Virtual talking heads and ambiant face-to-face communication

Gérard Bailly, Frédéric Elisei, Stephan Raidt

Summer school 2006
12-13 pages max

Virtual talking heads and ambiant face-to-face communication

Gérard BAILLY, Frédéric ELISEI and Stephan RAIDT
Institut de la Communication Parlée, 46 av. Félix Viallet, 38031 Grenoble - France

Abstract.. We describe here our first effort for developing a virtual talking head able to engage a situated face-to-face interaction with a human partner. This paper concentrates on the low-level components of this interaction loop and the cognitive impact of the implementation of mutual attention and multimodal deixis on the communication task.

Keywords. Talking heads, virtual conversational agents, audiovisual speech synthesis, face-to-face interaction.

Introduction

Building Embodied Conversational Agents (ECA) able to engage a convincing face-to-face conversation with a human partner is certainly one of the more challenging Turing test one can imagine [1]. The challenge is far more complex than the experimental conditions of the Loebner Prize[1] where dialog is conducted via textual information and of the perception test conducted by Ezzat et al [2] where a non interactive ECA was evaluated. Features of situated face-to-face communication including mixed initiative, back channeling, sense of presence, rules for turn taking should be implemented. The interaction loop should not only rely on a convincing animation but also requires a detailed scene analysis: the analysis and comprehension of an embodied interaction is deeply grounded in our senses and actuators and we do have strong expectations on how dialogic information is encoded into multimodal signals.

Appropriate interaction loops have to be implemented. They have to synchronize at least two different perception/action 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. The YTTM model [3] of turn-taking possesses three layered feedback loops (reactive, process control and content). The intermediate process control loop is responsible for the willful control of the social interaction (starts and stops, breaks, back-channeling, etc). In all interaction models, information- and signal-driven interactions should then be coupled to guarantee efficiency, believability, trustfulness and user-friendliness of the information retrieval.

[1] The Loebner Prize for artificial intelligence wards each year the computer programs that delivers the most human-like responses to questions given by a panel of judges over a computer terminal.

The work described here is dedicated to the analysis, modeling and control of multimodal face-to-face interaction between a virtual ECA and a user. We particularly study here the impact of facial deictic gestures of the ECA on user performance in simple search and retrieval tasks.

1. Eye gaze and human-computer interaction

1.1. 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 [4]. In conversation it is involved in the regulation of turn taking, accentuation and organization of discourse [5, 6]. We are also very sensitive to the gaze of others when directed towards objects of interest within our field of view or even outside [7]. In the Posner cueing paradigm [8, 9], 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. [10, 11], who have shown that observers react more quickly when the cue is an oriented face than when it is an arrow. Driver et al. [12] have shown that a concomitant eye gaze also speeds reaction time.

Eye gaze is thus capable of attracting visual attention whereas visual features associated with the objects themselves such as highlighting or blinking are not given so much attention, unless they convey important information for the recognition of a scene. As an example the work of Simons and Chabris [13] suggests that attention is 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. 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 [14] 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 [15] showed 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).

1.2. Interacting with humanoids and avatars

The faculty of interpreting eye gaze patterns of the others is thus crucial for humans and machines interacting with humans. For the "theory of mind" (TOM) as described by Baron-Cohen [16], the perception of gaze direction is an important element of the set of abilities that allow an individual, on the basis of the observation of his behavior, to infer the hidden mental states of another. Several TOM have been proposed [17, 18]. 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 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 [19] 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 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. This work still inspires a lot of works on humanoid robots where complex behaviors emerge from interaction with the environment and users despite the simple tasks to be fulfilled by the robot such as expressing empathy for Kismet [20] or following turn-taking for Robita [21, 22].

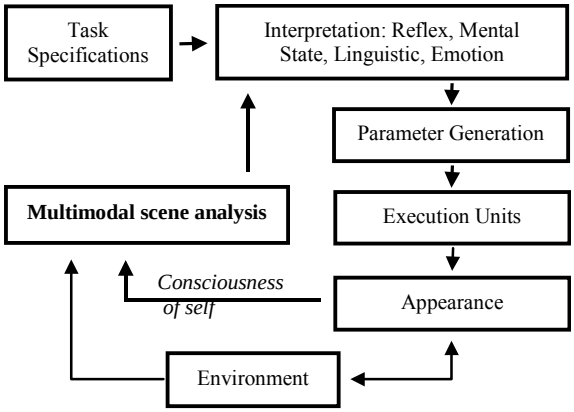


Figure 1: ECA-Human interaction scheme

2. 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 of deriving meaning from the implicit and explicit communicational gestures of a human interlocutor and are also not generally equipped to generate such gestures for communication purposes. Although no real evaluation has been performed, ECA eye gaze can be generated without grounding these gestures in the scene by simply reproducing statistical properties of saccadic eye movements [23]. Note however that Itti et al [24] 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 use of a so-called pertinence map.

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 show 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, 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 and the ECA can thus involve both physical and virtual objects - such as icons surrounding the virtual talking head – in their interaction.

Figure 2: Left. Animated 3D clone with independent head and eye movements. Right: face-to-face interaction platform with a 3D clone

3. Experimental Setup

Our experimental platform aims at implementing context-aware face-to-face interaction between a user and an ECA. Adequate processing of the activity of human partners and changes of ambient environment delivers information to the ECA that should properly condition its behavior.

3.1. Hardware

Sensors. The core element for the experiments described here is a Tobii 1750 eye-tracker[2] that consists of a standard-looking flat screen that discretely embeds infrared lights and a camera (see Figure 2). It monitors at up to 60Hz, the eye gaze of the user whose head can move and rotate freely within a 40cm square cube centered at 60cm away from the screen. A short calibration procedure typically leads to a mean accuracy of 0.5 degrees i.e. 5mm when eyes 50cm away from the 17" screen.

During interaction with the system, the user sits in front of the eye-tracker where our 3D talking head is displayed, as shown in Figure 2. Hardware and software specificities allow the user to interact with the system using eye gaze, mouse and keyboard. The Tobii eye tracker also delivers eye positions relative to its camera and we use this information for head tracking. Additional data input from a video camera and speech recognition is available for other experimental setups.

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

Actuators. The visual representation of our ECA is implemented as the cloned 3D appearance and articulation gestures of a real human [25, 26], (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. 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. Standard graphic hardware with 3D acceleration allows real-time rendering of the talking head on the screen. The ECA has also the gift of 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.

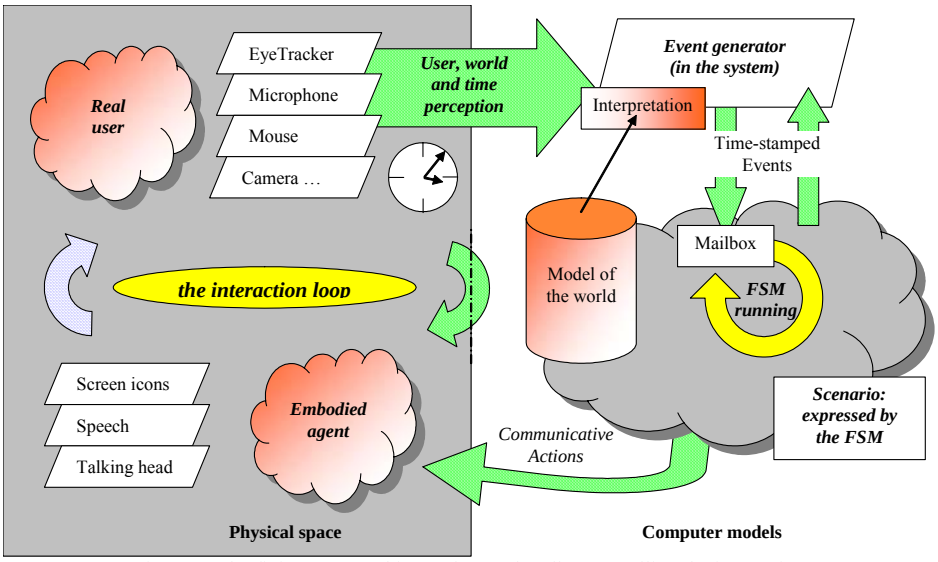


Figure 3: The finite state machine and event handler controlling the interaction.

3.2. Software: scripting multimedia scenarios

The perception-action control loop is described by an event-based language. This language allows the simple description and modification of multimedia scenarios. We developed a corresponding C++ code generator that permits to compile any scenario into an executable binary file. The C++ benefits, like variables, procedural and complex algorithms remain accessible through code inclusion inside any script. Compilation of the generated source code allows accurate recording of the involved events and timings.

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 a constantly refreshed internal model of the space of interaction (see Figure 3). Triggerable areas on the screen, such as selectable icons or parts 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 “quitting 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 which states are waiting on events. Pre-conditions consist of conjunctions or successions of expected multimodal events as for instance recognized keywords, mouse clicks or displacements, eye movements or gaze to active objects. Each of these events is time-stamped.

Pre-conditions can include tests on intervals between time-stamps of events. This allows, for example to associate 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. Post-actions can also generate phantom events, to simulate multimodal input or to share information. These phantom events are potential triggers for pre-conditions inside the following state of the FSM.

4. Experiments

4.1. The interaction scenario

To follow up Langton and Driver experiments on visual orienting in response to social attention [10, 11], 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 multimodal 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 played card.

4.1.1. Design of the card game

The card game consists of eight cards, the numbers of which are revealed once the played card at the lower middle of the screen is selected with a mouse click. The played card has then 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 played card. To anticipate memory effects, the numbers on the cards are shuffled before each turn. The target position is alternated randomly and uniformly distributed amongst the eight possibilities provided that the number of cycles is a multiple of eight. The background color for each position is not changed and thus not correlated with numbers.

4.1.2. Interaction loop

The ECA utter spoken commands and cue directions with an eye saccade combined with a head turn. The ECA alternates between mutual attention and deixis. His gaze and head orientation focus on three regions of interest: the face of the user, his current focus of interest and the target card. In the experiments described below, spoken instructions are not allowed and the ECA gaze alternates between the user’s eyes when fixating his face. When the user speaks, the ECA gaze pattern includes the speaker’s mouth in his attention loop [15].

4.1.3. Experimental conditions

We tested several experimental conditions corresponding to different levels of assistance given by the ECA. Screenshots of the game interface are given in Figure 4. 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 played card on the target position as fast as possible but no strategy is suggested.

4.1.4. Data acquisition

For all experiments the reaction time and the gaze behavior are monitored. The reaction time is 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 is displayed on the monitor with embedded eye-tracking, the visual focus of the user on the screen can be recorded. We thus compute which regions on the screen are looked at and how much time users spend on them. Monitored regions of interest are the eight cards on the sides and the ECA. Eye gaze towards the played card was not monitored, as it was constantly moving during the experiment.

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

4.2. Experiment I : does our clone have cues to direct social attention?

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 conducted by Langton et al [10, 11].

4.2.1. Conditions

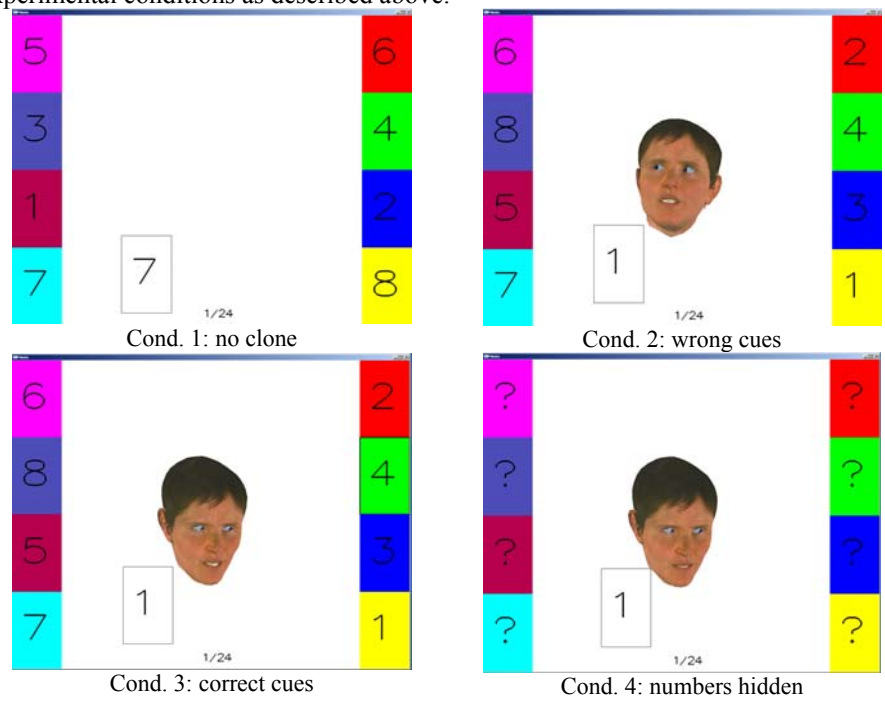
The first series of experiments consists of four different conditions, screenshots of which are displayed in Figure 4. For condition 1, no ECA is displayed. For condition 2, the ECA is visible and provides bad hints: it indicates randomly one of the non-matching positions with a facial gesture as soon as the user selects the played card. In condition 3, it provides good 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 ones to find the correct target position.

In conditions 2,3 and 4, the ECA rewards the user with randomly chosen utterances alternating between motivation and congratulation. The utterances are generated off-line to avoid computation delays.

We have strong expectations about the data to be collected: we expect a negative influence on the test person's performance when the clone gives misleading cues and a positive influence when giving good hints. The condition where no clone is displayed serves as a reference. From the fourth condition, we expect to measure the precision with which the gaze direction of the ECA could be perceived. As we expect the ECA to

strongly influence the users' attention, we keep the order of conditions as described above for all subjects awaiting the mentioned behavior to be free from learning effects.

Ten users (six male and four female) take part in the first series of experiments. Participants range from 23 to 33 years and most were students. All regularly use a computer mouse and none reported vision problems. The dominant eye is the right eye for all but one subject. Each user had to play the game with the four successive experimental conditions as described above.



Cond. 1: no clone

Cond. 2: wrong cues

Cond. 3: correct cues

Cond. 4: numbers hidden

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

4.2.2. Data processing

Reaction-times. Before evaluating the measured reaction times, extreme outliers (distance from median > 5 times inter quartile range) are detected and replaced by a mean value computed from the remaining valid data. Such outliers may be due to the dual screens setup chosen for the experiment. The mouse pointer may in fact leave the screen on one side to appear on the other screen. This happens occasionally when users overshoots the target card and loose time while moving the mouse pointer back into view. The distribution of reaction time is log-normal. We thus analyse the logarithms of the reaction times within each experiment and check with an ANOVA for significance at p=0.05. The significant differences between pairs of distributions are indicated in Figure 5 and Figure 7 with stars.

Number of cards inspected. The number of possible target positions looked at while searching for the correct target position was computed in order to analyse the search strategy of the subjects. This data depends heavily on the data delivered by eye tracker. If less than 60% of all cycles of a condition are not valid (eye gaze not detected, strong deviations between left and right eyes, etc), the data of this condition is entirely

rejected. We characterize the log-normal distribution of the number of cards inspected during a game. To avoid invalid operations (log of 0) an offset of one was added to all observed values before statistical analysis. An ANOVA analysis is then performed on valid data and significant differences between pairs of distributions are indicated in Figure 6 and Figure 8 with stars.

4.2.3. Results

Errors. For the conditions where the target cards are turned up, only one wrong selection occurred (condition 3). The pairing task can therefore be considered as accomplished successfully. Numerous errors occur during condition 4 where users could only rely on the deictic 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). Only one subject accomplished the task without any errors. 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 since the user benefits from additional information to localize the target during the other conditions.

Reaction times. The results are displayed in Figure 5. The mean reaction times are sorted for increasing difference between the compared conditions. Subjects are represented by their order of participation. Significance is marked with a red star above the subject number on the x-coordinate. The diagram shows that 5 out of 10 subjects show significantly shorter reaction times for condition 3 (with correct cues) compared to condition 2 (with wrong cues). Three subjects behave the same compared to the condition 1 (without the ECA). These users gain a substantial amount of 200 milliseconds (~10% of the mean duration of a game) at each drawing. Conditions 1 and 2 lead in fact to similar results: comparing the conditions without the ECA and the ECA giving wrong hints, one subject out of 10 show significant shorter reaction times whereas one show longer ones. 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 most of the subjects.

Number of cards inspected. The results are displayed in Figure 6. Due to the verification of the reliability of the eye tracker data, the data of subject 7 are excluded completely. Probably the subject changed his head position considerably during the experiment. For subject 8 the data of condition four, where no digits appear on the cards, is also excluded. Analysis of the means with an ANOVA at $p=0.05$ evidences a clear advantage for the condition 3 (with correct hints given by the ECA): 6 of the remaining 9 users check significantly fewer cards compared to condition 2 with misleading hints while 5 users also behave the same way when compared to the condition 1 without the ECA. On average these users inspect in fact 1,5 cards less with a correct gaze than with a wrong or no deictic gaze. We interpret this as a clear decrease of cognitive load since less cognitive resources are used for matching cards. Again no obvious interpretation emerges when comparing the conditions 1 and 2. The condition 4 (where the cards are only shown when selected) doubles the number of cards inspected. This is statistically significant for all except one subject.

Questionnaire. 4 of the 10 subjects think they are faster with the helpful assistance of the ECA and prefer this condition when playing.

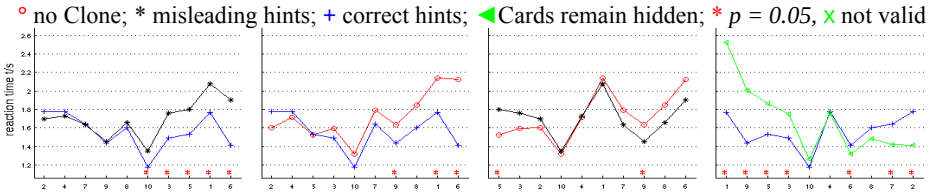


Figure 5: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. Mean reaction times for each user and for each session are displayed together with the statistical significance of the underlying distributions (stars displayed at the bottom when $p<0.05$).

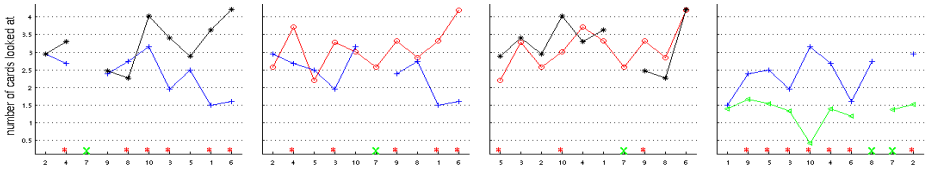


Figure 6: Same as Figure 5 but for nb. of cards inspected. The user order is the one of Figure 5.

4.3. Experiment II: multimodal deixis

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

4.3.1. Conditions

This second series of experiments consists likewise of four different conditions. They resembled the conditions of experiment I. As a major difference the head and gaze movements of the clone are accompanied by the uttering of the demonstrative adverb "là!" (engl.: "there!"). Condition 1 with no clone was replicated for reference. In conditions 2 (wrong cues) and 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 are now omitted. Condition 4 of experiment I is replaced by a condition with correct hints, where an additional delay of 200 ms is introduced between the gestural and the following acoustic deictic gestures in order to comply with data on speech and gesture coordination [27]. We expect this natural coordination to enhance the ability of the ECA to attract user attention. The data collection and treatment is identical to that used for the experiment I.

Fourteen users (ten male and four female) took part to this experiment. They range from 21 to 48 years and most are students. All regularly use a computer mouse and none report vision problems. The dominant eye is the right eye for 8 subjects and the left eye for the other 6 subjects.

As the influence of the clone is not as strong as expected in experiment I when providing bad hints, it was not clear if the order of presentation might have a major influence. Therefore the conditions are here presented in random orders.

4.3.2. Results

Errors. Only one click error between neighbouring cards occurred (subject six in the condition 2 with misleading hints).

Reaction times. The analysis of the reaction time evidences now a clear advantage for 7 subjects out 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 game) at each drawing. The proportion of users benefiting from this advantage and the amount of benefit are both more important than for Experiment I. When comparing the condition 1 versus conditions 2 and 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.

Number of cards inspected. Due to insufficient monitoring of the data collected with the eye tracker, the data of subject 2 is completely excluded from evaluation and data of subjects 1 and 9 only partly. Analysing the remaining data with an ANOVA for significance at p=0.05 it was found that 11 of the 13 subjects with usable data look at fewer cards for condition 3 compared to condition 2, and 10 of the 12 subjects with usable data when 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.

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.

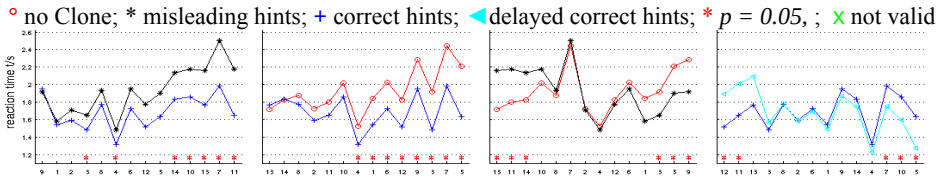


Figure 7: Comparing reaction times for four pairs of conditions. From left to right: condition 2 vs. condition 3; condition 1 vs. condition 3; condition 1 vs. condition 2; condition 4 vs. condition 3. Same conventions as for Figure 5.

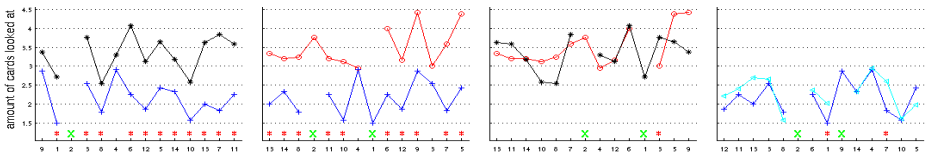


Figure 8. Same as Figure 7 but for number of cards inspected. The same user order is kept as for Figure 7.

5. 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 information available at the target position to take the final decision. Without such additional information the gestures of the ECA are not precise enough to decide 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 eyes 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. When misleading or no assistance is given by the ECA, no major differences are observed. 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 rewarding utterances in experiment I. Therefore these utterances fail as means to maintain attention and 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. No subject complains effectively for 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 (up to 90%). A more 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 can probably 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.

6. Conclusions

Our first implementation of 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 selection task. ECA guidance results in reduced reaction time and lower cognitive load for the given search and retrieval task. Subjects benefiting from ECA guidance have a substantial gain of 200ms (~10% of the task) and 1.5 cards 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 10] 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 [15], only few experimental data is currently available on gaze patterns when speaking [5, 28] and during face-to-face interaction. We recently identify and track facial degrees-of-freedom involved in face-to-face conversation [29] and conduct similar experiments on eye gaze.

Such interaction platforms involving actual real-time interaction between a user and autonomous (or semi-autonomous in Wizard-of-Oz experiments animating the ECA with control movements captured on a human operator) ECA 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.

7. 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 for their helpful comments. We are also grateful to all participants of the experiments.

8. References

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

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

[3] Thórisson, K. (2002) *Natural turn-taking needs no manual: computational theory and model from perception to action*, in *Multimodality in language and speech systems*, B. Granström,

D. House, and I. Karlsson, Editors. Kluwer Academic: Dordrecht, The Netherlands. p. 173–207.
[4] Carpenter, M. and Tomasello, M. (2000) *Joint attention, cultural learning and language acquisition: Implications for children with autism*, in *Communicative and language intervention series. Autism spectrum disorders: A transactional perspective*, A.M. Wetherby and B.M. Prizant, Editors. Paul H. Brooks Publishing: Baltimore. p. 30–54.
[5] Argyle, M. and Cook, M. (1976) *Gaze and mutual gaze*. London: Cambridge University Press.
[6] Kendon, A. (1967) *Some functions of gaze-direction in social interaction.* Acta Psychologica, **26**: p.22-63.
[7] Pourtois, G., Sander, D., Andres, M., Grandjean, D., Reveret, L., Olivier, E., and Vuilleumier, P. (2004) *Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals.* European Journal of Neuroscience, **20**: p.3507-3515.
[8] Posner, M. and Peterson, S. (1990) *The attention system of the human brain.* Annual Review of Neuroscience, **13**: p.25-42.
[9] Posner, M.I. (1980) *Orienting of attention.* Quarterly Journal of Experimental Psychology, **32**: p.3-25.
[10] Langton, S. and Bruce, V. (1999) *Reflexive visual orienting in response to the social attention of others.* Visual Cognition, **6**(5): p.541-567.
[11] 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.
[12] 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.
[13] Simons, D.J. and Chabris, C.F. (1999) *Gorillas in our midst: sustained inattentional blindness for dynamic events.* Perception, **28**: p.1059-1074.
[14] 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.
[15] 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.
[16] Premack, D. and Woodruff, G. (1978) *Does the chimpanzee have a theory of mind?* Behavioral and brain sciences, **1**: p.515-526.
[17] Baron-Cohen, S., Leslie, A., and Frith, U. (1985) *Does the autistic child have a "theory of mind"?* Cognition, **21**: p.37-46.
[18] 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.
[19] Scassellati, B. (2001) *Foundations for a theory of mind for a humanoid robot*, in *Department of Computer Science and Electrical Engineering*. MIT: Boston - MA. 174 pages.
[20] Breazeal, C. (2000) *Sociable machines: expressive social exchange between humans and robots*. Sc.D. dissertation, in *Department of Electrical Engineering and Computer Science*. MIT: Boston.
[21] Fujie, S., Fukushima, K., and Kobayashi, T. (2005) *Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system*. in *Interspeech*. Lisbon, Portugal. p.889-892.
[22] Matsusaka, Y., Tojo, T., and Kobayashi, T. (2003) *Conversation Robot Participating in Group Conversation.* IEICE Transaction of Information and System, **E86-D**(1): p.26-36.
[23] Lee, S.P., Badler, J.B., and Badler, N. (2002) *Eyes alive.* ACM Transaction on Graphics, **21**(3): p.637-644.
[24] Itti, L., Dhavale, N., and Pighin, F. (2003) *Realistic avatar eye and head animation using a neurobiological model of visual attention*. in *SPIE 48th Annual International Symposium on Optical Science and Technology*. San Diego, CA. p.64-78.

[25] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003) *Audiovisual speech synthesis.* International Journal of Speech Technology, **6**: p.331-346.
[26] 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.
[27] Castiello, U., Paulignan, Y., and Jeannerod, M. (1991) *Temporal dissociation of motor responses and subjective awareness.* Brain, **114**: p.2639-2655.
[28] Vertegaal, R., Slagter, R., van der Veer, G., 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.
[29] Bailly, G., Elisei, F., Badin, P., and Savariaux, C. (2006) *Degrees of freedom of facial movements in face-to-face conversational speech.* in *International Workshop on Multimodal Corpora*. Genoa - Italy. p.33-36.