To appear in *Applied Artificial Intelligence*, 1999.

The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents*

Justine Cassell and Kristinn R. Thórisson†
Gesture and Narrative Language Group, M.I.T. Media Laboratory,
Cambridge, Massachusetts, USA

Abstract

In this article we describe results from an experiment of user interaction with autonomous, human-like (humanoid) conversational agents. We hypothesize that for embodied conversational agents, nonverbal behaviors related to the process of conversation, what we call envelope feedback*, is much more important than other feedback, such as emotional expression. We test this hypothesis by having subjects interact with three autonomous agents, all capable of full-duplex multimodal interaction: able to generate and recognize speech, intonation, facial displays, and gesture. Each agent, however, gave a different kind of feedback: (1) content-related only, (2) content + envelope feedback, and (3) content + emotional*. Content-related feedback *includes answering questions and executing commands;* envelope feedback *includes behaviors such as gaze, manual beat gesture, head movements;* emotional feedback *includes smiles and looks of puzzlement. Subjects' evaluations of the system were collected with a questionnaire, and video tapes of their speech patterns and behaviors were scored according to how often the users repeated themselves, how often they hesitated, and how often they got frustrated. The results confirm our hypothesis that envelope feedback is more important in interaction than emotional feedback and that envelope feedback plays a crucial role in supporting the process of dialog. A secondary result from this study shows that users give our multimodal conversational humanoids very high ratings of lifelikeness and fluidity of interaction when the agents are capable of giving such feedback.*

Introduction

Research intended to answer questions about the various features of embodied agent-oriented systems—systems that employ an embodied character as interface—has to date been hampered by the lack of real computer systems capable of sustaining and supporting multimodal dialog with a human user. To assess topics such as lifelikeness, trust, effectiveness of communication,

---

This research was sponsored by the sponsors of the MIT Media Laboratory, TSG Magic, and RANNÍS. We thank Joey Chang, Youngme Moon, Erin Marie Panttaja, Roland Paul, Mukesh Prasad Singh, and Hannes Vilhjálmsson for the care and feeding of Gandalf, and other valuable contributions to this work.

K. R. Thórisson can be reached at kris@media.mit.edu.

Address correspondence to Justine Cassell, Gesture and Narrative Language Group, MIT Media Laboratory, 20 Ames Street, E15-315, Cambridge, MA 02139, USA. E-mail: justine@media.mit.edu

users' likability of the interaction, as well as the question of whether to employ humanoid figures to represent the system, these studies have turned to rhetoric (cf. Lanier 1995, Laurel 1992), Wizard-of-Oz techniques (Maulsby et al., 1993, Hauptman, 1989), mixed automation/Wizard-of-Oz (Thórisson, 1994), typed natural language (Neal & Shapiro, 1991, Wahlster, 1991), iconic embodiments of various types (King & Ohya, 1996, Maes, 1994), or simply ignored the issue of embodiment (Sparrell, 1993, Thórisson et al., 1992). As a result, one cannot justifiably generalize the results of these studies and systems to future systems employing computer-controlled characters capable of real-time dialog—tempting as it may be.

In this article we present an embodied animated agent capable of supporting real-time multimodal conversation with a user. We exploit the multimodal capabilities of this agent system in order to test the comparative importance of two aspects of interaction claimed to be crucial motivations for multimodality in interactive systems: turn-taking and other conversation-process-oriented interaction behaviors, and emotional responses. In what follows, we first lay out the debate over what features make an embodied animated agent most worthwhile to conversational systems. We then turn to the nonverbal, embodied behaviors that humans use in conversation. We describe how sets of these behaviors make up our two feedback conditions in the next section, on experimental design. Finally, we describe our results and draw conclusions about what embodied behaviors should be addressed first in the design of animated embodied interactive agents for conversation.

Feedback in Embodied Conversation

Responses to commands and questions are a given in any purposeful, conversational system—without appropriate response to content, there is little point to interactive systems (e.g. Moore & Paris, 1993, Cawsey, 1993). One sometimes hears the claim that there is no need for modes other than speech, since the speech channel carries most or all the necessary information in conversational systems (Ochsman & Chapanis, 1974). This has been countered by a growing body of research on believable, life-like embodied conversational agents (André & Rist, 1996, Bates, 1994, Cassell et al., 1994; Nass et al., 1993). However, even within the community that accepts embodiment to be potentially important in interactive systems, many different human characteristics have been put forth as *the* key to making embodied agents effective—fluid body movement, face and hand gestures, emotional expressions, posture, realistic looking skin. Nonetheless the young field of synthetic computer characters has not seen much research comparing these different putatively "most important" characteristics of interactive computer characters.

Two of these human characteristics have been particularly singled out as important to conversational systems: *emotional feedback* and *envelope feedback.*[*] For the most part, in this

[*] We call these behaviors "envelope" to convey the fact that they concern the outer envelope of communication, rather than its contents. A similar distinction between content and envelope is made by Takeuchi & Nagao (1993) when they refer to the difference between "object-level communication" (relevant to the communication goal) and "meta-level processing" (relevant to communication regulation).

conversation-oriented literature, emotional feedback has meant *emotional emblems*. Emotional emblems are facial displays that reference a particular emotion without requiring the person showing the expression to feel that emotion at the moment of expression (Ekman, 1979). In the literature on anthropomorphism, emotional feedback as displayed by the animated agent's emotional emblems in response to a user's input is held to be a feature that an embodied agent-based interface could—and *should*—add to human-computer interaction (cf. Elliott, 1997, Koda & Maes, 1996, Hasegawa et al., 1995, Nagao & Takeuchi, 1994, Takeuchi & Nagao, 1993, Britton, 1991). The emotional feedback used in such systems has been, in general, very simple: scrunched eyebrows to indicate puzzlement, a smile and raised eyebrows to indicate happiness. The second characteristic held to be important to the effectiveness of conversational systems is envelope feedback. In this literature, for the most part, envelope feedback has meant the nonverbal (and occasionally verbal) behaviors that exist in face-to-face conversation, such as manual gesture, back-channel feedback (Yngve, 1970), and gaze, and that the animated agent produces in response to the user's communicative actions. The designers of these systems, likewise, claim that these envelope behaviors are essential to embodied interactive conversational systems (Cassell et al., 1994, Lester et al., 1997, Pelachaud et al., 1991, Thórisson 1994, Cassell et al., forthcoming). The envelope feedback used in such systems is equally simple, primarily nods, and glances towards and away from the user, but emphasizes the issue of timing in the production of such feedback.

With many researchers focusing on these important issues, we ask the questions (1) Does nonverbal feedback help in animated agents, and if so (2) which kind of nonverbal feedback? In particular, which is more crucial: providing the system with the ability to provide (1) *emotional* feedback, or (2) feedback that is related to the *process* of the conversation? Our claim is that the importance of embodiment in computer interfaces lies first and foremost in its power as a *unifying concept for representing the processes and behaviors surrounding conversation*. If this is true, feedback that relates directly to the process of the conversation should be of utmost importance to both conversational participants, while any other variables, such as emotional displays, should be secondary.

Let us now turn to what is meant by these two types of nonverbal behaviors. Before dividing the behaviors into emotional and envelope feedback, we give a brief introduction to the kinds of nonverbal behaviors that are found in the context of conversation. We will first discuss facial and gaze behaviors, and then hand behaviors.

When we talk about nonverbal conversational behaviors, we are really most interested in precisely-timed changes in eyebrow position, expressions of the mouth, movement of the head and eyes, and gestures of the hands. For example, raised eyebrows + a smiling mouth, is taken to be a happy expression (Ekman & Friesen, 1984), while moving one's head up and down is taken to be a nod, and a quick two-phase gesture of the hand is called a beat (McNeill, 1992). Some facial displays[*] are linked to personality and remain constant across a lifetime (a "wide-eyed look"), some are linked to emotional state and may last as long as the emotion is felt (downcast

[*] Following the literature, we refer to these conjunctions of facial behaviors as "facial displays" rather than "facial expressions" to avoid the automatic connotation of emotional expression.

eyes during depression), and some are synchronized with the units of conversation and last only a very short time (eyebrow raises along with pitch-accented words). Likewise, some hand gestures can carry information in the absence of speech (these tend to be culturally specific, e.g., the "thumbs up" gesture), and some are synchronized with conversational units.

In addition to characterizing nonverbal behaviors by the muscles or part of the body in play, we can also characterize them by their function in a conversation. Of course, some nonverbal behaviors convey content, e.g., holding one's hands apart as one says “the fish was this big.” Some nonverbal behaviors convey transitory emotional states (emotional feedback to the conversation), and some have an envelope, or conversational-process-oriented function. For example, smiling when asked if one would like ice cream or looking puzzled when queried about one's nonexistent sister are nonverbal behaviors with an emotional feedback function. Quick nods of the head while one is listening to somebody speak, a glance at the other person when one is finished speaking, or a beat gesture when one is taking a turn speaking have an envelope function. These are not the only functions for nonverbal behaviors in conversation: people also use nonverbal cues to cement social relationships (polite smiles) and to fulfill grammatical functions (eyebrow raises on pitch-accented words).

Note that the same movements by the body can have two or more different functions. Smiles can serve the function of emotional feedback, indicating that one is happy, or they can serve a purely social function even if one is not at all happy. Nods of the head can replace saying "yes" (a content
function), or simply indicate that one is following, even if one does not agree with what is being said (an envelope function).* In this study we compare the contribution of two debated functions filled by nonverbal behavior: the emotional feedback function and the envelope function. The claim here is that envelope behaviors, which relate directly to the process of conversation, are the strongest argument for using an embodied agent in speech-based human-computer interaction.

Experimental Design

The System

To test this claim, we used a fully automated character generation system, called Ymir, capable of real-time, multimodal, face-to-face interaction with a user (Thórisson, this issue). Ymir is a testbed system especially designed for prototyping multimodal agents that understand human communicative behavior and generate integrated spontaneous verbal and nonverbal behavior of their own. Ymir is constructed as a layered system and provides a foundation for accommodating any number of interpretive processes, running in parallel, working in concert to interpret and respond to the user's behavior. Ymir thus offers opportunities to experiment with various computational schemes for handling specific subtasks of multimodal interaction, such as natural language processing and multimodal action generation and execution.

---

* Content nods tend to be fewer and produced more emphatically and slowly than envelope nods (Duncan, 1974).

In interactions with animated characters created in the prototype Ymir system, the character appears on one monitor and a model of the solar system on another (see Figure 1). The user wears a tracking suit that gives the computer a visual representation of the user's upper body. An eye tracker allows it to track the user's gaze. Speech is collected through a head-mounted microphone and analyzed by a grammar-based continuous speech recognizer and prosody analyzer. For more details of the prototype system, see Thórisson (this issue). The interaction revolves around users asking the animated character to show them the various planets and to answer their questions about those planets and about the solar system.

Figure 1: A subject gets ready to interact with Gandalf.

For the purposes of the current experiment, we created three animated embodied conversational agents within the Ymir environment. The three agents were identical except that each had a different face and voice *and*, depending on the condition, they each responded to conversation differently as follows: in the control condition, the autonomous character gave content-related feedback (speech and action) only. Content feedback is *speech* that pertains to the topic of the conversation, such as answers to questions, or *actions*, such as responses to requests (e.g., "Show me the sun"). The agent in the second condition added envelope feedback to these content responses, as defined below. The agent in the third condition combined emotional feedback with content responses, as defined below. If emotion or envelope feedback add nothing to content feedback, we should see no difference between these three conditions. If, however, emotional feedback or envelope feedback is beneficial for the fluidity of communication, interaction with the agents in these conditions should be more efficient and satisfying than with content feedback alone.

Kinds of Communicative Behaviors

Thus we examined users' evaluation of, and efficiency of their interaction with, three different kinds of communicative behaviors given by the autonomous humanoids: (1) Content-only feedback (CONT), (2) content + emotional feedback (EMO), and (3) content + envelope feedback (ENV).

The following agent behaviors were used in each category.

I. Content (CONT)

1. Executing commands & answering questions
2. Verbal acknowledgement as a part of carrying out an action ("okey-dokey, let's go to Jupiter!")

An example of an interaction with an agent in the content condition follows:

Gandalf: "What can I do for you?" [*Face looks at user. Eyes do not move.*]
User: "Will you show me what Mars looks like?" [*User looks at Gandalf.*]
Gandalf: "Why not—here is Mars" [*Face maintains orientation. No change of expression. Mars appears on monitor.*]
User: "What do you know about Mars?" [*User looks at map of solar system.*]
Gandalf: "Mars has two moons" [*Face maintains orientation. No change of expression.*]

II. Emotional feedback (EMO)*

3. Confused expression when it doesn't understand an utterance
4. Smiles when addressed by the user and when acquiescing to a request (for example to take the user to a particular planet)

An example of an interaction with an agent in this emotional condition follows:

Gandalf: "What can I do for you?" [*Gandalf smiles when user's gaze falls on his face, then stops smiling and speaks.*]
User: "Take me to Jupiter." [*User looks at screen and then back at Gandalf and so Gandalf smiles.*]
Gandalf: "Sure thing. That's Jupiter" [*Gandalf smiles as he brings Jupiter into focus on the screen.*]
User: [*Looks back at Gandalf. Short pause while deciding what to say to Gandalf.*]
Gandalf: [*Looks puzzled because the user pauses longer than expected*†*. Waits for user to speak.*]
User: "Can you tell me about Jupiter?"

---

* These emotional feedback emblems were chosen for two reasons. First, given the dearth of literature addressing this issue, we felt that it was important to carry out the cleanest manipulation possible, and therefore we used the two emotional expressions most commonly found in existent embodied agents. Second, we chose those expressions that were appropriate to the task. For example, we would not expect Gandalf to get angry at his user, and therefore frowns are inappropriate.

† The pause was set to 400 ms, which turned out to trigger the puzzled expression once every 3.5 turns, on average.

III. Envelope (ENV)

5. Turning head and eyes towards user when listening, and towards task when executing commands in the domain
6. Averting gaze and lifting eyebrows when taking turn
7. Gazing back at person when giving turn
8. Tapping fingers to show that it is "alive"
9. Beat gesture when providing verbal content

An example of an interaction with this envelope agent follows:*

User: "Is that planet Mars?"
Gandalf: "Yes, that's Mars." [*Gandalf raises eyebrows and performs beat gesture while saying "yes," turns to planet and points at it while saying "that is Mars," and then turns back to face user.*]
User: "I want to go back to Earth now. Take me to Earth." [*User looks at map of solar system so Gandalf looks at solar system.*]
Gandalf: "OK. Earth is third from the sun." [*Gandalf turns to planet as he brings it up on the screen, then turns to user and speaks.*]
User: "Tell me more." [*Gandalf takes about 2 seconds to parse the speech, but he knows within 250 ms when the user gives the turn, so he looks to the side to show that he's taking the turn, and his eyebrows go up and down as he hesitates while parsing the user's utterance.*]
Gandalf**:** "The Earth is 12,000 km in diameter." [*Gandalf looks back at the user and speaks.*]

Users' *experience* of the *agents' life-likeness* and *ease of interaction* was assessed by questionnaire. The efficiency of the interaction was measured by *relative number of utterances* (number of subject contributions to the discourse over the number of character contributions), *relative number of subjects' hesitations* (over the total number of their contributions), and *relative expressions of frustration* (over the total number of their contributions).

* For full details of the motivation for these behaviors, their timing, and exact movements, see Cassell, (in press).

Pretest of Emotional Expressions

Three different characters (distinctive face + distinctive voice) were constructed (see Figure 2). Each face was capable of displaying one neutral expression and the two emotional emblems, happy and confused, as well as the range of head, eye and hand movements. In previous research on emotional emblems in interactive systems, the faces have not been pretested in order to determine how viewers interpret the expressions. But, how do we know that emotional emblems actually display the emotion that we imagine they do? We ran a pilot test to confirm that the facial displays used were reliably identified in all characters, as described next.

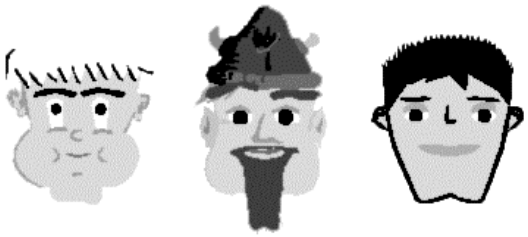

Figure 2: The faces used for the three agents, from left to right: Bilbo, Gandalf and Roland (neutral expressions).

Faces were tested by adopting a paradigm from Elliott (1997),who demonstrated that ambiguous sentences, e.g., "I see people like that all the time," would be interpreted differently when paired with faces and intonation that communicated different kinds of emotional content. For our purposes, synthesized speech was used without intonation, so as not to bias subjects. Three scenarios were presented with each ambiguous sentence, one corresponding to each facial display of neutral, happy, and confused. For example, subjects heard, "I see people like that all the time" and were asked to choose among the following three scenarios: (1) Nick is a good teacher and so he experiences satisfaction when he is stopped on the street by a former student who wanted to thank him for all he learned in his class; (2) Nick does not know what to think of crazies like the one he just saw; (3) Nick explains his client list to his new partner. One third of the subjects saw the test sentence uttered by Bilbo with a neutral face, one third heard the sentence uttered by Bilbo with a happy face, and one third heard it uttered by Bilbo with a confused face. Two other sentences were used to test the Gandalf and Roland faces. Subjects were given three 'warm-up' scenarios (without feedback from the experimenter) before the actual three test scenarios. In addition to matching the faces with descriptions, subjects rated the confidence of each of their matches from 1 (not confident) to 5 (highly confident).

A sample of 24 subjects revealed that the happy face was recognized as such by almost all subjects, for all three characters (88% of the time, with an average confidence level of 4), and confused and neutral faces were recognized 67% and 75% of the time, respectively, with confidence levels of 3 in each case, for all three characters. These faces were used in creating the animated conversational agent for each condition.

Hypotheses

We hypothesized that CONT and EMO would be equal on both measures, but ENV feedback would prove to make a significant difference for these measures. Eight hypotheses were tested:

{H1} No difference will be found for relative contributions from users between condition CONT and condition EMO.

{H2} Relatively fewer subject contributions will be found in condition ENV than conditions CONT and EMO.

{H3} No difference in hesitations will be found between conditions CONT and EMO.

{H4} Relatively fewer hesitations will be found in condition ENV than in conditions CONT and EMO.

{H5} No difference in overlaps in speech will be found between conditions CONT and EMO.

{H6} Relatively fewer overlaps in speech will be found in condition ENV than in conditions CONT and EMO.

{H7} No difference will be found in subjects' rating of the agent between conditions CONT and EMO.

{H8} Subjects in condition ENV will rate the agent higher than those in conditions CONT and EMO.

Data for hypotheses 7 and 8 was collected with a questionnaire (see the appendix). Data for hypotheses 1-6 were collected by analyzing video tape recordings of the subjects. Relative number of contributions, as well as hesitations and frustration responses, were scored according to predetermined scoring schemes (see the appendix). Video tapes were scored independently by two scorers in a double-blind design. Scoring reliability for the variables obtained from the videos (relative number of overlaps, hesitations, and contributions) was $r=0.95$ ($p < 0.001$).

Design

The experiment was a mixed between/within-subjects design: Dependent variables were: (1) relative number of user contributions per condition (number of user contributions divided by the number of agent contributions), (2) relative number of user hesitations per condition (number of hesitations over the number of contributions), (3) relative number of overlaps per condition (number of overlaps over the total number of contributions), and (4) the users' subjective assessment of the interaction based on their answers to nine questions (appendix). Independent variables were agent (Gandalf, Roland, Bilbo), feedback condition (ENV, CONT, EMO), order of faces, order of conditions. The first two were both within-subject variables. Each agent occurred equally often as the first, second or the third agent a subject interacted with. Each condition occurred equally often as the first, second or third condition to which a subject was exposed. These were thus between-subjects variables. In sum, then, each subject interacted with three different animated agents, each with a different set of conversational behaviors.

Subjects

A convenience sample of 12 volunteers between the ages of 22 and 37, both male and female, was tested. K. R. Thórisson acted as experimenter. A background questionnaire confirmed that the subjects were novice computer users with no visual problems or other handicaps. All were native English speakers.

Instructions

Subjects were told that the agents knew about the solar system and that they should "interact as normally as possible" with the agents in order to find out more about the solar system, and to get a tour of the animated model of the planets. Although the agent was capable of responding to unrecognized speech (with utterances such as "I'm sorry, I didn't get that"), to avoid an initial period of trial and error (and frustration), users were given a sheet with example utterances that the agents would understand.* Subjects had a trial of 4-8 turns with the agents before the experiment began. Subjects interacted, on average, for 7 minutes with each of the three agents, with a 5 minute break in between.

Results

Six of the eight hypotheses were confirmed (α is set at 0.05 for all hypotheses, N=12). The null hypothesis was tested with a repeated-measures multivariate analysis of variance (MANOVA) with the following dependent variables: all ratings of Gandalf in question 1 (see appendix), hesitations, contributions and overlaps, and was rejected ($F[24, 22] = 2.742$, $p < .02$). Overall, the results supported the significance of envelope feedback over emotional feedback and content-only feedback. Comparisons between individual means were done with paired t-tests and are summarized in Table 1. As can be seen in Table 1, in no case was emotional feedback rated significantly differently from the control condition. This allows us to pool the conditions of CONT and EMO for subsequent tests: CE=(CONT+EMO/2).

No effects were found for order of character, order of conditions, or interactions between these, on the dependent variables.

Hypothesis	Means	t	Significance	Confirmed
{H1} Contributions: EMO = CONT	EMO=1.52 CONT=1.33	-1.45	n.s. (two-tailed)	
{H2} Contributions: ENV < CONT, EMO	ENV=1.23 C+E/2=1.42	2.49	p < .016 (one-tailed)	
{H3} Hesitations: EMO = CONT	EMO=0.022 CONT=0.023	.07	n.s. (two-tailed)	
{H4} Hesitations: ENV < CONT, EMO	ENV=1.0 C+E/2=0.02	-2.86	p < .008 (one-tailed)	no
{H5} Overlaps: EMO = CONT	EMO=0.036 CONT=0.015	-1.55	n.s. (two-tailed)	
{H6} Overlaps: ENV < CONT, EMO	ENV=0.42 C+E/2=0.03	-2.05	p < .033 (one-tailed)	no
{H7} Agent Rating (Q1): EMO = CONT	EMO=40.67 CONT=44.83	1.86	n.s. (two-tailed)	
{H8} Agent Rating (Q1): ENV > CONT, EMO	ENV=46.83 C+E/2=42.75	-3.99	p < .001 (one-tailed)	
{H7} Helpfulness (Q2): EMO = CONT	EMO=3.23 CONT=3.02	-1.13	n.s. (two-tailed)	
{H8} Helpfulness (Q2): ENV > CONT, EMO	ENV=3.85 C+E/2=3.13	-4.29	p < .001 (one-tailed)	

*Users by no means said only utterances printed on the "cheat sheet," or, for that matter, only utterances that agents understood.

Table 1: Results from Paired t-Tests for Each of the Hypotheses. Rating hypotheses were tested with two questions from the Evaluation Questionnaire (appendix). DF = 11 for all t-values. EMO and CONT are pooled for all comparisons with ENV. Rightmost column lists which hypotheses were confirmed; n.s., not significant.

Relative Number of Contributions

There was a significant difference for relative number of contributions between the three conditions ($F[2, 10] = 4.86$, $p < 0.02$, repeated-measures MANOVA). Figure 3 shows a comparison between the number of contributions with CONT and EMO pooled. The figure indicates that, as predicted, subjects made fewer contributions per agent contribution when speaking with the envelope feedback agent than when speaking with the emotional or content-only animated agents. This measure can be taken as a rough estimate of conversational efficiency.

Subject Evaluation of Conditions

The questionnaire concerns subjects' ratings of the animated agents' language, interaction style, and lifelikeness. Subjects' rating of the characters' language abilities is surprisingly high: on a scale from 0 to 10 (humans getting a perfect 10), subjects gave agents in the ENV condition a mean of 7.25 (SD=1.86) for language understanding and 7.92 (SD=1.83) for language use. These numbers might simply indicate the users' satisfaction with the language part of the system. However, it is striking that the envelope feedback condition—in which language understanding and language generation were identical to the other conditions—was rated more highly on these dimensions, indicating that the presence of envelope feedback caused users to give the agents a higher score on language ability.

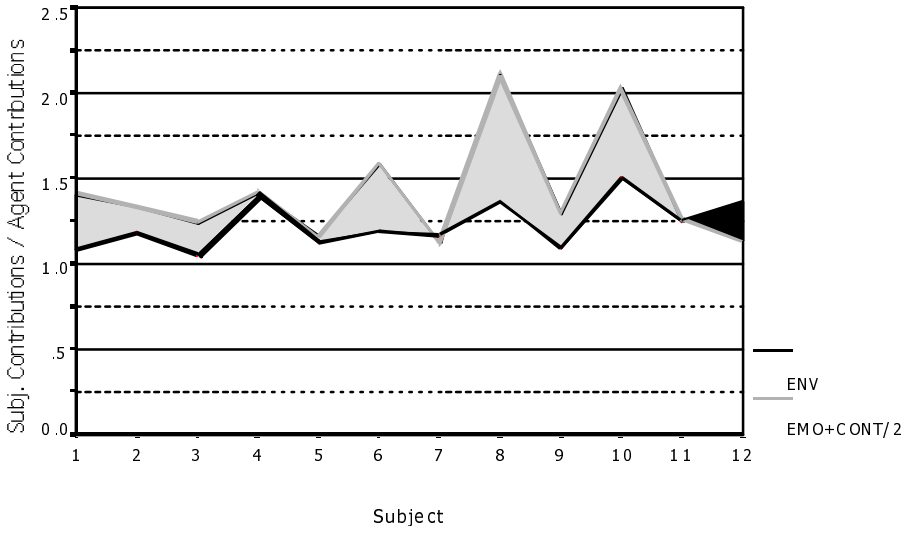

Figure 3: Difference in relative number of contributions between condition ENV (dark line) and CE (CONT+EMO/2). This difference was significant. Shaded area shows the amount of difference; the solid dark tail at right shows a clear reversal of the overall pattern for subject 12 only.

The questions that were significantly different between the ENV and CE conditions are summarized in Table 2. Along the majority of dimensions, the agent displaying envelope feedback was rated more highly than the agents in the other two conditions. We expected to get a floor effect for question 7, comparing the interaction with interacting with a goldfish. This turned out to be the case (mean score on question 1g = 4.92), lending strength to results obtained in the other questions.

Discussion

All but two hypotheses were confirmed. This supports the general claim set forth in this paper—that nonverbal behaviors are important to animated conversational agents and that apart from content, the most helpful kind of nonverbal behaviors are envelope (over emotional feedback). In particular, in terms of rating of interaction style, the agent that gives envelope feedback is rated as more helpful, more lifelike, and more smooth in its interaction style. In terms of user's behaviors, the user was more efficient in the envelope feedback condition, using fewer utterances to accomplish the task. In fact, no difference was found, along any dimension, between the agent that gave content feedback alone and the agent that gave content feedback plus emotional feedback. Note that the behaviors responsible for these results are slight nods of the head, gaze towards user and away, beat gestures when taking the turn, eye blinks—in short behaviors that are often thought of as "background processes" in human-human communication (Duncan, 1974), which for the most part last less than 500 ms, and whose onset is finely timed in the conversation.

QUESTION	ENV	C+E
1a. Language understanding	7.25	6.67
1b. Language use	7.92	6.83
1c. Smoothness of interaction	6.25	5.41
1f. Smoothness of interaction compared to interacting with a dog*	4.08	3.16
1i. Life-likeness compared to any computer character*	3.83	3.16

Table 2 Items in Question 1 (Appendix) that Were Significantly Different Between the ENV Condition and the Pooled CONT and EMO Conditions. *Scale from 1 through 5; others scaled from 0 through 10. C+E=(CONT+EMO)/2.

In two cases our original hypotheses were not confirmed. These two cases are interesting because they seem to point to what happens when an agent is so successfully human-like in its communication style that it raises user expectations past what can be currently met.

The two hypotheses that were not confirmed showed a reverse pattern of what was expected: first, subjects more frequently spoke at the same time as the agent in the envelope feedback condition and, second, more often showed evidence of hesitation or frustration in this condition. Some examination of the data, however, points to the following interpretation.

1. Subjects in the ENV condition tended to look more back and forth between the big screen (where the planets were displayed) and the animated character, tended to gesture more, and seemed to be more drawn into the interaction in general. In fact, participants tended to mimic the agents' behaviors: if the agent was rigid, they tended to stand still; if the agent was more animated, they tended to be more animated. This may be leading, in the current implementation, to a less predictable response pattern from the agent, resulting in more errors in judgement of the conversational state both on the part of the user and the agent. Nevertheless, the user prefers this kind of interaction, according to every measure, to one without this envelope feedback. We interpret this simply to indicate that the robustness of the agent's behaviors and interpretive processes needs to be improved, hardly a surprise given the early state of our prototype.
2. Although we originally assumed overlaps in speech between the user and the animated agent to indicate disfluency on the user's part, overlaps may, in fact, indicate that the user is treating the animated agent *exactly* like another, human, conversational partner. Thus, in the same way that humans may talk over one another, users are talking over the animated character, fully expecting the character to be able to keep up. Unfortunately, the agents' multimodal analysis is not powerful enough to keep up, and in cases of overlap, the agent doesn't understand the user's actions. This makes the user hesitate or turn towards the experimenter for assistance. In this situation, the envelope, process-related cues that the agent is providing—in particular, looking towards the user to give over the turn—are allowing the user to begin her/his turn more quickly than in the condition in which s/he must wait to see if the synthesized speech is really finished. That is, the agent's envelope feedback is increasing the naturalness of its behavior, which in turn, leads the user to become more spontaneous and natural.

One might argue that the emotional agent condition was not strong enough to test against the envelope agent. It is true that only two emotional expressions (happy, puzzled, neutral) were used by the emotional agent, and technical issues led to those faces not being as highly differentiated as one might wish. However, we would argue that it is quite common to find such a small number of expression constituting "emotion" in interactive systems and, further, unlike previous research, we did pretest those faces and found them to be distinguished reliably by users. Moreover, other research in our lab has shown that the same emotional agent, using exactly these expressions, is differentiated from the content-only face and is found to be more "friendly" by users, while the envelope agent is found to be more "helpful" (Cassell, in press). Nonetheless, it is clear that further research is needed with a larger variety of more finely drawn expressions.

Conclusions

Results such as these point strongly towards the need to further explore the role of envelope feedback in humanoid animated conversational agents. We have shown that users rate conversational agents employing envelope feedback behaviors more highly than agents without nonverbal behaviors or agents giving emotional feedback and that such envelope behaviors increase users' efficiency, and decrease their conversational mis-starts. We believe that these

results derive from the role of envelope feedback in allowing users to (1) apply the knowledge that they already have about human conversation to their interaction with the computer, and (2) trust that the conversation is proceeding smoothly, since they are updated about the process of conversation throughout the interaction.

These results do not rule out a role for emotional feedback in certain conditions or tasks. First, while we tested the simple set of emotional emblems that are implemented in many existent interactive systems, a more complex set of emotional expressions may be more effective in engaging users. Second, as Takeuchi & Naito (1995) remark, one reason that interface systems that incorporate emotional facial displays may take more effort rather than less, is that users may be trying to interpret the meaning of the human expressions. This is fine in cases where a system could be expected to be conveying the emotional, semantic function of facial displays—but these cases are rare. That is, whereas the woman who sells you your daily lottery ticket may naturally be expected to show pleasure, disappointment, or surprise as you read off the numbers on your ticket, the agent who retrieves your e-mail does not have an awful lot to get emotional about. Work circumstances such as these still make up the majority of situations in which animated agents are employed [although see Lester et al., (1997)]. Therefore, in the general case, we can say that more effort should be put into implementing envelope behaviors, which facilitate human-like, predictable, and smooth interaction with an embodied conversational agent.

In sum, we have demonstrated that conversational systems employing an anthropomorphic, human-like character may work better, and be better accepted, if they provide important envelope features from human face-to-face interaction, such as back-channel feedback, attentional cues, and beat gestures.

APPENDIX: SCORING & EVALUATION QUESTIONNAIRE

Scoring of Contributions

Relative number of subject contributions was estimated in the following way: for each utterance or action of the agent, the number of times a subject repeats the same request or makes a new request is counted.

Scoring of Hesitations & Frustration

The following behaviors will be counted as constituting interaction-related hesitations and frustration, on behalf of the subjects:

1. Restarts (related to agent's behavior—not to recalling a command).
2. Subject looks at experimenter while waiting for agent to respond.
3. Subject clearly indicates frustration with agent's response or lack of response, by gesture, or verbally, for example by asking experimenter what to do.

Evaluation Questionnaire

(Notice to the reader: The name appearing in the questions, Gandalf, Bilbo, or Roland, depended on which character the subject just interacted with.)

You have just interacted with Gandalf, one of three computer-enacted characters we are developing. Following are questions related to your experience with Gandalf. Please answer them to the best of your ability.

Estimated time: 2 minutes.

Your cooperation is appreciated.

1. Please rate Gandalf on the following issues:

1a. On a scale from 0 to 10, assuming that a human gets a score of 10, Gandalf's understanding of your spoken language gets a score of _____.

1b. On a scale from 0 to 10, assuming that a human gets a score of 10, Gandalf's use of spoken language gets a score of _____.

When two people interact face-to-face, their interaction is most of the time very smooth, with minimal hesitations and misunderstandings.

1c. On a scale from 0 to 10, assuming that a human gets a score of 10, the smoothness of the interaction with Gandalf gets a score of _____.

1d. Compared to interacting with a dog, Gandalf's understanding of spoken language is ...

o...much better. o...somewhat better. o...about equal. o...slightly worse. o ...much worse.

1e. Compared to interacting with a dog, Gandalf's use of language is ...

o...much better. o...somewhat better. o...about equal.

o...slightly worse. o...much worse.

1f. Compared to interaction with a dog, the smoothness of the interaction with Gandalf is ...

o...much better. o...somewhat better. o...about equal. o...slightly worse. o...much worse.

1g. Compared to interacting with a fish in a fishbowl, interacting with Gandalf is ...

o ...much more interesting.

o ...somewhat more interesting.
o ...about equal
o ...somewhat less interesting.
o ...much less interesting.

1h. Compared to any real animal (excluding humans), Gandalf seems...
o ...incredibly life-like.
o ...very life-like.
o ..somewhat life-like.
o ...not very life-like.
o ...not life-like at all.

1i. Compared to the most life-like character in any computer game or program you have seen, Gandalf seems...
o ...incredibly life-like.
o ...very life-like.
o ..somewhat life-like.
o ...not very life-like.
o ...not life-like at all.

2. How helpful to the interaction did you find ...

2a. ...the content of Gandalf's speech?
o Very helpful.
o Somewhat helpful.
o Neither helpful nor unhelpful.
o Unhelpful.
o Counterproductive.

2b. ...Gandalf's head motions?
o Very helpful.
o Somewhat helpful.
o Neither helpful nor unhelpful.
o Unhelpful.
o Counterproductive.

2c. ...Gandalf's expressions?
o Very helpful.
o Somewhat helpful.
o Neither helpful nor unhelpful.
o Unhelpful.
o Counterproductive.

2d. ...Gandalf's gaze?
o Very helpful.
o Somewhat helpful.

o Neither helpful nor unhelpful.
o Unhelpful.
o Counterproductive.

2e. ...Gandalf's hand gestures?
o Very helpful.
o Somewhat helpful.
o Neither helpful nor unhelpful.
o Unhelpful.
o Counterproductive.

References

André, E. and T. Rist, 1996. Coping with temporal constraints in multimedia presentation planning. *Pro. of AAAI-96*, pp. 142-147.

Bates, J. 1994. The role of emotion in believable agents. *Commun. ACM*, 37(7): 122-125.

Britton, B.C.J. 1991. Enhancing computer-human interaction with animated facial expressions. Master's thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Cassell, J. In press. Embodied conversation: Integrating face and gesture into automatic spoken dialog systems. In *Spoken dialogue systems*, ed. S. Luperfoy. Cambridge, Mass.: MIT Press.

Cassell, J., C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Beckett, B. Douville, S. Prevost, and M. Stone. 1994. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Comput. Graphics*, 28(4): 413-420.

Cassell, J., O. Torres, and S. Prevost. In press. Turn taking vs. discourse structure: How best to model multimodal conversation. In *Machine conversations*, ed. Y. Wilks. The Hague: Kluwer.

Cawsey, A. 1993. Planning interactive explanations. *Int. J. Man-Machine Stud.* 38: 169-199.

Duncan, S. 1974. Some signals and rules for taking speaking turns in conversations. *In Nonverbal communication*, ed. S. Weitz. New York: Oxford University Press.

Ekman, P. 1979. About brows: Emotional and conversational signals. In *Human Ethology: Claims and limits of a new discipline*, eds. M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, 169-249. New York: Cambridge University Press.

Ekman, P., and W. V. Friesen. 1984. *Unmasking the face*. Palo Alto, Calif.: Consulting Psychologists Press.

Elliott, C. 1997. I picked up Catapia and other stories: A mulitmodal approach to expressivity for "emotionally intelligent" agents. In *Proc. First International Conference on Autonomous Agents*, 451-457. February 5-8, Marina del Rey, Calif.

Hasegawa, O., K. Itou, T. Kurita, S. Hayamizu, K. Tanaka, K. Yamamoto, and N. Otsu. 1995. Active agent oriented multimodal interface system. In *Proceedings of IJCAI*, 82-87. August 19-25, Montréal, Québec, Canada.

Hauptman, A.G. 1989. Speech and gestures for graphic image manipulation. *In Proc. SIGCHI 89*, 241-245. New York: ACM Press.

King, W. J. and J. Ohya. 1996. The representation of agents: Anthropomorphism, agency and intelligence. In *Proc. CHI '96: ACM Conference on Human Factors in Computing Systems,* 289-290. New York: ACM Press.

Koda, T., and P. Maes. 1996. Agents with faces: The effects of personification of agents. Presented at Human-Computer Interaction '96, August, London, UK.

Lanier, J. 1995. Agents of alienation. *Interactions* 2(3) (July): 66-72.

Laurel, B. 1992. Anthropomorphism: From Eliza to Terminator 2. Panelist discussion (A. Don, moderator). *In CHI '92 Conference Proceedings: ACM Conference on Human Factors in Computing*, 67-70. May 3-7, Monterey, Calif.

Lester, J., J. Voerman, S. Towns, and C. Callaway. 1997. Cosmo: A life-like animated pedagogical agent with deictic believability. In *Working Notes of the IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent*, 61-69. August, Nagoya, Japan.

Maes, P. 1994. Agents that reduce work and information overload. *Commun. ACM* 37(7): 31-40.

Maulsby, D., D. Greenberg, and R. Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. *Proc. InterCHI '93*, 277-84. April 24-29, Amsterdam, Netherlands.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Moore, J.D., and C. L. Paris. 1993. Planning text for advisory dialogs: Capturing intentional and rhetorical information. *Comput. Linguistics* 19(4): 651-695.

Nagao, K., and A. Takeuchi. 1994. Social Interaction: Multimodal conversation with social agents. In *Proc. 12th National Conference on Artificial Intelligence (AAAI-94),* vol. 1, 22-28. August 1-4, Seattle, Wash.

Nass, C., J. Steuer, E. Tauber, and H. Reeder. 1993. Anthropomorphism, agency and ethopoeia: Computers as social actors. In *InterCHI '93 Adjunct Proceedings*, 111-112.

Neal, J. G., and S. C. Shapiro. 1991. Intelligent multi-media interface technology. In *Intelligent User Interfaces*, eds. J.W. Sullivan, and S.W. Tyler, pp. 11-45. New York: Addison-Wesley.

Ochsman, R. B., and A. Chapanis. 1974. The effects of 10 communication modes on the behavior of teams during co-operative problem solving. I*nt. J. Man-Machine Stud*. 6: 579-619.

Pelachaud, C., N. I. Badler, and M. Steedman. 1991. Linguistic issues in facial animation. In *Computer animation '91*, eds. N. Magnenat-Thalmann and D. Thalmann, 15-30. New York: Springer-Verlag.

Sparrell, C.J. 1993. Coverbal iconic gesture in human-computer interaction. Master's thesis, MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

Takeuchi, A., and K. Nagao. 1993. Communicative facial displays as a new conversational modality. In *Proc. InterCHI, '93*, 187-193. April 24-29, Amsterdam, Netherlands.

Takeuchi, A., and T. Naito. 1995. Situated facial displays: Towards social interaction. In *Human factors in computing systems: CHI '95 Conference Proceedings*, 450-454. May 7-11, Denver, Colo.

Thórisson, K. R. 1994. Face-to-face communication with computer agents. In *AAAI Spring Symposium on Believable Agents Working Notes*, 86-90. March 19-20, Stanford University, Stanford, Calif.

Thórisson, K. R. This issue. A Mind model of multimodal communicative creatures and humanoids. *Applied Artificial intelligence*.

Thórisson, K. R., D. B. Koons, and R. A. Bolt. 1992. Multi-modal natural dialogue. *SIGCHI Proc. '92*, 653-654. (Also in SIGGRAPH Video Review, CHI '92 Technical Video Program 76(1).)

Wahlster, W. 1991. User and discourse models for multimodal communication. In *Intelligent user interfaces*, eds. J.W. Sullivan and S.W. Tyler, 45-67. New York: Addison-Wesley.

Yngve, V. H. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting,* 567-78. Chicago, ILL.: Chicago Linguistics Society.