Does Non-Verbal Behavior of an Embodied Agent Matter?

Helmut Prendinger
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
helmut@nii.ac.jp

Chunling Ma, Junichiro Mori, Mitsuru Ishizuka
Dept. of Information and Communication Eng.
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{macl,jmori,ishizuka}@miv.t.u-tokyo.ac.jp

Abstract

This paper reflects on some of our research on embodied agents from the viewpoint of non-verbal behavior. In previous studies we aimed to investigate the utility of embodied interface agent by applying novel evaluation methods. One study tracks bio-signals in order to evaluate the impact of affective agent behavior on the stress level of users [13]. In another study, users' eye movements were recorded to demonstrate the benefit of an embodied interface agent as a navigational guide [6]. Since the encouraging results of these two studies mostly relied on the use of non-verbal agent behaviors, including non-verbal means to express affect and empathy as well as deictic gestures, we want to answer the question posed in the title of the current paper in the affirmative: non-verbal agent behavior matters for effective human–computer interaction.

Keywords: Life-Like Characters, Utility, Bio-signal and Eye Movement Tracking, Non-Verbal Behavior

1. Introduction and Motivation

Non-verbal behavior is an essential part of human-to-human communication and social experience. This fact is also gaining increasing importance for human-computer systems that strive to capitalize on the naturalness and efficiency of human conversation. Since non-verbal behavior requires a greater bandwidth than e.g. textual messages, researchers started to propose different types of embodiment to improve human–computer interaction. The embodiment of a computer is either realized by means of a synthetic interface agent [2, 12] or a physical robotic agent [1].

In this paper, we will describe two of our previous studies that used embodied (synthetic) agents with non-verbal behavior in order to improve human–computer interaction. In the *first study*, we investigate the impact of affective behavior of an embodied agent – assuming the role of a virtual quiz master in a mathematical game – on users' affective states, which are derived from physiological data [13]. The main hypothesis of this study can be formulated as: If an embodied interface agent provides affective (verbal and non-verbal) feedback to the user, it can effectively reduce user stress. It is well known that physiological signals (or bio-signals) such as skin conductance, muscle tension, and heart rate provide important information regarding the intensity and quality of a person's experience, and can thus be used to infer a user's emotion or affective state (see, e.g. [10]). By tracking users' bio-signals, we may find answers to questions such as "Does the interaction with an embodied agent have an influence on users' affective state?" or "Which particular verbal and non-verbal behaviors of an embodied agent cause frustration or relaxation in the user?"

In the *second study*, we will track and analyze eye movements and bio-signals while users are following the web page based presentation of different rooms of an apartment [6]. Three types of presentations will be contrasted:

- An embodied interface agent presents the apartment using speech and gestures;
- The apartment is presented by means of a text-box and read out by speech;
- The presentation is given by speech only.

Although gaze point and focus of attention are not necessarily always identical, a user's eye movement data provide rich evidence of the user's visual and (overt) attentional processes [3]. The movements of the human eye can be used to answer questions such as "To which part of the embodied agent (face or body) is the user attending to?" or "Can the agent's verbal or gestural behavior direct the user's focus of attention?"

In our experience, bio-signal and eye movement data can offer valuable information relevant to the utility of embodied agents and the usability of interfaces employing those agents. The tracking of the physiological activity of users

lends itself to reliably capturing the moment-to-moment experience of interface users, which is hard to assess by using post-experiment questionnaires.

The rest of the paper is organized as follows. The two following sections report on our studies on embodied behaviors with affect display and empathy (Sect. 2) and deictic gestures (Sect. 3). Section 4 concludes the paper.

2. Displaying Affect and Empathy

2.1. Method

2.1.1. Theory and Game Design We implemented a simple mathematical quiz game where subjects are instructed to sum up five consecutively displayed numbers and are then asked to subtract the i-th number of the sequence ($i \leq 4$). The instruction is given by the "Shima" character, an animated cartoon-style 2D agent, using synthetic speech and appropriate gestures. The numbers are also displayed in a balloon adjacent to the agent. Subjects compete for the best score in terms of correct answers and time. Subjects were told that they would interact with a prototype interface that may still contain some bugs. This warning was essential since in some quiz questions, a delay was inserted before showing the 5th number. The delay was assumed to induce frustration as the subjects' goals of giving the correct answer and achieving a fast score are thwarted.

In order to measure user frustration (or stress), we took users' galvanic skin response (GSR) signal which is an indicator of skin conductance.[1] It has been shown that skin conductance varies linearly with the overall level of arousal and increases with anxiety and stress (see Picard [10]).

2.1.2. Subjects and Design Participants of the experiment were twenty male students of the School of Engineering at the University of Tokyo, on average 24 years of age, and all of them native speakers of Japanese. According to the independent variables, *affective* vs. *non-affective* feedback of an embodied agent, two versions of the quiz game have been prepared:

- *Affective version*. Depending on whether the subject selects the correct or wrong answer from the menu displayed in the game window (see the numbers in Fig. 1), the character expresses 'happy for' and 'sorry for' emotions both verbally and non-verbally, e.g., by "smiling" (for happiness) and "hanging shoulders" (for sorriness). When a delay in the game flow happens, the character expresses empathy for the subject after the subject answers the question that was affected by the delay (see Fig. 1).
- *Non-affective version*. The agent does not give any affective feedback to the subjects. It simply replies "right" or "wrong" to the answer of the subjects. If a delay happens, the agent does not comment on the occurrence of the delay, and simply remains silent for a short period of time.

If a delay occurs (in the affective version), the agent expresses empathy to the subjects by displaying a gesture that Japanese people will easily understand as a signal of the interlocutor's apology (see Fig. 1), and uttering: "I apologize that there was a delay in posing the question" (English translation). Note that the apology is given *after* the occurrence of the delay, immediately after the subject's answer (and not during the delay period).

In order to show the effect of the agent's behavior on the physiological state of subjects, we consider specific segments. (i) 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; (ii) the DELAY-RESPONSE segment refers to the period when the agent expresses empathy 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; (iii) the RESPONSE segment refers to the agent's response to the subject's correct or wrong answer to the quiz question.

2.1.3. Procedure and Apparatus The subjects were recruited directly by the experimenter and offered 1,000 Yen for participation, and additionally 5,000 Yen for the best score. Subjects have been randomly assigned to one of the two versions of the game. The experiment was conducted in Japanese, and lasted for about 25 minutes (15 min. for game play; 10 min. for experimenter instructions, attaching the sensors, etc). Subjects came to the testing room individually and were seated in front of a computer display, keyboard, and mouse. After briefing the subjects about the experiment and asking them to sign the consent form, they were attached to GSR and blood volume pulse sensors on the first three fingers of their non-dominant hand.

Before subjects actually started to play the game, the character shows some quiz examples that explain the game. This period also serves to collect physiological data of subjects that are needed as a baseline 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 duration of delays was 6–14 seconds. While subjects played the game the experimenter remained in the room and monitored their physiological activity on a laptop computer. The experimenter and laptop were hidden from the view of the subjects. After the subjects completed the

1 We also recorded subjects' blood volume pulse (BVP) signal from which the heart rate of subjects can be calculated. Unfortunately, the low reliability of our method used to gather the BVP signal precluded its consideration in the analysis.

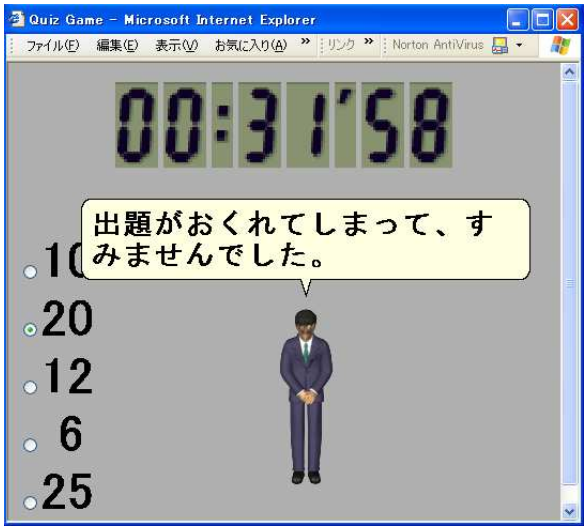

Figure 1. Shima character: "I apologize that there was a delay in posing the question."

quiz, the sensors have been removed from their hand, and they were asked to fill out a short questionnaire, which contained questions about the difficulty and their impression of playing the game. Finally, subjects were told to keep checking a web page that will announce the best score.

The game was displayed on a 20 inch color monitor, running Internet Explorer with browsing buttons deactivated. The Microsoft Agent package [7] was used to control agent animations and synthetic speech. Two flat speakers produced the sound. Physiological signals have been recorded with the ProComp+ unit and visualized with BioGraph2.1 software (both from Thought Technology Ltd. [14]).

2.2. Results

The first observation relates to the use of delays in order to induce stress in subjects. All eighteen subjects showed a significant rise of skin conductance in the DELAY segment, indicating an increased level of arousal. The data of two subjects of the non-affective version were discarded because of extremely deviant values. In the following, the confidence level α is set to 0.05.

The general hypothesis about the positive effect of embodied agents with affective behavior on a subjective measure, here the users' stress level, can be divided into two specific hypotheses (*Empathy* and *Affective Feedback*).

- *Empathy* Hypothesis: Skin conductance (stress) is lower when the character shows empathy after a delay occurred, than when the character does not show empathy.
- *Affective Feedback* Hypothesis: When the character tells whether the subject's answer is right or wrong, skin conductance is lower in the affective version than in the non-affective version.

To support the Empathy Hypothesis, the differences between the mean values of the GSR signal (in micro-Siemens) in the DELAY and DELAY-RESPONSE segments have been calculated for each subject. In the non-affective version (no display of empathy), the difference is even negative (mean $= -0.08$). In the affective version (display of empathy), GSR decreases when the character responds to the user (mean $= 0.14$). The t-test (two-tailed, assuming unequal variances) showed a significant effect of the character's emphatic behavior as opposed to non-affective behavior ($t(16) = -2.47$; $p = 0.025$). This result suggests that an animated agent expressing empathy may undo some of the frustration (or reduce stress) caused by a deficiency of the interface.

The Affective Feedback Hypothesis compares the means of GSR values of the RESPONSE segments for both versions of the game. Note that the character responses of all queries, not only the queries affected by a delay, 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, affective behavior of the character has seemingly no major impact on subjects' skin conductance.

3. Deictic Gestures

3.1. Method

3.1.1. Experimental Design A presentation of an apartment located in Tokyo has been prepared using a web page based interface [15]. The apartment consists of six rooms: living room, bedroom, dining room, den, kitchen, and bathroom. Views of each room are shown during the presentation, including pictures of some part of the room and close-up pictures of e.g. a door handle or sofa. Three versions of the apartment show have been designed for the experiment:

- *Agent (& speech) version.* A character called "Kosaku" presents the apartment using synthetic speech and deictic facial and hand gestures (see Fig. 2). The character is controlled by a version of MPML [11].
- *Text (& speech) version.* The presentation content of each scene is displayed by a text box and read out by Microsoft Reader.
- *Voice (only) version.* Synthetic speech is the only medium used to comment on the apartment.

The main purpose of programming the Text and Voice versions was to provide interfaces that can be compared to the Agent version in terms of the user's eye movements and

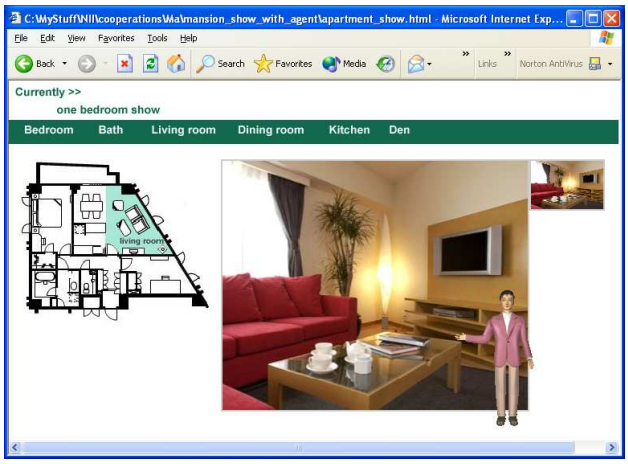

Figure 2. A embodied agent presents the living room of the apartment.

physiological activity. (A condition "Agent without deictic gestures" was not found interesting when designing the experiment.) The same type and speed of (synthetic) voice was used in all versions.

3.1.2. Subjects Fifteen subjects (3 female, 12 male), all students or staff from the University of Tokyo, participated in the study. Five subjects were randomly assigned to each version. The age of subjects ranged from 24 to 33 (mean 28.75 years). They were recruited through flyers and received 1,000 Yen for participation.

3.1.3. Apparatus The presentation of the apartment was hosted on a computer with a 17 inch (42.5 cm) monitor (the main monitor). A second computer was used to control the eye tracking system, a NAC Image Technology Eyemark Recorder model EMR-8B [8]. The system has a sampling rate of 60 Hz. The subject's head posture was maintained with a chin rest, with the eyes at a distance of 24 inch (60 cm) from the main monitor. A digital video recorder that captured the data from the third camera was connected to the computer that processed the eye movements. (The subjects were also connected to a bio-signal encoder that collected skin conductance and heart rate information. Results regarding those physiological data will be discussed in another publication.)

3.1.4. Procedure The subjects were first briefed about the experiment. They were told that an apartment will be shown to them, and that they would be asked general questions about the apartment afterwards. They were also instructed to watch the demonstration carefully since they should be able to report features of the apartment to others. Calibration of the eye tracker was performed by instructing subjects to fixate nine points in the screen area. After that, the subjects were shown the presentation that lasted for 8 minutes. Finally, the subjects were freed from the equipment, and asked to fill out a questionnaire in order to report on their perception of the interface and to answer some content-related questions concerning the presented material.

3.1.5. Data Analysis For analysis, the recorded video data of a presentation were first divided into individual scenes. A scene is a presentation unit where a referring entity (agent, text box, or voice) describes a reference object (an item of the room). Only the Agent and Text versions feature a visible referring entity. In Fig. 2, the scene consists of the agent performing a hand gesture to its right and introducing the living room. In order to be able to compare the three versions, scenes where the agent or text box moves from one location were left out. For each scene (41 in total), the following four screen area categories were defined:

- The area of a (visible) referring entity is either the smallest rectangle demarcating the agent or the text box (the agent area is further subdivided into face and body areas).
- The area of the reference object is the smallest rectangle demarcating the object currently described.
- The layout area (a designated, permanent reference object) is the field on the screen that displays the layout of the apartment.
- Other screen areas.

A program has been written that first maps eye-tracking data to xy-coordinates of the video sequence, and then counts the gaze points in each of the four categories.

When eye movements are relatively steady for a short period in one area, they are called *fixations* whereas rapid shifts from one area to another are called *saccades* [3]. During a saccade, no visual processing takes place. If a cluster of gaze points has less than 6 entries, it was categorized as part of a saccade [5]. All data accounted for in the analysis are derived from the activity of subjects' left eyes.

3.2. Results

3.2.1. Focus of Attention Hypothesis The ability of the interface to direct a subject's focus of attention to reference objects has been tested in two ways, spatial and spatio-temporal. The *spatial* analysis counts the gaze points that fall within areas of interest, specifically the reference object area and the layout area. Except for the introductory episode, the layout is not explicitly referred to during the presentation although it may serve as an orientation aid for users. The hypothesis is tested by restriction to those scenes where the referring entity (agent, text, voice) refers to some

item of the apartment. A between-subjects analysis of variance (ANOVA) showed that users focus on the reference objects more in the Voice version than in either of the Agent or the Text version ($F(2,9) = 8.2$; $p = 0.009$). The result for the map area, while not statistically significant, shows a tendency toward a similar distribution of gaze points ($F(2,9) = 2.8$; $p = 0.11$). (For a comparison between gaze points in the agent and text box areas, see the Locked Attention Hypothesis.) Those results suggest that gaze points are not randomly distributed across the screen area but depend on the presence of a visible presentation medium. When an agent or a text box is present, users' attentional focus is more evenly shared between the presentation medium and the presented material, as in human–human communication.

3.2.2. Locked Attention Hypothesis This hypothesis compares the portions that subjects focus on the agent (face or body) or the text box that reveals text line by line. The mean for the agent is 18% of the total number of gaze points, and the mean for the text box is 32%. The t-test (one-tailed, assuming unequal variances) showed that subjects look significantly more often at the text box ($t(6) = -2.47$; $p = 0.03$). This result can be seen as evidence that users spend considerable time for processing an object that gradually reveals new information. Locked attention can prevent users from attending to other salient information [4].

3.2.3. Shift of Attention Hypothesis While a spatial analysis can indicate where attention is spent, it cannot reveal the nature of *how* users traverse the interface when watching a presentation. In order to address those more complex aspects of intelligent interfaces, we performed a (preliminary) *spatio-temporal* analysis of eye movement data. Figure 3 depicts a screen shot of the original view (taken by the outward directed camera of the EMR-8B system) of a subject in the Agent version. The dark colored dots are gaze points drawn by our program. The numbers have been added to the screen shot by hand. The frames around the agent (face, body) and the layout have been re-drawn for clarity. When the agent speaks the sentence written as the title of Fig. 3, the subject's focus of attention is first on the agent's face, next on the layout area, then it traverses back to the agent's face, and finally shifts to the layout area.

A more detailed description of one subject's attentional shifts is shown in Fig. 4. The rectangles above the sentences of the introductory episode of the apartment presentation indicate the focus of the subject's attention. The surface structure of the sentences is synchronized with attentional focus. Observe that the subject initially shifts attention between the agent and the living room (the current reference object), and when the agent says "The space of this apartment is 78 square meters", the subject focuses on the layout that depicts the size of the apartment. In the following, the subject partly attends to the agent's gesture, and after some occasional shifts to other areas, fixates on the layout. When the agent explains how the rooms are marked, the subject is apparently not attending to the layout during the utterance of the sentence.

The attentional shifts in the example of Fig. 4 suggest that users can perceive embodied agents to possess a certain degree of competence, such as directing the user to locations of interest. Even more importantly, it demonstrates

Figure 3. "To your left is the layout of the apartment. As you can see, the apartment includes: bedroom, living room, dining room, den, kitchen and bathroom."

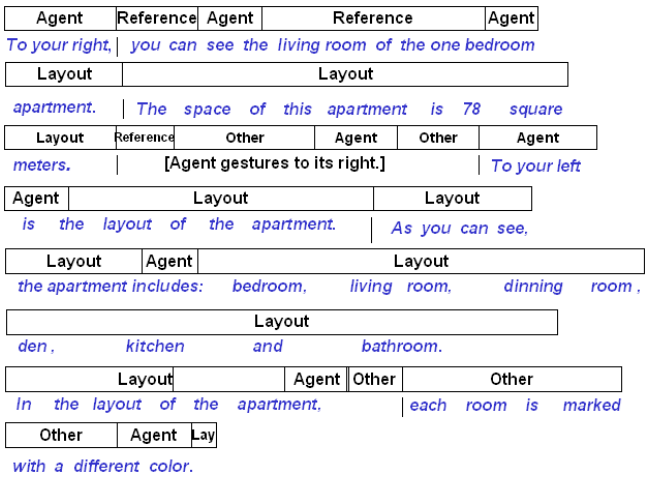

Figure 4. Example of attentional shifts in the introductory episode of the presentation.

how a user re-directs attentional focus back to the agent after being directed to a reference object, which supports the interpretation of users expecting agents to provide them conversational cues and other meaningful information.

As a first attempt to provide a systematic spatio-temporal analysis of eye movements for embodied agent based interfaces, we propose a Instructor–Reference–Instructor (IRI) triple as a basic unit for evaluation. A IRI denotes a situation where the user first attends to an instructor, a referring entity like an agent or a text box, then focuses on a reference object, and afterwards shifts attention back to the instructor. IRIs appear to be important interaction patterns in conversation, including direction-giving tasks [9], and strong indicators of the instructor agent being conceived of as a social actor. As a preliminary evaluation, we compared the number of IRIs of the Agent and Text versions for the episode displayed in Fig. 4 (plus one sentence). Here, both the living room and the layout qualify as reference objects. Figure 4 e.g. contains 4 IRIs. The t-test on the small sample was not significant ($t(5) = 1.75$; $p = 0.07$). The means are: Agent (4.34) and Text (2). While this outcome indicates a tendency, further analysis with more episodes is needed to support the hypothesis that embodied agents trigger conversational behavior in users.

3.2.4. Agent Face–Body Hypothesis This hypothesis has been tested by summarizing gaze points that are contained in either the agent face or the agent body region. It could be shown that subjects were looking mostly at the agent's face (mean = 83.1%; stdev = 6.8), which supports the claim that users interact socially with interface agents [16].

4. Conclusions

This paper tried to answer the question whether non-verbal behavior of an embodied agent matters in human–computer interaction. Based on two studies that employ embodied agents we demonstrated that non-verbal behavior is of key importance in (at least) two ways: (i) Non-verbal agent behavior is crucial to the expression of affect and empathy, and may thus be used as a stress-reducing channel in human–computer interaction; (ii) Deictic gestures in addition to speech can be used to direct the user's focus of attention and provide navigational aid to the user.

A shortcoming of the work described in this paper is that the impact of non-verbal behavior was not tested for (truly) interactive face-to-face communication between a human and an embodied agent. Here, we refer the interested reader to the work described in [9]. Another issue relates to the risk of selecting inappropriate non-verbal behaviors and possible negative effects on human–agent interaction. Similarly, repetitive verbal and non-verbal behaviors (as in the case of apologizing in the quiz example) will have to be treated with more care in the future.

Acknowledgements

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

References

[1] C. L. Breazeal. *Designing Sociable Robots*. The MIT Press, Cambridge, Massachusetts, 2002.

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, 2000.

[3] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, UK, 2003.

[4] P. Faraday and A. Sutcliffe. An empirical study of attending and comprehending multimedia presentations. In *Proceedings of ACM Multimedia 96*, pages 265–275, Boston MA, 1996.

[5] J. H. Goldberg and X. P. Kotval. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645, 1999.

[6] C. Ma, H. Prendinger, and M. Ishizuka. Eye movement as an indicator of users' involvement with embodied interfaces at the low level. In *Proceedings of AISB-05 Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction – Situational and Environmental Information Enforcing Involvement in Conversation*, 2005.

[7] Microsoft. *Developing for Microsoft Agent*. Microsoft Press, Redmond, WA, 1998.

[8] NAC Image Technology, 2004. URL: `http://eyemark.jp`.

[9] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of Association for Computational Linguistics (ACL-03)*, pages 553–561, 2003.

[10] R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, MA, 1997.

[11] H. Prendinger, S. Descamps, and M. Ishizuka. MPML: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages and Computing*, 15(2):183–203, 2004.

[12] H. Prendinger and M. Ishizuka, editors. *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg, 2004.

[13] H. Prendinger, J. Mori, and M. Ishizuka. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International Journal of Human-Computer Studies*, 62(2):231–245, 2005.

[14] Thought Technology Ltd., 2002. URL: `http://www.thoughttechnology.com`.

[15] Tokyo Mansions, 2004. URL: `http://www.themansions.jp/`.

[16] M. Witkowski, Y. Arafa, and O. de Bruijn. Evaluating user reaction to character agent mediated displays using eye-tracking technology. In *Proceedings AISB-01 Symposium on Information Agents for Electronic Commerce*, pages 79–87, 2001.