Evaluation of an Embodied Conversational Agent with Affective Behavior

Junichiro Mori, Helmut Prendinger, and Mitsuru Ishizuka
Department of Information and Communication Engineering
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: {jmori,ishizuka}@miv.t.u-tokyo.ac.jp; prendinger@acm.org

Abstract

The aim of the experimental study described in this paper is to investigate the effect of an Embodied Conversational Agent (ECA) on the affective state of users. The agent expresses affect by verbal and nonverbal behaviors such as linguistic style and gestures. In this study, we focus on users' emotional state that is derived from physiological signals of the user. Our results suggest that an agent with appropriate verbal and nonverbal behaviors may decrease the intensity of users' negative emotions.

Keywords. Affective behavior, verbal and nonverbal behavior, evaluation method

1 Introduction

In human–human interaction, nonverbal behaviors such as gesture and posture support the meaning of the linguistic message, and convey important information about personality and emotional state. For a person's individuality, his or her linguistic style and way to gesture are fundamental factors. Since Embodied Conversational Agents (ECAs) are mostly designed as synthetic counterparts of 'real' humans, we may assume that linguistic style and gestures are key factors for them to be perceived as individuals.

In this paper, we do not directly address the issue of individuality and associated questions as to its representation and parameters. However, our experiment considers the cultural background of the target users by using an ECA that is equipped with gestures specific to that culture. Specifically, our agent shows behaviors whose meaning is readily understood by Japanese, who constitute the intended audience of our interaction scenario.

In order to measure the effect of the interaction with an ECA on user emotions, we take physiological signals from the user. Unlike standard evaluation methods such as questionnaires, the use of physiological data may support a more accurate assessment of the affective state of users (Schreirer et al., 2002; Picard, 1997). In particular, the recorded history of users' bio-signals allows to precisely relate emotion occurrence with the (user-computer) interaction state. Furthermore, using both bio-signals and questionnaires enables to detect possible discrepancies between the interaction as perceived by the user and the user's physiological state.

2 The Agent

In our experiment, a mathematical quiz game (described below), we use a 2D cartoon-style character called "Shima agent" as our ECA.

The Shima agent is controlled by the Microsoft Agent package that provides the following features:

- Controls to animated facial and body gestures.
- A Text-to-Speech (TTS) engine for synthetic speech which is also displayed in a balloon adjacent to the agent.
- A limited form of synchronizing gestures with speech (typically co-occurrence or overlapping of speech output and gesture).

The Shima agent is designed with behaviors of a typical Japanese businessman and hence the agent's actions are familiar to Japanese and easily understood. For example, one animation shows the agent bowing, a gesture which Japanese people perceive as a signal of the interlocutor's apology.

In our actual implementation, the actions of Shima are specified by using MPML (Multi-modal Presentation Markup Language), a scripting language that allows for easy handling of the verbal and nonverbal control of Microsoft Agent based characters (Ishizuka et al., 2000). The agent is embedded in a web browser environment, where its behavior can be triggered by the user's mouse of keyboard input.

3 Design of the Experiment

We implemented a simple mathematical quiz game where subjects are instructed to sum up five successively displayed numbers and are then asked to subtract the i-th number of the sequence ($i \leq 4$). Subjects compete for the best score in terms of correct answers and time (a monetary award was given for both participation and best score). Subjects were told that they would interact with a prototype interface that might still contain some bugs. Before game start, some quiz examples that explain the game was given to subjects. This period also serves to collect physiological data of subjects that are needed to normalize data obtained during game play. In six out of a total of thirty quiz questions, a delay was inserted before showing the 5th number. The delay, between 6 and 14 sec. (9 sec. on average), is assumed to induce frustration as the subjects' goals

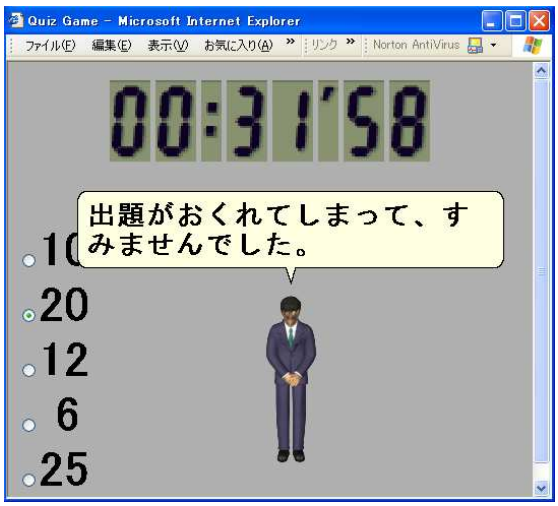

Figure 1: The "Shima" character apologizes.

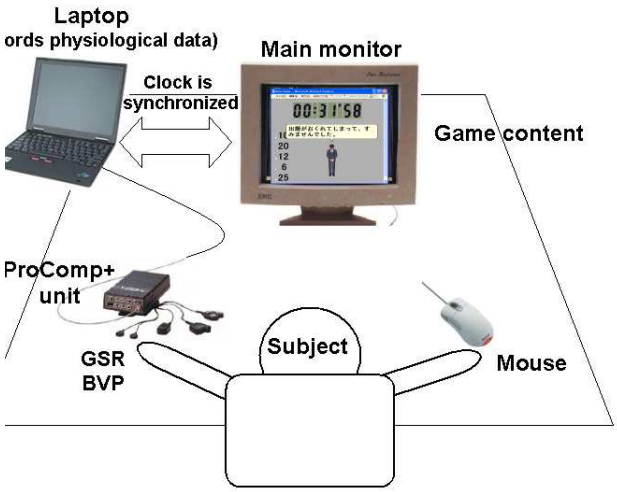

Figure 2: Experimental setup.

of giving the correct answer and achieving a fast score are thwarted, called 'primary frustration' in behavioral psychology (Lawson, 1965).

In the experiment, subjects were twenty male students, all of them native speakers of Japanese. We randomly assigned subjects to one of two versions of the game (ten in each version), the *affective version* or the *non-affective version*.

- *Affective version*. Depending on whether the subject selects the correct or wrong answer from the menu displayed in the game window, the agent expresses 'happy for' and 'sorry for' emotions both verbally and nonverbally.

  If a delay in the game play happens, the agent expresses empathy for the user after the subject answers the question that was affected by the delay. Note that the apology is given *after* the occurrence of the delay, immediately after the subject's answer.

- *Non-affective version*. The agent does not give any affective feedback to the subjects. It simply replies "right" or "wrong" to the user's answer and does not comment on the occurrence of the delay.

In the affective and non-affective version, the verbal and nonverbal behaviors of our ECA differ with respect to linguistic style and gesture.

- *Linguistic style*. In the non-affective version, the agent utters "seikai" or "fu-seikai" (romanized Japanese for English "right" and "wrong"). In the affective version on the other hand, the agent replies by saying "seikai-*desu*" and "fu-seikai-*desu*", which contain the postfix "desu". Attaching "desu" indicates a more formal usage of Japanese language and gives a soft and polite impression to the utterance.

- *Facial and bodily gestures*. In the affective version, the agent expresses a 'happy for' (the subject's right answer) emotion or a 'sorry for' (the subject's wrong answer) emotion by displaying a smiling face, and hanging shoulders together with sad facial expression, respectively. On the other hand, in the non-affective version, the agent does not show any affective gestures.

  Moreover, when a delay in asking the query happens, the agent expresses empathy for the subject by verbal and nonverbal means. Fig. 1 shows the agent displaying a gesture that Japanese people perceive as a sign of apology (bowing with both hands in the lab), and says (in romanized Japanese): "Shutsudai ga okurete shimatte sumimasen deshita". The English translation of this apology is "I apologize that there was a delay in posing the question". It is important to note that both the verb conjugations and words used in this sentence convey the speaker's politeness and sorriness. Besides indicating the speaker's respect for the addressee, politeness levels in Japanese can be used to empathize the speaker's emotional state, such as sorriness. By contrast, the agent ignores the occurrence of the delay in the non-affective version.

Subjects are attached to two types of sensors on the first three fingers of their non-dominant hand (see Fig. 2) that measure skin conductivity (SC) and heart rate (HR). Signals are recorded with the ProComp+ unit and visualized using Thought Technology software.

- The galvanic skin response (GSR) signal is an indicator of SC. It has been shown that SC varies linearly with the overall level of arousal and increases with anxiety and stress (see the discussions in Picard (1997, p. 162) and Healey (2000, p. 25, 40)).

- The blood volume pressure (BVP) signal is an indicator of blood flow. (HR was automatically calculated from BVP with our software.) BVP increases with negatively valenced emotions such as fear and anxiety, and decreases with relaxation (Picard (1997, p. 162) and Healey (2000, p. 27)).

In order to obtain named emotions from signals, SC and HR can be mapped to the emotion model of (Lang, 1995) which shows that emotions can be located as coordinates of affective valence and arousal in a two-dimensional space. In our experiment, however, BVP data could be taken reliably in only six out of twenty cases.[1] In effect, we could

[1] In particular, our method to gather BVP data was unreliable, and not the sensor data.

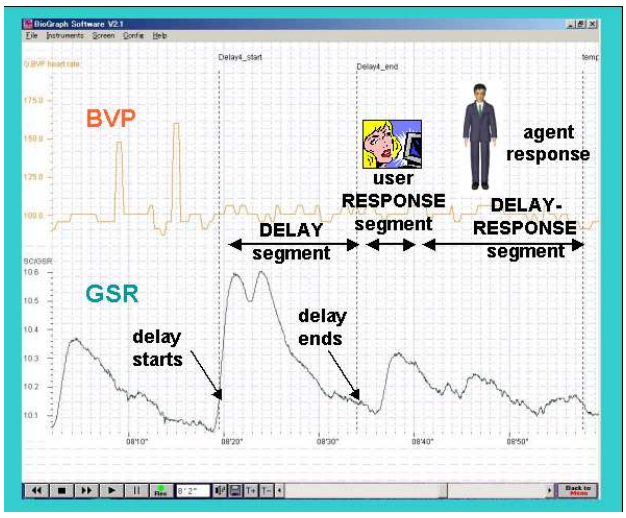

Figure 3: Three relevant segments of the game flow, exemplified by the bio-signals of one user.

not use Lang's model and had to restrict our findings to different levels of stress rather than specific emotions.

In order to show the effect of the agent's behavior, we have been interested in three specific segments (see Fig. 3):

- The DELAY segment refers to the period after which the agent suddenly stops activity while the question is not completed until the moment when the agent continues with the question.
- The DELAY-RESPONSE segment refers to the period when the agent expresses gesture concerning the delay, or ignores the occurrence of the delay – which follows the agent's response (regarding the correctness of the answer) to the subject's answer.
- The RESPONSE segment refers to the agent's response to the subject's correct or wrong answer to the quiz question.

4 Results of the Experiment

The first observation relates to the use of delays in order to induce frustration and stress in subjects. All eighteen subjects showed a significant rise of SC in the DELAY segment, indicating an increased level of arousal.

Our general hypothesis about the positive effect of embodied agents with affective behavior on users can be divided into three specific hypotheses.

- Hypothesis 1 (*Empathy*): SC is lower when the agent shows empathy after a delay occurred, than when the agent does not show empathy.
- Hypothesis 2 (*Affective feedback*): When the agent tells whether the subject's answer is right or wrong, SC is lower in the affective version than in the non-affective version.
- Hypothesis 3 (*Score*): Subjects interacting with the affective version score better in the game than subjects interacting with the non-affective version.

Table 1: Mean scores for questions about interaction experience in affective (*A*) and non-affective (*NA*) game version. Ratings range from 1 (disagreement) to 10 (agreement).

Question	*NA*	*A*
I experienced the quiz as difficult.	7.5	5.4
I was frustrated with the delays.	5.2	4.2
I enjoyed playing the quiz game.	6.6	7.2

To support Hypothesis 1 (empathy), we calculated the differences between the mean values of SC in the DELAY and DELAY-RESPONSE segments for each subject. (The data of two subjects of the non-affective version were discarded because of extremely deviant values.) In the non-affective version (no display of empathy), the difference is even negative (mean $= -0.08$). In the affective version (display of empathy), SC decreases when the character responds to the user (mean $= 0.14$). In the following, the α level is set to 0.05. The t-test (two-tailed, assuming unequal variances) showed a significant effect of the character's affective (emphatic) behavior as opposed to non-affective behavior ($t(16) = -2.47$; $p = 0.025$). This result suggests that an embodied agent expressing empathy may undo some of the frustration (or reduce stress) caused by a deficiency of the interface.

Hypothesis 2 (affective feedback) compares the means of SC values of the RESPONSE segments for both versions of the game (the agent responses of all queries are considered here). However, the t-test showed no significant effect ($t(16) = 1.75$; $p = 0.099$). When responding to the subject's answer, the agent's affective behavior has seemingly no major impact.

Hypothesis 3 (score) could not be supported in the present game. The average score in the affective version was 28.5 (from 30 answers), and 28.4 in the non-affective version. We may interpret this result in the light of the findings in (van Mulken et al., 1998), who show that interface agents have no significant effect on objective measures (in their case, comprehension and recall). Another reason might be that the mathematical task was too simple, so that the agent's behavior had no effect on game performance.

In addition to taking subjects' physiological data we asked subjects to fill out a short questionnaire after they completed the quiz. Table 1 shows the mean scores for some questions. None of the differences in rating reached the level of significance. Only the scores for the first question suggest a tendency ($t(17) = 1.74$; $p = 0.1$) somewhat related to the one observed by (van Mulken et al., 1998), namely, that a character may influence the subjects' *perception* of difficulty. This indicates that affective behavior influences the subjects' impression of difficulty. In their experiment though, van Mulken and coworkers compare "persona" vs. "no-persona" conditions rather than "affective persona" vs. "non-affective persona" conditions.

The scores for the second question indicate that subjects underestimate the extent to which they were frustrated in both versions of the game. Since the GSR signal significantly increased during the delay period, subjects were obviously frustrated during those periods which is not reflected in their answer to this question (non-extreme scores)

in the questionnaire. This highlights the importance of using a more objective evaluation method, such as physiological user data assessment, which may detect user experiences that can hardly be revealed by using only questionnaires. Furthermore, bio-signal assessment is not affected by a well-known problem of the standard questionnaire method, namely that subjects answer the way they believe the experimenter expects them to answer.

The scores for the third question are slightly in favor of the affective-version but, as said above, not significantly so.

Although the obtained results are still somewhat restricted, we believe that embodied conversational agents with affective behavior have the potential to alleviate user frustration similar to human interlocutors, and the assessment of user's physiological data is an adequate method to show the effects of agents.

5 Current and Future Work

We currently extend our work to process physiological data in real time and base the agent's behavior on the current emotional state of the user. As shown in the experiment, it is possible to assess the user's arousal by taking physiological signals. But the great challenge of online emotion recognition is to integrate information about the user's cognitive state (goals, beliefs, standards) (Ortony, 2003) and physiological user data. An integrated approach may increase the reliability of inferring emotions and could be used to distill named emotions as opposed to coordinates of the valence-arousal axis. Our goal is to develop an adaptive ECA interface (Conati, 2002; Hudlicka & McNeese, 2002). In order to realize such an adaptive interface, we are currently developing a decision network that achieves tailored agent reactions depending on more features of the interaction, such as user goals and personality, and interaction task.

6 Conclusions

The aim of the experimental study described in this paper is to show the impact of an Embodied Conversational Agent on the physiological and (derived) emotional state of users. The agent expresses verbal behaviors (synthetic speech and linguistic style) and nonverbal behaviors (facial and bodily gestures). We focus on users' emotional state that is derived from physiological signals of the user. Our results suggest that an ECA with appropriate verbal and nonverbal behaviors may positively affect users' emotional state.

This research intends to evaluate recent efforts to generate (and script) ECAs with life-like behavior (see Prendinger & Ishizuka (2003) for an up-to-date collection of character scripting languages and applications). Certainly, the design of the quiz game was driven by considerations of evaluating specific aspects of ECA behavior, such as affective behavior and emphatic feedback to 'frustrating events', rather than ECAs in general. However, it is reasonable to assume that even (or especially) more complex applications will likely frustrate the user at some point, and the strategy to decrease user stress discussed in this paper will be readily applicable.

Although the described experiment was originally not designed to filter out agent features and parameters for individuality, it suggests the positive impact of verbal and nonverbal emotion and empathy expression as opposed to the lack of those behaviors.

Acknowledgements

This research is supported by the JSPS Research Grant (1999-2003) for the Future Program ("Mirai Kaitaku").

References

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, *16*, 555–575.

Healey, J. A. (2000). *Wearable and Automotive Systems for Affect Recognition from Physiology*. PhD thesis, Massachusetts Institute of Technology.

Hudlicka, E. & McNeese, M. D. (2002). Assessment of user affective and belief states for interface adaption: Application to an Air Force pilot task. *User Modeling and User-Adapted Interaction*, *12*, 1–47.

Ishizuka, M., Tsutsui, T., Saeyor, S., Dohi, H., Zong, Y. & Prendinger, H. (2000). MPML: A multimodal presentation markup language with character control functions. In *Proceedings Agents'2000 Workshop on Achieving Human-like Behavior in Interactive Animated Agents* (pp. 50–54).

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50(5)*, 372–385.

Lawson, R. (1965). *Frustration: The Development of a Scientific Concept*. New York: MacMillan.

Ortony, A. (2003). On making believable emotional agents believable. In R. Trappl, P. Petta & S. Payr (Eds.), *Emotions in Humans and Artifacts*. The MIT Press.

Picard, R. W. (1997). *Affective Computing*. The MIT Press.

Prendinger, H. & Ishizuka, M. (Eds.). (2003). *Life-like Characters. Tools, Affective Functions and Applications*. Cognitive Technologies. Springer Verlag. To appear.

Schreirer, J., Fernandez, R., Klein, J. & Picard, R. W. (2002). Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, *14*, 93–118.

van Mulken, S., André, E. & Müller, J. (1998). The Persona Effect: How substantial is it? In *Proceedings Human Computer Interaction (HCI-98)* (pp. 53–66). Berlin: Springer.