Face-to-face interaction with a conversationnal agent: eye-gaze and deixis

Stephan Raidt, Frédéric Elisei, Gérard Bailly

Face-to-Face Interaction with a Conversational Agent: Eye-Gaze and Deixis

Stephan Raidt **Frédéric Elisei** **Gérard Bailly**

Institut de la Communication Parlée,UMR CNRS n°5009,
INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
(+33).(0)4.76.57.45.33

{raidt,elisei,bailly}@icp.inpg.fr

ABSTRACT

We present a series of experiments that involve a face-to-face interaction between an embodied conversational agent (ECA) and a human interlocutor. The main challenge is to provide the interlocutor with implicit and explicit signs of mutual interest and attention and of the awareness of environmental conditions in which the interaction takes place. A video realistic talking head with independent head and eye movements was used as a talking agent interacting with a user during a simple card game offering different levels of help and guidance. We analyzed the user performance and how he perceived the quality of assistance given by the embodied conversational agent. The experiment showed that users profit from its presence and its facial deictic cues.

Keywords

Face-to-Face Interaction; Mutual Attention; Facial Animation, Eye-gaze; Multimodal Interaction, Deixis;

1. INTRODUCTION

Two complementary perspectives coexist implicitly in the development of Embodied Conversational Agents (ECA). The dialogic perspective [5] focuses on the study of communicative interaction, with strong semantic and linguistic components, between human and/or software agents in mediated information systems. This perspective considers that the ultimate goal of interaction is information retrieval with ECA being the communication interface.

The sociable perspective [3, 4] puts forward the embodiment. In this later perspective our analysis and comprehension of an interaction is deeply grounded in our senses and actuators and we do have strong expectations on how dialogic information – if any - is encoded into multimodal signals. Of course users' mental representations and states, common belief spaces built when interacting with ECA is a complex construct that takes into account both communicative and sociable dimensions of interaction. Appropriate interaction loops have to be implemented. They have to synchronize low-frequency dialogic loops - that require analysis, comprehension and synthesis of dialog acts with time-scales of the order of a few utterances - with more high-frequency interaction loops - that require prompt reactions to the scene analysis such as involved in eye contact or exogenous saccades. Both information- and signal-driven interactions should be then coupled to guarantee efficiency, believability, trustfulness and user-friendliness of the information retrieval.

The work described here is dedicated to the analysis, modeling and control of multimodal face-to-face interaction between an embodied virtual conversational agent and a user. We particularly study here the impact of mutual attention in a series of simple deictic tasks.

Conference'04, Month 1–2, 2004, City, State, Country.

2. EYE GAZE AND ATTENTION

The cognitive demand of a task has a striking impact on the human audiovisual analysis of scenes and their perception. The eye gaze pattern during the examination of pictures is highly influenced by the interest of the observer as shown by Yarbus [18] for example. He instructed a subject to answer seven different questions about the depicted situation in Repin's picture "An Unexpected Visitor". Resulting eye gaze patterns show that eyes tend to be attracted by those parts of the scene containing relevant information for the answers to these questions.

Similarly Vatikiotis-Bateson et al [16] show that eye gaze patterns of perceivers 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).

The work of Simons and Chabris [15] suggests that attention may be essential to consciously perceive any aspect of a scene. Major changes to objects or scenes may be ignored ('change blindness') and objects may not even be perceived ('inattentional blindness') if they are not in our focus of attention.

While not even salient visual features as for instance highlighting or blinking are given much attention, unless they convey important information for the recognition of a scene, visual attention can indirectly be guided using visual cues.

In the Posner cueing paradigm [10, 11], observers' performance in detecting a target is typically better in trials in which the target is present at the location indicated by a former visual cue than in

trials in which the target appears at the uncued location. The outstanding prominence of the human face in this respect was shown by Langton et al. [7, 8], who have shown that observers react more quickly when the cue is an oriented face than when it is an arrow. Driver et al. [6] have shown that a concomitant eye gaze alone also speeds the reaction time.

The work presented below extends these studies towards practical use of eye saccades as cues to guide the direction of social attention. It has to be mentioned yet, that the benefit achieved in the above-cited studies is very small (20ms) compared to the total reaction time (around 300ms) and the conditions are quite constricted. The experiment presented here in contrast is very complex and therefore less sensitive to differences of this magnitude.

3. Face-to-Face Interaction with an ECA

For a beneficial application of an ECA in face-to-face interaction with a human, it has to be equipped with the means to derive meaning from the implicitly and explicitly communicational gestures of a human interlocutor. Likewise, it needs the ability to reproduce such gestures for communication purposes. To convince a user of it's usefulness, an ECA must 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. Conveying its ability to interpret human behavior, the system encourages the interlocutor to show the appropriate natural activity. Therefore it is important that the ECA knows how to display what would correspond to mental states in humans. This allows to understand the machine processes of the system in terms of human expressiveness and to assign them a corresponding meaning. Thus the system may maintain an interaction based on human patterns.

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).

We concentrate on eye gaze patterns as a crucial modality of human activity to attribute beliefs, goals, and percepts to other people. The set of abilities that allow an individual to infer these hidden mental states based on observed actions and behavior is called a "theory of mind" [12]. Several TOM have been proposed [2, 9]. Baron-Cohen proposes for example an Eye Direction Detector (EDD) and an Intentionality Detector (ID) as basic components of a Shared Attention Mechanism (SAM) that is essential to the TOM's bootstrap. The actual implementation of these modules requires the coordination of a large number of perceptual, sensorimotor, attentional, and cognitive processes. Scassellati [14] developed an *embodied theory of mind* to link high-level cognitive skills to the low-level motor and perceptual abilities of a humanoid robot. The low-level motor abilities comprised coordinated eye, head and arm movements for pointing. The low-level perceptual abilities comprised essentially detection of salient textures and motion for monitoring pointing and visual attention.

We see here that even the unique control of eye gaze in face-to-face interaction is very complex and requires the coordination and cooperation of multiple processes. Some of these processes are more particularly dedicated to the analysis of the multimodal scene whereas some others are more particularly concerned with interpreting the communicative intentions of the user that the information system may respond to.

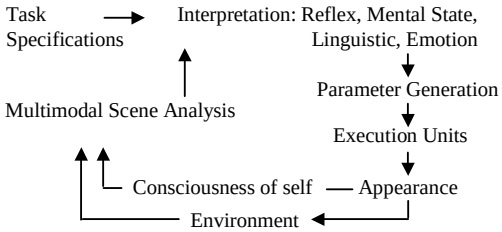


Figure 1: ECA-Human interaction scheme

4. Research Aims

Our perspective is to develop an embodied TOM to link high-level cognitive skills to the low-level motor and perceptual abilities of a virtual conversational agent and to demonstrate that such a TOM will provide the information system with enhanced user satisfaction, efficient and robust 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 can also interact with other virtual objects (e.g. virtual icons) surrounding the virtual talking head.

Figure 2: Face-to-face platform for interacting with a 3D clone

5. Developing our Face-to-Face Platform

During interaction with the system, the user sits in front of a standard-looking flat screen where a 3D talking head faces him, as shown in Figure 2. Hardware and software specificities allow the user to interact with the system using eye gaze, a mouse and speech. The 3D clone can look at the user, talk to him, and react to his eye gaze. These elements form the basis of a grounded virtual face-to-face situation.

5.1 The Hardware

The flat screen used for the display of the ECA is a Tobii1750 eye-tracker[1] that discretely embeds infrared lights and a camera. It allows us to detect, at up to 60Hz, the eye gaze of the user whose head can move and rotate freely in a fairly unrestricted 3D

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

volume having the shape of a 40cm square cube centered at 50cm away from the screen. Effective accuracy is obtained through a single short calibration procedure that each user must follow. Standard graphic hardware with 3D acceleration allows real-time rendering of the talking head on the screen.

5.2 The Talking Head

We use the cloned 3D appearance and articulation gestures of a real human [1, 13], (see Figure 3). The eye gaze of the clone can be controlled independently to look at the user, at the spot same on the screen as does the user (giving signs of mutual attention) or to actively direct the user's attention to 2D objects on the screen. Hereby the vergence of the eyes is controlled and provides a crucial cue for inferring spatial cognition. The virtual neck is also articulated and can accompany the eye-gaze movements. The audiovisual messages can either be recordings of human speaker or be synthesized from text input. In the present work synthetic signals were generated off-line to avoid reaction delays.

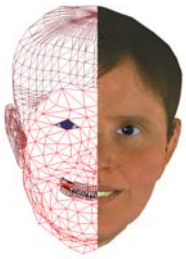


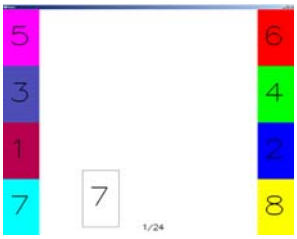
Figure 3: Animated 3D clone with independent head and eye movements

6. Face-to-Face Interaction Experiment

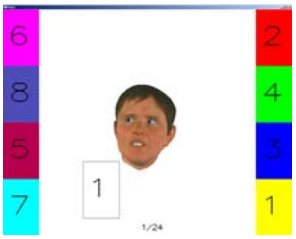
6.1 The Experimental Scenario

To follow up the findings of Langton and Driver about the special ability of human faces and eyes to direct attention, we designed an experiment with an ECA in a complex scene. Our aim is to investigate the effect of the presence of an ECA able to orientate his head and gaze on the user performance during a retrieval task. We chose a virtual card game, where the user is asked to locate and indicate the correct target position of a play card.

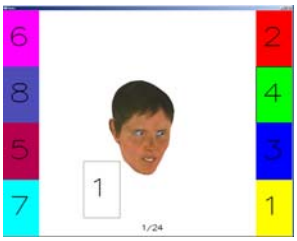
On each side of a computer screen, four cards are shown, that respectively reveal a number once the play card at the lower middle of the screen was selected by a mouse click of a test person. Likewise by clicking with the mouse the play card has to be put down on one of the eight possible target positions represented by the cards at the sides. The correct target position is the one showing the same digit as the play card. To anticipate memory effects the values shown on the cards are shuffled before each turn. The target position is alternated randomly, but uniformly distributed amongst the eight possibilities to compensate possible influences of the respective positions on the user performance. General information about the task is displayed on the screen at the beginning.



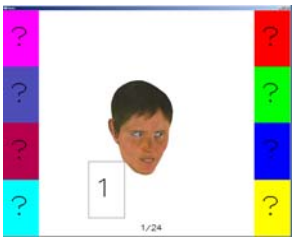
1.Condition: no clone



2.Condition: wrong cues



3.Condition: correct cues



4.Condition: numbers hidden

Figure 4: Experimental conditions: The experiment is divided into four conditions with different levels of help and guidance by the clone.

To analyze the effect of the presence of the ECA, four experimental conditions of 24 turns respectively were realized (see Figure 4). Condition specific instructions appear as text on the screen before the start of each condition. In the first condition, no clone is displayed. In the second condition, the 3D head is visible and gives a random eye-saccade accompanied by a head turn to one of the non-matching cards just after the eight numbers are revealed. In the third condition, the saccade of the 3D head indicates the correct target position. In the fourth condition, the cards are no longer revealed. Here the values cannot be read. Question marks remain displayed instead, while the saccade by the clone remains correct. In this condition the saccade of the clone is the only cue about the position where to put down the play card. In all four conditions, the oral messages are identical. The test person is invited to find the matching card and random congratulations are uttered when it is found.

6.2 Data Collection

Ten users (six male and four female) took part in the experiment described above. Participants ranged in age from 23 to 33 years and most were students. All regularly use a computer mouse and none reported vision problems. The dominant eye was detected to be the right eye for all but one subject.

Each user had to play the game with the four successive experimental conditions described above. Specific written instructions given on the screen, followed by a corresponding training session with three card pairs to match preceded each condition and thus informed the user about the respective gaze behavior of the clone. No strategy was suggested, but the user was instructed to find the matching pair as fast as possible.

During the four successive playing sessions, the time needed to match the right pair was measured as the time between the click on the play card and the time of the click on its target position. Furthermore, the gaze of the test persons was recorded. This allows to calculate the time spent on the different screen objects and to determine how many objects have been looked at, before the target position for the card was found and selected with a

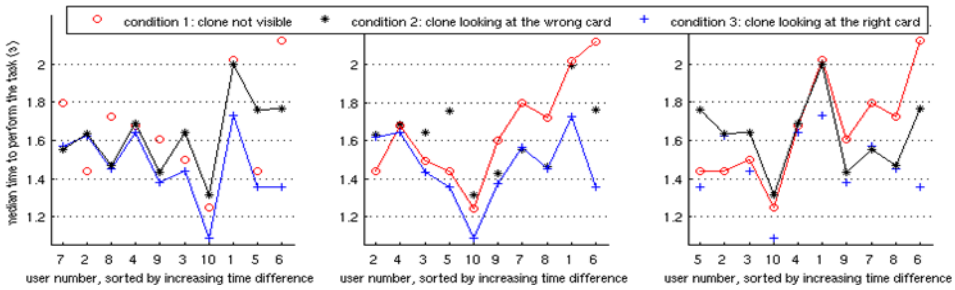


Figure 5: Comparison of performance time median:
left: condition with correct hints (+) and condition with misleading hints (*)
middle: condition with correct hints (+) and condition without clone (°)
right: condition with misleading hints (*) and condition without clone (°)

mouse click. A system log of all the posted messages was also recorded.

After the experiment, which lasted less than 20 minutes, participants ranked various subjective aspects of the experiment on a five-point scale, quantifying the following points:

- clone quality: authenticity of neck and eye movements, accuracy of gaze in condition 4 (numbers hidden)
- estimation of personal performance: velocity, influence by the distracting eye gazes in condition 2, usefulness of the clone when giving correct indications
- experimental condition: preference of clone behavior, preference of behavior for best performance

6.3 Analysis of the Experiment

6.3.1 Expected Outcome

Corresponding to the structure of the experiment we expected a negative influence on the test person's performance when the clone gave misleading cues, and a positive influence when giving matching cues with the condition where no clone is displayed as a reference. The fourth condition, where no numbers are shown on the cards the clone's saccades being the only information available, was expected to reveal the precision with which the gaze direction of the clone could be perceived.

6.3.2 Influence of the Experimental Conditions on User Performance

The test persons' subjective ratings of the clone's gaze precision being middle to very bad are confirmed by the rather high error rate of 15% (34 errors/240 turns) observed during condition four. Especially one subject (number 9) had an extremely high personal error rate of 37% (9 errors/24 turns). Closer examination however showed that almost all wrong choices were made between neighboring cards. This may be due to the momentarily basic implementation of synchronization between gaze and head orientation that is not derived from measured data yet. Another reason may be the unanimated eyelids.

During the other experimental conditions (1 to 3) only one error occurred and they can therefore be considered as being successfully performed. This is consequential to the possibility to verify the correctness of choice by comparing the numbers shown on the play card and the target position before selecting.

For the evaluation of user performance one possibility is to measure the time between taking up and putting down the play card at the target position by mouse clicks respectively. However the described task is rather complex and therefore improvements in performance of a magnitude as found by Langton et al. [7, 8] are very unlikely to be measured. The dragging of the play card with the mouse and selection by mouse click are quite complex movements that demand precise coordination and offer various possi bilities for time delays.

Hence the number of visited target positions was considered as another possibility for performance evaluation. However, searching for the target position without using the assistance by the clone, the user may have to look at 1 up to 8 cards before finding the correct one. Including repetitions this may even be more. Therefore the results of both methods of evaluation are highly depending on chance that cannot be controlled.

However the measurements show an influence of the ECA corresponding to our expectations as show in Figure 5. We chose to compare the medians as the task is susceptible to produce outliers in the measured performance time. All test persons completed the task faster in the condition with the clone giving correct cues compared to the conditions with the clone giving misleading cues and the condition without the clone (one exception). This corresponds to our expectations that users profit from helpful deictic visual gestures of an ECA.

There is no clearly negative influence of the misleading cues compared to the experimental condition without the clone as was expected. We found no explanation for this. Regarding the temporal order of succeeding conditions we assumed a learning effect, which could however not be confirmed analyzing the evolution of the subjects' respective performance times.

For an interpretation of the measured gaze data, we first have to verify that we recorded it sufficiently. Summing up the time spend on surveyed objects on the screen showed that this is indeed the case. Only about 10% of the gaze was not dedicated to surveyed objects, which may be explained by the fact that the play card, as a moving object was not surveyed. We conclude that effects of periphery view were of no major importance to user performance.

Comparing the medians of performance time and the time spent looking at the clone, no general influence can be found. It seams that the test persons developed different strategies to complete the task. This impression is fostered by the subjects' estimation of the influence of the clone's correct cues. Some subjects rated this influence very high and a corresponding high time span of their gaze was dedicated to the clone. Figure 6 shows the measured

data for such a subject (compare to Figure 7) who profited from the ECA both in performance time and number of cards visited.

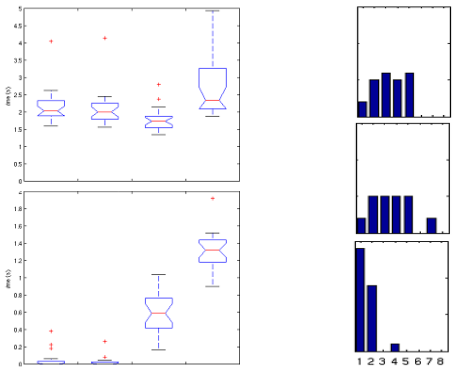


Figure 6: Example of measured results (subject 1): left: The upper diagram shows performance times as box plots for conditions 1 to 4 from left to right. The lower diagram shows the time that the user spend looking at the clone during these conditions.
right: Number of cards visited before selecting target position (condition 1 to 3 from top bottom).

Other subjects chose exactly the contrary possibility, spending little time on the clone and rating his influence accordingly as minor.
All subjects agreed in their estimation that they were little influenced by the clone giving misleading cues (see Figure 7). This ability to successfully ignore the clone may be encouraged by the preceding condition without the clone. Here the test persons have to go through all possible target positions until the correct one is found and some users seem to maintain this strategy once adopted even when they might profit from the clone.

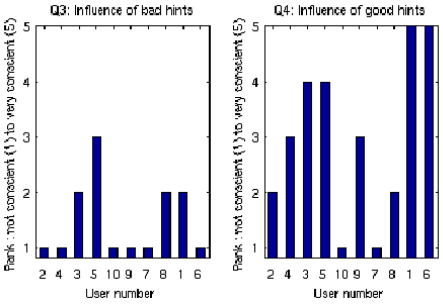


Figure 7: User estimation about the influence of the clone behavior on their performance

6.4 Discussion and Perspectives

We described here our first effort to design a system enabling face-to-face interaction between a 3D talking head capable of mutual attention and shared reality. We addressed an application that resembles our daily use of computer and mouse for information retrieval on screen, e.g. through file directories or icons. With this interesting choice, we collected representative data on the real interaction of our participants with the computer.
Validating the interest of our platform in human computer interaction, we found that various strategies were used to perform the task. It might be interesting to explore in which case the cognitive load is higher. One way to achieve this would be to remove the possible pause between two successive cards, and to instruct users that they must act as fast as possible, without pausing. The number of cards could be increased, or cards could be revealed only when looked at. To check for habituation or fatigue, extra turns might be added, so that conditions might be repeated in various orders for the same participant.
However users could already benefit from this very basic implementation of facial deictic gestures in an ECA with accompanying speech. As no general correlation between user performance and the time spent on the clone itself could be found, the presence of a clone alone already seams to motivate the user and result in better performance.
Despite the limitations of a display without stereovision and the lack of spatial accuracy in interpreting the clone eye gaze, our 3D clone could successfully be used as an extra modality. Of course, 3D rendering on a screen lacks the depth sensation and fails to transmit faithfully all the directions in the 3D world. It remains to be tested whether the neck and eye gaze animations could be improved to get better results. For this objective, we might use our interaction system to record eye gaze of real humans in the symmetrical situation.

7. CONCLUSIONS

Our first approach to implement an animated 3D talking clone as an ECA able to maintain face-to-face interaction with a human interlocutor proofed the capability to direct a user's attention with head and eye movements of the clone. With our experimental setup, we have developed a rather flexible hybrid hard- and software platform that allows us to place users in a multi-modal face-to-face interaction with our talking agent and to record their activity for statistical analysis.
We demonstrated that users can benefit already from a very basic implementation of facial deictic gestures in an ECA with accompanying speech. Although the card game scenario leaves a lot of freedom to the users, who apparently developed different strategies and although the current implementation of the ECA does not yet realize very sophisticated reactions to the user's activities, all test persons profited from the clone's presence.
We believe that the study and modeling of the components of human face-to-face interaction are crucial elements to obtain an intuitive, robust and reliable communication interface able to establish an interaction loop. While most experimental data on speech and gaze examine attention of the listener, almost no experimental data is currently available on gaze patterns when speaking [17]. This motivates our investigation of interactive real-time eye-gaze patterns of human speakers in face-to-face communication with a special focus on turn taking and the speaking and listening states.

ACKNOWLEDGEMENTS

We gratefully acknowledge the patience of Hélène Lœvenbruck, the human model for our clone. We thank Alain Arnal and Christophe Savariaux for their technical assistance with the audiovisual capture platform, as well as Matthias Odisio, Pauline Welby and the reviewers for their helpful comments. We are also grateful to our experiment participants. This work would not have

been possible without the results from several past projects involving students.

REFERENCES

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

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

[3] Breazeal, C. (2002) *Designing Sociable Robots*.: The MIT Press.

[4] Brooks, R.A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M. (1999) *The Cog Project: Building a Humanoid Robot" in Computation for Metaphors, Analogy, and Agents*, in *Lecture Notes in Artificial Intelligence*, C. Nehaniv, Editor. Springer: New York. p. 52–87.

[5] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (2000) *Embodied Conversational Agents*. Cambridge: MIT Press.

[6] Driver, J., Davis, G., Riccardelli, P., Kidd, P., Maxwell, E., and Baron-Cohen, S. (1999) *Shared attention and the social brain : gaze perception triggers automatic visuospatial orienting in adults.* Visual Cognition, **6**(5): p.509-540.

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

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

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

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

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

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

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

[14] Scassellati, B. (2001) *Foundations for a theory of mind for a humanoid robot*, in *Department of Computer Science and Electrical Engineering*. MIT: Boston - MA. p.174.

[15] Simons, D.J. and Chabris, C.F. (1999) *Gorillas in our midst: sustained inattentional blindness for dynamic events.* Perception, **28**: p.1059-1074.

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

[17] Vertegaal, R., Slagter, R., Veer, G.v.d., and Nijholt, A. (2001) *Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes.* in *Conference on Human Factors in Computing Systems*. Seattle, USA. p.301 - 308.

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