoS ONE **3**(8): e3022.

Breazeal, C. (2000). Sociable machines: expressive social exchange between humans and robots. Sc.D. dissertation. Department of Electrical Engineering and Computer Science. MIT Boston, MA.

Buchan, J. N., M. Paré and K. G. Munhall (2007). "Spatial statistics of gaze Fixations during dynamic face processing." Social Neuroscience **2**(1): 1-13.

Buchan, J. N., M. Paré and K. G. Munhall (2008). "The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception." Brain Research **1242**: 162-171.

Carpenter, M. and M. Tomasello (2000). Joint attention, cultural learning and language acquisition: Implications for children with autism. Communicative and language intervention series. Autism spectrum disorders: A transactional perspective. A. M. Wetherby and B. M. Prizant. Baltimore, Paul H. Brooks Publishing. **9:** 30–54.

Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000). Embodied Conversational Agents. Cambridge, MIT Press.

Castiello, U., Y. Paulignan and M. Jeannerod (1991). "Temporal dissociation of motor responses and subjective awareness." Brain **114**: 2639-2655.

Chen, M. (2002). Leveraging the asymmetric sensitivity of eye contact for videoconference. SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, Minneapolis, Minnesota, 49-56.

Driver, J., G. Davis, P. Ricciardelli, P. Kidd, E. Maxwell and S. Baron-Cohen (1999). "Gaze perception triggers reflexive visuospatial orienting." Visual Cognition **6**: 509-540.

Duncan, S. (1972). "Some signals and rules for taking speaking turns in conversations." Journal of Personality and Social Psychology **23**(2): 283-292.

Elisei, F., G. Bailly and A. Casari (2007). Towards eyegaze-aware analysis and synthesis of audiovisual speech. Auditory-visual Speech Processing, Hilvarenbeek, The Netherlands, 120-125.

Evinger, C., K. Manning, J. Pellegrini, M. Basso, A. Powers and P. Sibony (1994). "Not looking while leaping: the linkage of blinking and saccadic gaze shifts." Experimental Brain Research **100**: 337-344.

Fujie, S., K. Fukushima and T. Kobayashi (2005). Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. Interspeech, Lisbon, Portugal, 889-892.

Geiger, G., T. Ezzat and T. Poggio (2003). Perceptual evaluation of video-realistic speech. Cambridge, MA, Massachusetts Institute of Technology**:** 224.

Giles, H. and R. Clair (1979). Language and Social Psychology. Oxford, Blackwell.

Goodwin, C. (1980). "Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning." Sociological inquiry - special double issue on language and social interaction **50**(3-4): 272-302.

Haddington, P. (2002). Congruent gaze points, mutual gaze and evasive gaze : some ways of using gaze in stance-taking sequences in a conversation. Studia Linguistica et Litteria Septentrionalia. Studies presented to Heikki Nyyssönen. E. Kärkäinen, J. Haines and T. Lauttamus, Department of English , University of Oulu**:** 107-125.

Itti, L., N. Dhavale and F. Pighin (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE 48th Annual International Symposium on Optical Science and Technology, San Diego, CA, 64-78.

Kaur, M., M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo and C. Sekhar Mantravadi (2003). Where is "it"? Event synchronization in gaze-speech input systems. International Conference on Multimodal Interfaces, Vancouver - BC, 151-158.

Kendon, A. (1967). "Some functions of gaze-direction in social interaction." Acta Psychologica **26**: 22-63.

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

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

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

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. Mapping the Mind: Domain specificity in cognition and culture. L. A. Hirschfeld and S. A. Gelman. Cambridge, Cambridge University Press**:** 119–148.

Lewkowicz, D. J. (1996). "Perception of auditory-visual temporal synchrony in human infants." Journal of Experimental Psychology: Human Perception & Performance **22**: 1094-1106.

Matsusaka, Y., T. Tojo and T. Kobayashi (2003). "Conversation robot participating in group conversation." IEICE Transaction of Information and System **E86-D**(1): 26-36.

Miller, L. M. and M. D'Esposito (2005). "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech." The Journal of Neuroscience **25**(25): 5884 –5893.

Morgan, J. L. and K. Demuth (1996). Signal to Syntax: an Overview. Mahwah, NJ - USA, Lawrence Erlbaum Associates.

Novick, D. G., B. Hansen and K. Ward (1996). Coordinating turn-taking with gaze. ICSLP, Philadelphia, PA, 1888-1891.

Os, E. d., L. Boves, S. Rossignol, L. t. Bosch and L. Vuurpijlc (2005). "Conversational agent or direct manipulation in human–system interaction." Speech Communication **47**(1-2): 194-207.

Peters, C. and C. O'Sullivan (2003). Attention-driven eye gaze and blinking for virtual humans. Siggraph, San Diego , CA

Peters, C., C. Pelachaud, E. Bevacqua, M. Mancini and I. Poggi (2005). A model of attention and interest using gaze behavior. Intelligent Virtual Agents, Kos, Greece, Springer Verlag, 229-240.

Picot, A., G. Bailly, F. Elisei and S. Raidt (2007). Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. International Conference on Intelligent Virtual Agents (IVA), Paris, 272-282.

Posner, M. and S. Peterson (1990). "The attention system of the human brain." Annual Review of Neuroscience **13**: 25-42.

Posner, M. I. (1980). "Orienting of attention." Quarterly Journal of Experimental Psychology **32**: 3-25.

Pourtois, G., D. Sander, M. Andres, D. Grandjean, L. Revéret, E. Olivier and P. Vuilleumier (2004). "Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals." European Journal of Neuroscience **20**: 3507-3515.

Povinelli, R. J. and X. Feng (2003). "A new temporal pattern identification method for characterization and prediction of complex time series events." IEEE Transactions on Knowledge and Data Engineering **15**(2): 339-352.

Premack, D. and G. Woodruff (1978). "Does the chimpanzee have a theory of mind?" Behavioral and brain sciences **1**: 515-526.

Raidt, S. (2008). Gaze and face-to-face communication between a human speaker and an embodied conversational agent. Mutual attention and multimodal deixis. PhD Thesis. GIPSA-Lab. Speech & Cognition dpt. Institut National Polytechnique Grenoble - France: 175 pages.

Raidt, S., G. Bailly and F. Elisei (2006). Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points? Language Ressources and Evaluation Conference (LREC), Genova, Italy, 2544-2549.

Revéret, L., G. Bailly and P. Badin (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing, China, 755-758.

Riva, G., F. Davide and W. A. Ijsselsteijn (2003). Being there: concepts, effects and measurements of user presence in synthetic environments. Amsterdam, IOS Press.

Rochet-Capellan, A., R. Laboissière, A. Galvan and J. L. Schwartz (2008). "The speech focus effect on jaw-finger coordination in a pointing task." Journal of Speech, Language, and Hearing Research **51**: 1507–1521.

Rutter, D. R. and K. Durkin (1987). "Turn-taking in mother–infant interaction: An examination of vocalizations and gaze." Developmental Psychology **23**(1): 54-61.

Salvucci, D. D. and J. H. Goldberg (2000). Identifying fixations and saccades in eye-tracking protocols. Eye Tracking Research and Applications Symposium, Palm Beach Gardens, FL, 71-78.

Scassellati, B. (2001). Foundations for a theory of mind for a humanoid robot. Department of Computer Science and Electrical Engineering. MIT Boston - MA: 174 pages.

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

Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano and K. G. Munhall (1998). "Eye movement of perceivers during audiovisual speech perception." Perception & Psychophysics **60**: 926-940.

Wallbott, H. G. and K. R. Scherer (1991). "Stress specifics: Differential effects of coping style, gender, and type of stressor on automatic arousal, facial expression, and subjective feeling." Journal of Personality and Social Psychology **61**: 147-156.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. Eye Movements and Vision. L. A. Riggs. New York, Plenum Press. **VII:** 171-196.