Does a Virtual Talking Face Generate Proper Multimodal Cues to Draw User's Attention to Points of Interest?

Stephan Raidt, Gérard Bailly, Frédéric Elisei

Does a Virtual Talking Face Generate Proper Multimodal Cues to Draw User's Attention to Points of Interest?

Stephan Raidt, Gérard Bailly & Frederic Elisei

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{raidt,bailly,elisei}@icp.inpg.fr

Abstract

We present a series of experiments investigating face-to-face interaction between an Embodied Conversational Agent (ECA) and a human interlocutor. The ECA is embodied by a video realistic talking head with independent head and eye movements. For a beneficial application in face-to-face interaction, the ECA should be able to derive meaning from communicational gestures of a human interlocutor, and likewise to reproduce such gestures. Conveying its capability to interpret human behaviour, the system encourages the interlocutor to show 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 interpret 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.
During a first experiment we investigated the ability of our talking head to direct user attention with facial deictic cues (Raidt, Bailly et al. 2005). Users interact with the ECA during a simple card game offering different levels of help and guidance through facial deictic cues. We analyzed the users' performance and their perception of the quality of assistance given by the ECA. The experiment showed that users profit from its presence and its facial deictic cues. In the continuative series of experiments presented here, we investigated the effect of an enhancement of the multimodality of the deictic gestures by adding a spoken instruction.

1. Introduction

Two complementary perspectives coexist implicitly in the development of Embodied Conversational Agents (ECA). The dialogic perspective (Cassell, Sullivan et al. 2000) 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 (Brooks, Breazeal et al. 1999; Breazeal 2002) puts forward the embodiment. In this later perspective the analysis and comprehension of an interaction is deeply grounded in our senses and actuators. We do in fact have strong expectations on how dialogic information is encoded into multimodal signals. These perspectives are complementary and should benefit from each other. Users' state and mental representations as well as common belief spaces built when interacting with others are complex constructs that take into account both communicative and sociable dimensions of interaction. Appropriate interaction loops have to be implemented. They have to synchronize at least two different dialogic loops. On the one hand there are low-frequency dialogic loops. They require analysis, comprehension and synthesis of dialog acts with time scales of the order of a few utterances. On the other hand there are interaction loops of higher frequency. These include the prompt reactions to the scene analysis such as involved in eye contact, or exogenous saccades. Similarly the YTTM model (Thórisson 2002) of turn-taking possesses three layered feedback loops (reactive, process control and content). The intermediate process control loop of the YTTM is responsible for the willful control of the social interaction (starts and stops, breaks, back-channeling, etc). In all models, information- and signal-driven interactions should then be 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 a virtual ECA represented by a talking head and a user. We particularly study here the impact of facial deictic gestures of the ECA enhanced by a concomitant spoken instruction on user performance in simple search and retrieval tasks.

2. The Interaction Scenario

To follow up the findings of Langton (Langton, Watt et al. 2000) and Driver (Driver, Davis et al. 1999) about the special ability of human faces and eyes to direct attention, we designed an interaction scenario where an ECA should use these means to direct the user's attention in a complex virtual scene. Our aim was to investigate the effect of facial deictic gestures on the user's performance during a search and retrieval task. We chose an on-screen card game, where the user is asked to locate the correct target position of a playing card.
The card game consists of eight cards, the numbers of which are revealed once the playing card at the lower middle of the screen is selected with a mouse click. During each turn the playing card has to be put down on one of the eight possible target cards placed on the sides of the screen. The correct target card is the one with the same digit as the playing card. To anticipate memory effects the numbers on the cards are shuffled before each turn. The target position is alternated 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 target position on the user performance. The color of the target depends only on the position and not on the digit (see Figure 1).
We compared 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 1. The ECA can utter spoken commands and can indicate directions with an eye saccade combined with a

Condition 1: no clone | Condition 2: wrong cues | Condition 3: correct cues | Condition 4: digits hidden

Figure 1: Experimental conditions: The experiment is divided into four conditions with different levels of help and guidance by the clone. When the clone is present it can give helpful or misleading facial cues. When the numbers on the cards are not shown (condition 4) these cues are the only possibility to locate the correct target card.

head turn. Each experimental condition comprises three training cycles to allow the subjects to become accustomed to the task, which are followed by 24 measurement cycles. The characteristics of the upcoming condition are described as text on the screen before the training cycles and thus inform the user about the expected gaze behavior of the clone. General information explaining the task is given as text on the screen at the beginning of the experiment. The user is instructed to put down the playing card on the target position as fast as possible but no strategy is suggested. Users were not informed about the data to be measured.

For the evaluation of the experiments the reaction time and the gaze behavior have been monitored. The reaction time was measured as the time span between the first mouse click on the playing card and the click on the correct target position. As the card game was displayed on a monitor with embedded eye-tracking[1], the visual focus of the user on the screen could be recorded. We thus computed which objects on the screen have been looked at and how much time users spent on them. Possible objects were the eight target cards on the sides and the face of the ECA. Eye gaze towards the playing card was not monitored, as it was constantly moving during the experiment.

At the end of the experiment, which lasted about 15 minutes, participants were asked to answer a questionnaire. They had to rank various subjective aspects of the experiment on a five-point MOS scale, and to choose which condition they considered as most appropriate and fastest.

2.1. Experiment I: Impact of Facial Cues

This experiment aims at evaluating the capacity of our ECA for attracting user's attention using facial cues and quantifying the impact of good and bad hints on the user's performance. This work builds on the psychophysical experiments on visual priming done by Langton et al (Langton, Watt et al. 2000; Langton and Bruce 1999). We resume here the scenario and findings of the first series to provide a framework for the discussion of the second series of experiments.

2.1.1. Conditions

The first series of experiments consisted of four different conditions, screenshots of which are displayed in Figure 1. For condition 1, no ECA is displayed. For condition 2, the ECA is visible and provides misleading hints: it indicates randomly one of the non-matching positions with a facial gesture as soon as the user selects the playing card. In condition 3, it provides supportive hints: it indicates the correct target position. For condition 4, cards remain upside down and the correct visual cues provided by the ECA are the only clue to find the correct target position.

In all conditions where the ECA is displayed it encourages the user with randomly chosen utterances alternating between motivation and congratulations. The utterances are generated off-line to avoid computation delays.

2.1.2. Expected Outcome

We had strong expectations about the data to be collected. Corresponding to the design of the experiment we expected a negative influence on the test person's performance when the clone gives misleading cues and a positive influence when giving supportive hints. The condition where no clone is displayed was supposed to serve as a reference. From the fourth condition, we expected to measure the precision with which the gaze direction of the ECA could be perceived.

2.1.3. Data Processing

Before evaluating the measured reaction times, extreme outliers were detected and replaced by a mean value computed from the remaining valid data. We analysed the reaction time means within each experimental condition and checked with an ANOVA for significance at p=0.05 for the respective subjects. The significant differences between pairs of distributions are indicated in Figure 2 with stars. They are used as a measure of performance of the subjects.

The number of looked at possible target positions was determined to analyse the search strategy of the subjects. Before analysing the data collected by the eye tracker its reliability was tested and invalid data rejected. An ANOVA analysis at p=0.05 was then performed on the valid data and significant differences between pairs of distributions are indicated in Figure 3 with stars right above the subject number.

2.1.4. Results

During the conditions that permitted comparing the digit on the playing card and the target card only one wrong selection occurred. These can therefore be considered as accomplished successfully. Numerous errors occurred (15% errors) during condition 4 where users had to rely

[1] Tobii 1750 Eeye-tracker

o no Clone; + Clone gives misleading hints; □ Clone gives correct hints; ∧ Cards remain hidden;
* *p = 0.05* ; **X** data not valid

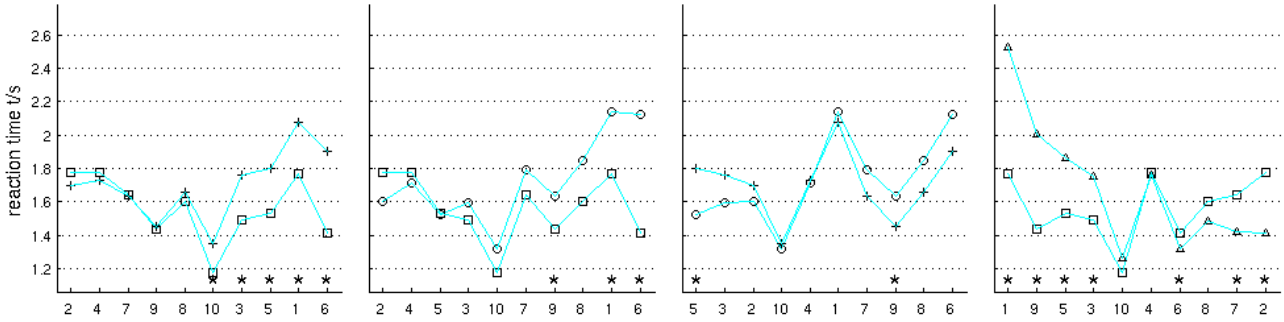

Figure 2: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. The x-coordinate lists the subjects whereas the digit represents the order of participation. Mean reaction times for each user and for each session are displayed together with the statistical significance of the underlying distributions (stars displayed at the bottom when p<0.05).

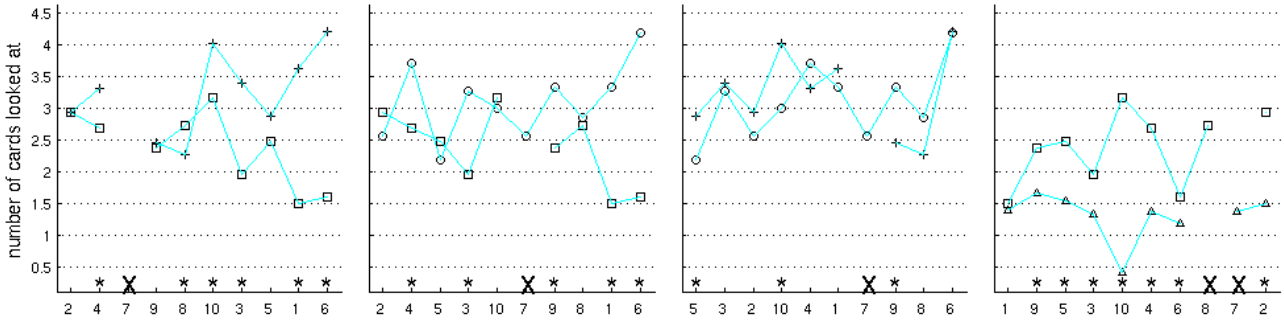

Figure 3: Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 2.

on the gestures of the ECA. This indicates that users have difficulties to precisely interpret the gaze direction of the ECA. Nevertheless, as all of these errors occurred between neighbouring cards, we consider the assistance given by the facial gestures as sufficient as long as the user has additional information to confirm the choice as it is the case during the other conditions.

The means of reaction time are displayed in Figure 2. They are sorted for increasing difference between the compared conditions for the respective subjects that are represented by their number of participation order. Significance is marked with a star above the subject number on the x-coordinate. The diagram shows that 5 out of 10 subjects showed significantly shorter reaction times during the condition 3 (with correct cues by the ECA) compared to the condition 2 (with wrong cues) and that three subjects did so compared to the condition 1 (without the ECA). These users gain a substantial amount of 200 milliseconds (~10% of the mean duration of a turn) at each drawing. Comparison of conditions 1 and 2 leads in fact to similar results. One subject out of 10 shows significant shorter reaction times when no ECA is present whereas one shows longer ones compared to when the ECA is giving wrong hints. As several selection errors occurred during the condition 4 (with cards remaining hidden until selection), it is obvious that this entails longer reaction times for half of the subjects.

The analysis of the eye tracker data is shown in Figure 3. Some of the data was not sufficiently reliable and had to be excluded from evaluation. This concerns subject 7 who had to be excluded completely and subject 8 (marked with X above the subject number on the x-coordinate). The remaining data is sufficiently reliable. Analysis of the means with an ANOVA at p=0.05 evidences a clear advantage when correct hints are given by the ECA. On average users that profit from the hints inspect 1.5 cards less when given a correct gaze cue than when given wrong or no deictic cues. We interpret this as a clear decrease of cognitive load since less cognitive resources are used for checking cards.

From the examination of the answers to the questionnaire it can be retained that 4 of the 10 subjects think they are faster with the helpful assistance of the ECA and prefer this condition to play the card game.

2.2. Experiment II: Impact of Multimodal Deixis

This experiment aims at evaluating the possible benefit from the enhancement of the ECA's multimodal deixis by a spoken instruction.

2.2.1. Conditions

This second series of experiments is based on experiment I and consists likewise of four different conditions. As a major difference the head and gaze movements of the clone are accompanied by the uttering of the demonstrative adverb "là!" (engl.: "there!"). This represents an enhancement of the multimodality of deixis

o no Clone; + Clone gives misleading hints; □ Clone gives correct hints; ∧ Correct hints with delay ;
* *p = 0.05* ; **X** data not valid

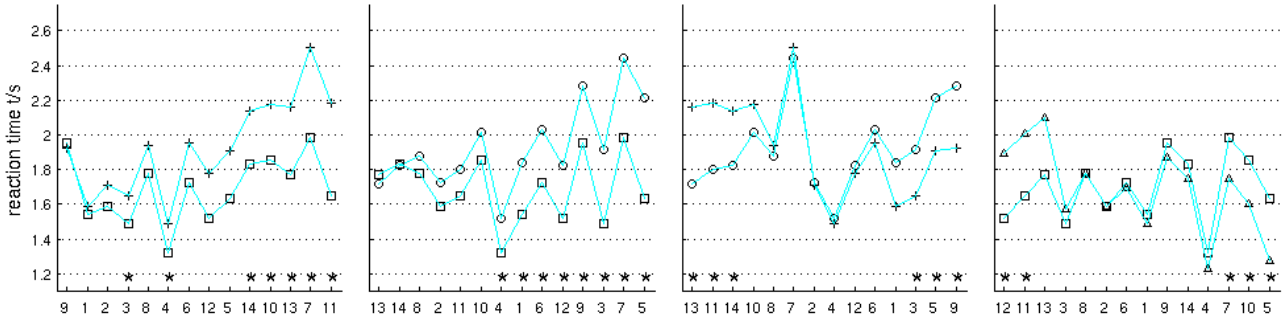

Figure 4: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. The x-coordinate lists the subjects whereas the digit represents the order of participation. Mean reaction times for each user and for each session are displayed together with the statistical significance of the underlying distributions (stars displayed at the bottom when p<0.05).

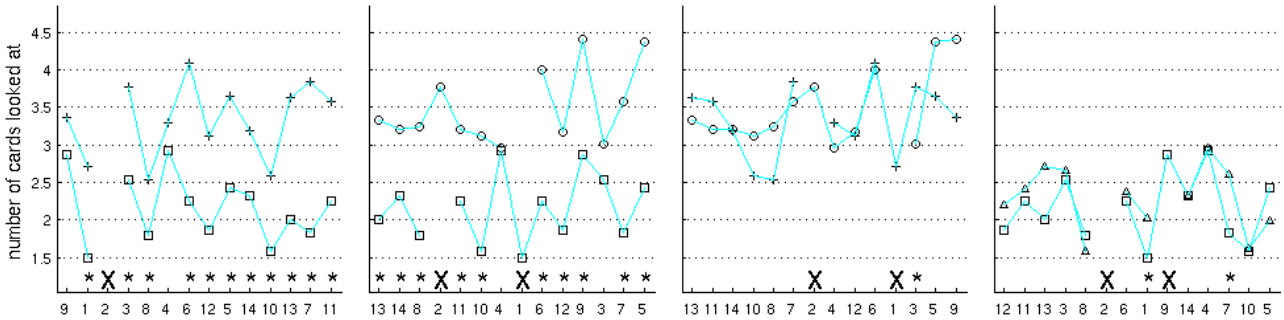

Figure 5. Comparing the number of cards inspected during the search for the correct target card. Conditions compared and order of subjects are the same as in Figure 4.

given by the ECA. No equivalent utterance was present in experiment I where only utterances of motivation were given between turns. To be able to compare the two experimental series, condition 1 with no clone present was replicated for reference. In conditions 2 (wrong cues) and condition 3 (good cues) speech onset is initiated 100ms 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 given during the first series of experiments are now omitted. Condition 4 of experiment I (numbers on cards not shown) is replaced by a condition with correct hints, where an additional delay of 200 ms was introduced between the facial 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. The data collection and treatment was done as described for experiment I. As the influence of the clone when providing bad hints in experiment I was not as strong as expected, it was not clear if the order of presentation might have a major influence. Therefore the conditions are here presented in random orders.

Fourteen users (ten male and four female) took part in this experiment. They range from 21 to 48 years 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.

2.2.2. Results

Before evaluation the reliability of the measured data was examined as described for experiment I. During experiment II only one click error between neighbouring cards occurred (subject six in the condition 2 with misleading hints). As can be seen in Figure 4, the analysis of the reaction time evidences a clear advantage for 7 subjects of 14 during the condition 3 (with correct hints) against the condition 2 (with misleading hints), and for 8 subjects of 14 compared to the condition 1 (without the ECA). These users now gain on average a substantial amount of almost 400 milliseconds (~20% of the mean duration of a turn) at each drawing. The proportion of users benefiting from this advantage and the amount of benefit are both more important than they were in experiment I. Similar to the findings in experiment I, when comparing conditions 1 and 2, 3 subjects show faster reaction times in condition 2 while 3 other subjects just behave the opposite way. When comparing delayed vs. synchronized spoken instructions, 2 subjects show shorter reaction time for condition 3 while 3 show longer reaction times.

For the analysis of the data collected with the eye tracker, subject 2 was completely excluded from evaluation due to insufficient monitoring of eye gaze, subjects 1 and 9 only partly (marked with X above the subject number on the x-coordinate in Figure 5). Analysing the remaining data with an ANOVA for significance at p=0.05 it was found that 11 of the 13 subjects with valid data looked at fewer

cards for condition 3 (correct hints) compared to condition 2 (misleading hints), and 10 of the 12 subjects with valid data compared to condition 1 (without the ECA). On average these users have in fact to inspect 1.5 cards less with a correct gaze than with a wrong or no deictic gaze. These numbers are in line with the data of experiment I. Again data between condition 1 and 2 were statistically significant for only 1 subject. No clear tendency can be reported when considering influence of delay on performance except that the delayed stimuli cause 2 subjects to look at more cards than for the synchronous condition 3.

Answering the questionnaire, 11 of the 14 subjects estimate that they have the best reaction times when correct hints are given by the ECA. Most of the subjects declare 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 reaction time. The movements of the ECA are judged realistic.

3. Discussion and Perspectives

When considering the number of cards inspected and the number of wrong selections in condition 4 of experiment I, the current control and rendering of deixis gestures of the ECA are sufficient to localize objects as long as there is supplementary information available at the target position to take the final decision. Without such additional information the gestures of the ECA seem not to be precise enough to allow a decision 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: eyelids should be notably enlarged to widen the aperture available for the iris. These additional cues may contribute significantly to the estimation of eye direction.

When considering reaction time, 30% to 50% of the users benefit from the assistance given by the ECA. When considering the number of cards a user had to check visually to find the correct target position, the percentage is slightly higher. No major differences are observed between the conditions of misleading and of no assistance given by the ECA. The influence of the ECA when giving misleading hints is however less strong than expected and most users seem able to willingly ignore its gesturing. No clear correlations between the data emerge that would enable a more detailed comprehension of the individual strategies followed to fulfill the task.

Several subjects complained for being disturbed by utterances of motivation in experiment I. Therefore these utterances fail as means to maintain attention and to make the interaction more natural. A more appropriate feedback should be short and clear according to the instruction given to subjects as to react fast. Furthermore it should contribute to attract the attention to the object of interest. The characteristics of experiment II take these complaints into consideration. No subject complains effectively about spoken instructions in experiment II.

The results of experiment II show that the benefit in reaction time from the assistance of the ECA using multimodal deixis could still be improved. An important finding is the reduced number of looked at cards for more than 80% of the subjects. The majority of participants manage to complete the task looking at significantly less cards when the ECA is giving helpful assistance. This means that even if they do not improve their reaction time, the search process is more efficient and probably more relaxed. We conclude that this helpfully diminishes the cognitive load of the task. The answers to the questionnaire confirm this finding as the good ratings for naturalness of the ECA and the preference of the condition where it is giving correct hints are outlined more clearly for experiment II compared with experiment I.

The experimental scenario presented here could probably be further improved by displaying more objects on the screen and using smaller digits. 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. Therefore the benefit of the assistance by the ECA should become more prominent. 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.

4. Conclusions

Our first implementation of a talking head as an embodied conversational agent able to maintain face-to-face interaction with a human interlocutor proves here its capability to direct user attention using multimodal deictic gestures. We demonstrate that users can largely benefit from a very basic implementation even in a rather simple search and retrieval task. ECA guidance results in reduced reaction time and lower cognitive load for the given task. Subjects benefiting from ECA guidance have a substantial gain of 200ms (~10% of a turn) in reaction time and 1.5 cards less to check compared with improper or no guidance in experiment I. The impact could be enhanced (up to 400ms in reaction time) by appropriate and well timed speech commands which especially entail reduction of the cognitive load by reducing the search space and number of matches. We confirm here that the rather modest impact of visual cues found in psychophysical experiments (the 20ms benefit in up/down directions found by Langton and Bruce 1999) is enhanced by more ecological interactions.

We believe that the study, modeling and implementation 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 effective and efficient interaction loop. While most experimental data on speech and gaze examine attention of the listener using recorded videos (Vatikiotis-Bateson, Eigsti et al. 1998), only few experimental data is currently available on gaze patterns when speaking (Vertegaal, Slagter et al. 2001; Argyle and Cook 1976) and during face-to-face interaction.

Such interaction platforms involving actual real-time interaction between a user and autonomous ECA (or semi-autonomous in Wizard-of-Oz experiments animating the ECA with control movements captured on a human operator) is highly valuable for recording characteristic control signals, investigating the influence of embodiment

and assessing the benefit of enhanced strategies on performance, learning and acceptability.

5. 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, Denis Beautemps for helping us with statistical analysis, as well as Matthias Odisio, Pauline Welby and the reviewers of the abstract for their helpful comments. We are also grateful to all participants of the experiments.

6. References

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

Breazeal, C. (2002). *Designing Sociable Robots*, The MIT Press.

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

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.

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

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.

Raidt, S., G. Bailly and F. Elisei (2005). Multimodal face-to-face interaction with a talking face: mutual attention and deixis. *sOc-EUSAI, International Conference on Smart Objects and Ambient Intelligence*, Grenoble - France: 247-252.

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.

Vertegaal, R., R. Slagter, G. van der Veer and A. Nijholt (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *Conference on Human Factors in Computing Systems*, Seattle, USA, ACM Press New York, NY, USA: 301 - 308.