Basic components of a Face-to-Face interaction with a Conversational Agent: Mutual Attention and deixis

Stephan Raidt  
Gérard Bailly  
Frédéric Elisei

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal  
46, av. Félix Viallet, 38031 Grenoble Cedex, France  
{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 the quality of assistance given by the embodied conversational agent was perceived. The experiment showed that users can profit from its presence and its facial deictic cues.

1. Introduction

In an environment that offers ambient intelligence, users interact with an information system that registers and keeps track of the user’s activity and demands. The system is aware of the user(s) and the environmental conditions of the interaction. But how the user(s) can be aware of this awareness, communicate intuitively with the system and consider this intrusion of ubiquitous information in private life as friendly and cooperative? We claim here for the embodiment of ambient intelligence whenever possible. All the unique affordances of the human body are recruited in the metaphor of face-to-face conversation as amplified by Cassell [5]. The issue of embodiment is here crucial: physical embodiment with companion robots guarantees presence and a full range of actions in the physical world, such as poking or grasping various objects as in [16] whereas virtual characters may invade displays (screens, windows, etc) in the vicinity of the user(s), accompany him more easily and provide multimodal information on the same media.

Two complementary perspectives coexist implicitly in the development of Embodied Conversational Agents (ECA). The communicative 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. We do have strong expectations on how dialogic information is encoded into multimodal signals. Of course user’s scene representation, mental state and common belief spaces built when interacting with an ECA 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 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 require prompt reactions to the scene analysis such as involved in eye contact or exogenous saccades. Information- and signal-driven interactions should then both 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 an embodied virtual conversational agent and a user. We particularly study here the impact of mutual attention and deictic gestures in a series of simple deictic tasks.

2. Theoretical Background

2.1. Eye gaze and attention

The cognitive demand of a task has a striking impact on the human analysis of audiovisual 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 [21] 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 [19] 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). The importance of eye gaze to the perceived quality of face-to-face communication was shown by Garau et al. [7]. They compared the user’s appreciation of face-to-face communication via a crossed monitor setup. The four conditions were video display of interlocutor, representation of interlocutor by an avatar with a gaze model for looking at or away from partner, an avatar with random gaze and pure audio communication. Despite the fact that movements of lips of the avatar were estimated from the audio
For a beneficial application of an ECA in face-to-face interaction with a human, the ECA has to be equipped with the means to derive meaning from the implicit and explicit communicational gestures of a human interlocutor. Likewise, it needs the ability to also produce such gestures for communication purposes. To convince a user of its awareness, 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. Giving the 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 the user to understand the system’s actions 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). 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 as well as afford efficient and robust interaction. The motor abilities are principally extended towards speech communication i.e. adapting content and speech style to the pragmatic needs of communication (e.g. confidentiality), to the speaker (notably age and possible communication handicaps) and to the 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 screen objects, such as icons, surrounding the virtual talking head.

\[ \text{Figure 1: ECA-Human interaction scheme} \]

### 3. Interaction with an ECA

For a beneficial application of an ECA in face-to-face interaction with a human, the 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. Giving the 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 the user to understand the system’s actions 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). 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 as well as afford efficient and robust interaction. The motor abilities are principally extended towards speech communication i.e. adapting content and speech style to the pragmatic needs of communication (e.g. confidentiality), to the speaker (notably age and possible communication handicaps) and to the 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 screen objects, such as icons, surrounding the virtual talking head.
eye gaze, a mouse and speech. The 3D clone can look at the user, talk to him, react to his or her eye gaze, and look at screen objects. These elements form the basis of a grounded face-to-face situation.

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

4.1. The Hardware

The flat screen used for the display of the ECA is a Tobii 1750 eye-tracker\(^1\) that discretely embeds infrared lights and a camera. It allows us to detect, at up to 50 Hz, the eye gaze of the user. The user can move and rotate his or her head freely in a fairly unrestricted 3D volume (30 x 15 x 20 cm, centered at 50 cm away from the screen). The average accuracy is 0.5 degrees for measured gaze and 3 milliseconds for the measured time stamps. Each user has to follow a single short calibration procedure before interaction. A standard graphic hardware with 3D acceleration achieves real-time rendering of the talking head on the screen.

4.2. The Talking Head

We use the cloned 3D appearance and articulation gestures of a real human [1, 15] as shown in Figure 3. The eye gaze and head orientation of the clone can be controlled independently from visual speech synthesis. The eye and head movements are controlled by respectively two and three polar angles. Synthetic trajectories were used, but we plan to use data collected from face-to-face interaction recordings. The clone can be programmed to look at the user, at the same spot on the screen as the user as a sign of mutual attention (exploiting the eye tracker data) or at 2D objects on the screen intended to actively direct the user's attention. The vergence of the eyes is controlled and provides a crucial cue for inferring spatial cognition. The audiovisual messages can either be created from audio recordings of human speaker or be synthesized from text input.

\(^{1}\) Please consult http://www.tobii.se/ for technical details.

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

5. The interaction experiment

5.1. The Experimental Scenario

To follow up the findings of Langton and Driver about the special ability of 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 reveal numbers once the play card at the lower middle of the screen is selected by a mouse click made by the user. The card is then attached to the mouse cursor and the user has to drop the play card with a click 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.

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 and therefore 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 should find the matching card and random congratulations are uttered upon success.
5.3. Analysis of the Experiment

5.3.1. Expected Outcome

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

5.3.2. 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 showed that almost all wrong choices were made between neighboring cards. Apart from the constrictions of 3D rendering on a screen, this may be due to the momentarily basic implementation of synchronization between gaze and head orientation and their trajectories that are not derived from measured data yet. Another explanation may be the unanimated eyelids [8, 9].

During the other experimental conditions (1 to 3) only one wrong choice of target position 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 with a mouse click.

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 it can not be expected to measure improvements in performance of a magnitude as found by Langton et al. [8, 9]. The cognitive demands are higher and the dragging of the play card with the mouse and selection by mouse click are quite complex movements that demand precise coordination. The task offers thus various possibilities for time delays.

The number of visited target positions was considered as another possibility for performance evaluation. When the clone indicates the target position this should ideally reduce to one essay. Accounting for the difficulties to derive precise directions from the clone’s eye gaze as discussed before, two essays might be necessary. Searching for the target position without using the assistance by the clone, the user may have to look at up to 8 cards before finding the correct one. Including repetitions this may even be more. The number of visited cards and therefore both possibilities of performance evaluation are highly depending on chance that cannot be controlled.
For the evaluation of the reaction time the means of the different conditions are compared. The measurements show that five subjects completed the task significantly faster (p=0.05) in the condition with the clone giving correct cues compared to the conditions with the clone giving misleading cues (Figure 5 left). Comparing the condition with correct cues to the condition without the clone, a significant benefit in reaction time could be found for four subjects (Figure 5 middle). It is remarkable that the subjects that are influenced by the cues of the clone are not always the same ones over conditions. The misleading cues of the clone resulted in significantly longer reaction time for only one subject whereas another subject improved for this condition due to a learning effect (Figure 5 right). A negative influence of the misleading cues compared to the experimental condition without the clone can therefore not clearly be stated.

For an interpretation of the gaze data, measured with the eye tracker, 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 could not be surveyed. However, we can not be sure that effects of periphery view were of no major importance to user performance.

A crucial correlation was expected between performance time and the time spent looking at the clone. However this could not be found as a general influence. It seems that the test persons developed different strategies to complete the task. This impression is fostered by the subject’s 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 who profited from the ECA both in performance time and number of cards visited. This agrees with the subject’s personal estimation (see Figure 7 right).

Other subjects chose exactly the opposite possibility, spending little time on the clone and rating its influence accordingly as minor. One subject even reported after the experiment that he deliberately tried to ignore the clone in order to outperform it. All subjects agreed in their estimation that they were little influenced by the clone giving misleading cues (see Figure 7 left). The 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 5: Performance time means across subjects: Each of the three figures highlights the difference in performance time between two conditions. Subjects are sorted by increasing time-difference between the respective conditions. 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 (°). (p=0.05).

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 spent looking at the clone during these conditions. Right: Number of cards visited before selecting target position (condition 1 to 3 from top bottom).
5.4. 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.

Since we expected a strong impact of the clone’s presence, we structured the experiment with the four conditions of different levels of help and guidance described above. It might be interesting to see if a different order of conditions (e.g. clone giving correct hints first) would change the attentiveness of users to the clone’s gaze. Nagao and Takeuchi [11] found that the order of presentation of the first stimulus can influence the results of the entire experiment. They tested the success of interaction with an information system using an animated face with facial expression (eye gaze was not included) or an unanimated one.

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, there is either a major influence of periphery view, or the mere presence of a clone already motivates the user.

Despite the limitations of a display without stereovision and the lack of spatial accuracy in interpreting the clone’s 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 plan to use our interaction system to record eye gaze of real humans in the symmetrical situation.

6. 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.

The experiment presented here is very complex. The card game scenario leaves a lot of freedom to the users and they apparently developed different strategies to fulfill the task. Therefore it is obvious that the benefit in reaction time of 20ms for a total reaction time of around 300ms as achieved in the experiments by Langton et al. [8, 9] cannot be assessed here. However, almost all test persons effectively profit from the presence of the clone.

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 [20]. 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.

7. Acknowledgments

We gratefully acknowledge the patience of Hélène Løvenbruck, 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.

References

Joint sOc-EUSAI conference
