

Embodied Conversational Agents as a UI Paradigm: A Framework for Evaluation

Jun Xiao¹, John Stasko¹ and Richard Catrambone²

¹College of Computing / Gvu Center, ²School of Psychology

Georgia Institute of Technology

Atlanta, GA 30332 USA

+1 404 385 2447

{junxiao, stasko}@cc.gatech.edu, rc7@prism.gatech.edu

ABSTRACT

Research on embodied conversational agent interfaces has produced widely divergent results. We suggest that this is due to insufficient consideration of key factors that influence the perception and effectiveness of agent-based interfaces. Thus, we propose a framework for the evaluation of conversational agent interfaces that can systematize the research. The framework emphasizes features of the agent, characteristics of the user, and the task the user is performing.

We have conducted experiments within this framework. The first study manipulated the agent's appearance (lifelike versus iconic) and the nature of the user's task (carrying out procedures versus providing opinions). We found that the perception of the agent was strongly influenced by the task while features of the agent that we manipulated had little effect. The second study (in progress) manipulates the initiative of the agent (proactive versus reactive). Initial analysis of the data showed that the participants strongly preferred proactive agents while initiative of the agent had little effect on their task performance.

Keywords

Embodied conversational agents, empirical study, agent-based interfaces, evaluation

1. INTRODUCTION

Embodied conversational agents who answer questions and perform tasks through conversational, natural language-style dialogs with users contrast the traditional view of computers as enabling tools for functional purposes. Many people believe that such interfaces have great potential to be beneficial in HCI for a number of reasons. Agents could act as smart assistants, much like travel agents or investment advisors [2]. A conversational interface appears to be a more natural dialog style because the user does not have to learn complex command structure and functionality [9]. Furthermore, an embodied agent could use intonation, gaze patterns, facial expressions and gestures, in addition to words, for conveying information and affect [1][12]. The human face seems to occupy a privileged position for conveying a great deal of information, including relatively subtle information, efficiently [6]. Finally, embodied conversational

agent interfaces could make a computer more human-like, engaging, entertaining, approachable, and understandable to the user, thus harboring potential to build trust and establish relationships with users, and make them feel more comfortable with computers.

However, relatively little careful empirical evaluation on embodied conversational agent interfaces has been performed, and the results from this research have been contradictory or equivocal [2]. We believe the question of whether embodied conversational agent interfaces are useful or useless is too general because it seems to depend on specific behaviors of the interface agents, characteristics of the users, and the kind of tasks users are trying to perform. Dehn and van Mulken suggest taking a more fine-grained perspective in their review of the various empirical studies conducted on animated interface agents [4]. Our goal is to contribute to the community's understanding of embodied conversational agent interfaces by identifying the dimensions of the design space, discovering the correlations and tradeoffs between factors, and distinguishing factors that should and can be improved.

One fundamental issue in the quality of agent interfaces is competence [11]. It appears obvious that perceptions of embodied conversational agent interfaces will be strongly influenced by the competence of the supporting software system and the quality of the replies and suggestions made by the agent. We are using a "Wizard of Oz" methodology in which we provide the back-end intelligence to the agent [3], which allows us to either factor out competence as an issue or control competence as a condition to see its effect on user performance and impression.

2. RELATED WORK

A few studies have revealed that anthropomorphic agents are attention grabbing and people make natural assumptions about the intelligence and abilities of those agents. King and Ohya found that a dynamic 3D human form whose eyes blinked was rated more intelligent than any other form, including non-blinking 3D forms, caricatures, and geometric shapes [7].

One common trend discovered in studies is that embodied conversational agents appear to command people's attention, both in positive and negative senses. Takeuchi and Nagao created conversational style interaction systems that allowed corresponding facial displays to be included or omitted [20]. According to their metrics, the conversations with a face present were more "successful." Across two experiments, they found that the presence of a face provided important extra conversational cues, but that this also required more effort from the human interacting with the system and sometimes served as a distraction.

Other studies have shown that the attention garnered by an embodied conversational agent had a more positive, desired effect. Walker, Sproull, and Subramani found that people who

interacted with a talking face spent more time on an on-line questionnaire, made fewer mistakes, and wrote more comments than those who answered a text questionnaire [21].

Koda created a Web-based poker game in which a human user could compete with other personified computer characters including a realistic image, cartoon male and female characters, a smiley face, no face, and a dog [8]. She gathered data on people's subjective impressions of the characters and found that people's impressions of a character were different in a task context than in isolation and were strongly influenced by perceived agent competence.

The work of Nass, Reeves and their students at Stanford has focused on the study of "computers as social actors." They have conducted a number of experiments that examined how people react to computer systems and applications that have certain personified characteristics [14,15,17]. Their chief finding is that people interact with and characterize computer systems in a social manner, much as they do with other people. This occurs even when the participants know that it is only a computer with which they are interacting. More specifically, Nass and Reeves found that existing, accepted sociological principles (e.g., individuals with similar personalities tend to get along better than do those with different personalities) apply even when one of the two participants is a machine.

The studies cited above, and others, suggest that people are inclined to attribute human-like characteristics to computer agents and that a variety of factors might influence how positively the agents are viewed. As mentioned earlier though, research in this area has been hampered by a lack of a coherent framework to guide the development of hypotheses, the construction of experiments, and the interpretation of results.

3. A FRAMEWORK FOR EVALUATION ON EMBODIED CONVERSATIONAL AGENTS

To effectively and systematically investigate the use of embodied conversational agents, one needs to consider the key factors that will affect the usefulness of such interfaces. We propose an investigative framework composed of three key components: characteristics of the user, attributes of the agent, and the task being performed.

We believe that serious empirical study in this area must systematically address each of these factors and understand how it affects human users. Below, we provide examples of individual variables within each factor that could potentially influence user performance and impressions.

3.1 Factor 1: Features of the User

Potential users vary, of course, in many ways. However, there are certain features that may be quite likely to affect how useful a user finds an embodied conversational agent. These features include:

Personality: Researchers have identified what are referred to as the "Big Five" traits that seem to be quite useful in describing human personalities: extraversion, openness, agreeableness, neuroticism, and conscientiousness (e.g.[13]). While any such breakdown is debatable, it seems reasonable to examine whether users' positions on these, or other, trait dimensions is predictive of how they will respond to agents. We can also have users rate the agents on these dimensions.

For instance, one might hypothesize that an introverted person might find a proactive agent to be intimidating while a more extroverted person would enjoy interacting heavily with the agent. It would be useful to collect information on personality traits to see if they correlate with our various measures of agent usefulness.

Background knowledge: A user who has a good deal of background knowledge in a domain might prefer an agent that is reactive and that the user can call upon when he or she needs some low-level bit of information or has a low-level task that needs to be done. Conversely, a user who is learning how to carry out tasks in a particular domain might welcome strategy advice from an agent, particularly if the agent can analyze the strategy and provide reasons for why the strategy might be altered.

Capability: The ability of a user to cognitively understand the causes of agents' actions as well as the planning capacity of the user for problem solving may vary greatly across individuals. Other non-cognitive abilities may also need to be considered. For example, in order to develop interactive learning tools for language training with profoundly deaf children, visible speech instructions are crucial [12].

Goal: Users who intend to get a solution of good quality may evaluate the usefulness of agents based on the accuracy and completeness of the agent's help, whereas users who intend to get a quick reference may evaluate the usefulness of agents based on the completion time and efficiency of the operation. Other indicators such as whether the user feels comfortable with the agent and how engaging the interaction is may be applicable in other situations such as learning and entertainment.

Psychological States: Users' moods and emotional states have both positive and negative impact on their attitude and behavior towards agents in a conscious and subconscious manner. Comforting words from an agent may be valued when the user is struggling with a math puzzle overnight, whereas surprises from an agent may not be appreciated if the user is rushing to meet a deadline.

Gender: Although background knowledge and the Big Five personality measures are likely to account for much of the user-determined usefulness of agents, it is also possible that gender will play a role. There has been some research on gender differences in advice taking, so it seems prudent to consider gender effects on the evaluations of agents.

Other variables: Other user-related variables include age, computer experience, previous experience with agent interfaces and culture.

3.2 Factor 2: Features of the Agent

Like users, embodied agents can vary on a wide variety of features. These features include:

Visual Appearance: Empirical evidence provided by Dryer suggests that rounder shapes, bigger faces, and happier expressions are perceived by humans as extraverted and agreeable, while bold colors, big bodies, and erect posture characters are perceived by human as extraverted and disagreeable [5]. In other words, visual stimuli may influence users' perception of agents' personalities and should be carefully chosen.

Fidelity: Earlier studies suggest that more realistic-appearing, 3D human representations are perceived as being more intelligent,

which could be viewed positively or negatively. Furthermore, realistic-appearing agents are more difficult to implement, so if user performance is improved by the presence of an agent, but does not vary according to appearance, simpler caricature style characters would be advantageous.

Expressiveness: Within realistic-appearing agents, we might vary the level of facial expressions, gestures, emotions, and movements of a particular character. Animated, expressive agents again may be viewed as more realistic and intelligent, but they might also unduly draw the viewer's attention and thus be distracting and annoying.

Personality: A further important component of an agent's profile is its personality. Should it be a dominant expert or humble servant? Should we adapt the personality of the agent according to the preferences of different users? As we have mentioned before, design decisions on the agent's personality should be made consistently with other characteristics of the agent, such as appearance.

Presence: Is an agent always present on the screen or does the agent only appear when it is engaged in a dialog by the user? One might hypothesize that an ever-present agent would make users uneasy by producing an effect of being watched or evaluated all the time.

Role: Should an agent act as a partner in the task or should it contribute only in clearly specified ways? For instance, an agent might be able to offer strategy guidance for design tasks. Alternatively, it might provide only lower-level procedural "how to" information.

Initiative: Related to the "role" dimension is the degree to which an agent initiates interactions. Should it proactively make suggestions and offer guidance or should it respond only when directly addressed? A proactive agent might be viewed as being "pushy" and might bother users, or it could be viewed as being extremely helpful and intelligent if it acts in situations in which the user is unsure of how to proceed or is so confused that he or she is unable to form a coherent help request.

Speech quality: Does the quality of the agent's speech affect user impressions of the agent? We speculate that poor quality spoken output might negatively influence user views of an agent. Research on user perceptions of speech quality already exists [9], and that work can provide guidance in designing our agent experiments.

Other variables: Other agent-related variables to consider are "gender" and competence.

3.3 Factor 3: Features of the Task

Tasks can also vary in many different ways. Some tasks can be opinion-like (e.g., choosing what to bring on a trip) while others are more objective (e.g., solving a puzzle) in terms of assessing the quality of a solution. Some involve a good deal of high-level planning (e.g., writing a talk) while others are more rote (e.g., changing boldface words into italics). Tasks must be classified along some or all of the dimensions listed below:

Intent: The user could have a learning goal or alternatively may be carrying out a set of steps in a familiar domain. In the latter, the user might need help with low-level details whereas in the former the user is looking for guidance as to the structure of the domain.

Objectiveness: The situation might be an opinion-based one in which the user is seeking advice and recommendations on some topic (e.g., which items to pack for a trip to Europe). Alternatively, the user might be carrying out an objective task such as simply acquiring facts (e.g., finding the keystroke combination for a particular command in a software application).

Domain: The domain in which the user is working (e.g., editing a paper vs. building a garage) might matter even if all other relevant features (e.g., objectiveness, intent) are held constant.

Focus: An agent's assistance might be directly involved with the primary task upon which a user is engaged. On the other hand, agents might be helpful with "side" tasks such as looking up a phone number quickly while a user attends to some other primary task. Would people perceive an agent as being more useful in one of the scenarios compared to the other?

Timing: While some tasks, such as monitoring events, require regular or constant attention, other tasks such as guiding a presentation require some degree of cooperation between the agent and the user. Some tasks may have a significant delay between their initiation and completion while in other tasks the delay of an agent's actions may arouse user's suspicion.

Other variables: Other task-related variables to consider are duration and consequences of the quality of task performance and available resources and environment.

3.4 Interactions

The number of variables within each factor is certainly larger than the number we have identified here. No doubt these factors will also interact. For instance, a novice attempting to carry out a task in a particular domain might welcome proactive comments/advice from an agent while someone with more experience could get annoyed. Thus, a person packing for her first trip abroad could be pleased to get advice from an agent (or a critique of her packing choices) while a seasoned traveler would be offended by suggestions. While such predictions seem reasonable for a "recommendation" task like packing, the predictions might be reversed for a more objective task such as text editing. Here, a novice, at least one who is interested in learning, might not want help from an agent unless explicitly asked because the novice wants to be an active learner and thereby increase his or her chances of remembering the information. Conversely, an expert would be happy to have the agent take over a set of lower level editing tasks while the expert can concentrate on the overall flow of the argument in the text.

3.5 Approaches to Assessing Usefulness

With respect to measuring the usefulness of an embodied agent, we have to consider which dependent measures are most appropriate. Our framework utilizes two main usefulness dimensions: performance and satisfaction.

Towards the more objective end, a user's performance on a task in terms of accuracy and time – when such measures are meaningful – can give one indication of usefulness. Thus, time and errors would be appropriate measures for a text-editing task. Towards the more subjective end, a user is likely to have a number of affective reactions to an embodied agent. These reactions might manifest themselves in terms of how much users liked the agent, how intrusive they found the agent, how they perceived the agent's personality, and how willing they are to use the agent in the

future. We can certainly assess a user's liking and satisfaction towards an embodied agent (for all tasks for that matter), but if the user can carry out the tasks more effectively with the agent, then how important are liking and satisfaction? On the other hand, long-term use of an embodied agent might be predicted by liking and satisfaction.

The likelihood of a user following an embodied agent's advice might be another interesting measure of the usefulness of an embodied agent. While advice following would certainly be at least partly a function of the quality of the advice, it will also be impacted by how the user feels about the agent (how many children ignore the advice of their parents merely because it is the parents giving the advice?).

4. INITIAL EXPERIMENT

A task that required the user to debate the merits of his or her opinion might lead the user to feel the agent had more of a personality compared to a task in which the user made use of the agent more as a reference tool. Conversely, users might find the agent to be more useful in its role as a reference source rather than as an entity that provides opinions. In addition, the more life-like the agent appeared, the more likely the user might be to ascribe qualities such as personality and intelligence to the agent, but objective performance would likely not be affected by appearance. In the first study, we independently manipulated both the agent fidelity (animated, stiff, iconic) and the task objectiveness (travel task versus editing task). What follows is a brief summary of the study presented more fully in [16]

4.1 Design

The animated agent (see the left side of Figure 1) was a 3D, female appearance (though somewhat androgynous) that blinked, moved its head, and produced certain facial expressions in addition to moving its mouth in synchronization with the synthesized voice. The stiff agent had the same face as the animated agent but only moved its mouth. The iconic agent (see the right side of Figure 1) was a light-bulb icon that had arrows appear whenever it spoke.

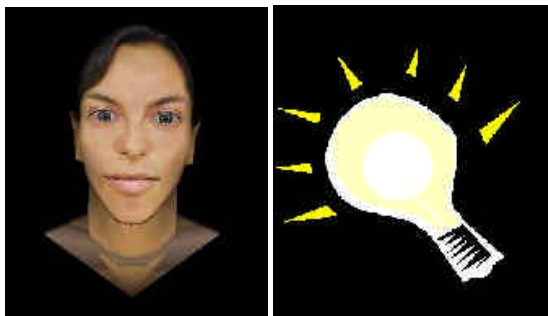


Figure 1: Appearance of Agent in Animated and Stiff Conditions (left) and Iconic Condition (right)

The travel task involved a hypothetical situation in which the participant had a friend who was flying overseas on his first international trip. The task was to recommend six items for the person to take with him from a pool of 12 items and to rank the six items in order of importance. It was chosen to be a type of creative, opinion-based task in which interacting with an agent might be viewed as an opportunity to think more deeply by discussing.

The editing task had participants use an unfamiliar text editor to modify an existing document by making a set of prescribed changes to the document. They were instructed that if at any time they could not remember the keystrokes for a particular function, they could ask the agent for help. The editing task was chosen to represent an opportunity to use an agent primarily as a reference source rather than as a guide or teacher.

As mentioned earlier, the agent was controlled through a Wizard of Oz technique. One experimenter was in the room with the participant to introduce the experimental materials; and a second experimenter was in an adjacent room, monitoring the questions and responses made by the participant. The second experimenter insured that the agent responded in a consistent, predefined manner using a prepared set of replies.

The primary dependent variables in the experiment were the responses to the individual items in questionnaires that addressed a number of qualities of the agent and the answers to the open-ended questions posed by the experimenter. Objective measures were also collected. For the travel task, we measured whether participants changed their rankings as a function of the agent's feedback. For the editing task, we measured how long it took participants to complete the tasks.

4.2 Results

With respect to more objective measures, the analysis shows that participants were more likely to change the rankings of items that the agent disagreed with compared to items that the agent agreed with ($F(1, 33) = 38.37$, $MSE = .07$, $p < .0001$). There was no effect of type of agent. The time (in seconds) to do the editing task did not differ as a function of agent too (animated: 714.8, motionless: 568.7, cartoon: 671.1; $F(2, 31) = 1.78$, $MSE = 37637.22$, $p = .19$)

As regard to questionnaire responses, though participants felt, on average, that the agent helped with the tasks and was worthwhile, there was no effect of agent type for any of the questions. For two of the questionnaire items, worthwhile and intrusive, there was an effect of task (worthwhile: $F(1, 31) = 15.68$, $MSE = .45$, $p = .0004$; intrusive: $F(1, 31) = 20.28$, $MSE = .23$, $p = .0001$). The agent was rated more worthwhile and less intrusive after the editing task compared to the travel task.

The analysis of the responses to interview questions shows that the participants' reactions to the agent again vary as a function of task. Although virtually all participants found the agent helpful for both tasks, participants were much less likely to consider the agent to have a personality after doing the editing tasks. In addition, the agent was perceived as more intelligent after the travel task than after the editing task. Finally, one striking difference in behavior in the interviews was whether a person referred to the agent using words such as "agent" or "it," versus the gender pronouns "she," "her," "he," or "him." Eleven of the 39 participants used the gender pronouns. This behavior reinforced the notion of how people often treat computers as social actors [15]. The study participants included 15 women and 24 men. Curiously, eight of the 11 participants who used the gender pronouns were women and only three were men. Thus, over half the women in the study referred to the agent this way and only 13% of the men did so.

5. CURRENT EXPERIMENT

Our current study investigates the affect of an agent's initiative on people's perception of the agent. Participants were instructed to complete an editing task similar to the one of the first experiment. Each participant worked with an agent under one of the two conditions: 1) the agent gave instructions without asking when the agent could infer about the participant's actions 2) the agent only responded to questions when being asked by the participant. A third condition had participants use a paper help sheet rather than an agent.

Initial analysis of both the objective and subjective data showed that the participants working with the proactive agent felt that it was helpful while initiative of the agent had little effect on their task performance.

6. CONCLUSION

Prior evaluation work on embodied conversational agents has been suffered from a lack of systematicity in examining key factors and used dependent measures that often did not appropriately assess subjective experience and objective performance. We developed a three-factor approach to studying embodied conversational agents. We performed experiments within this framework and will refine our framework to guide future empirical studies. We hope other researchers find the framework useful and that it will allow future experiments to provide more definitive answers about the features of agents, users, and tasks that predict success of embodied conversational agent systems.

REFERENCES

1. André, E. & Rist, T. Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems, in *Proceeding of the Second International Conference on Intelligent User Interfaces* (2000), 1-8.
2. Cassell, J. Embodied conversational interface agents. *Communications of ACM* 43, 4 (April 2000), 70-78.
3. Dahlback, N., Jonsson, A. & Ahrenberg, L. Wizard of oz studies – why and how, in *Proceedings of the 1993 International Workshop on Intelligent User Interfaces*, (Orlando, FL, 1993), 193-200.
4. Dehn, D.M. & van Mulken, S. The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies* 52, 1-22.
5. Dryer, D. C. Getting Personal with Computers: How to design personalities for agents, in *Applied Artificial Intelligence* 13 No. 3 (1999), 273-295.
6. Fridlund, A.J. & Gilbert, A.N. Emotions and facial expression. *Science* 230 (1985), 607-608.
7. King, W.J. & Ohya, J. The representation of agents: Anthropomorphism, agency and intelligence, in *CHI '96 Conference Companion*, (Vancouver, B.C., April 1996), 289-290.
8. Koda, T. Agents with faces: A study on the effect of personification of software agents. Master's thesis, MIT Media Lab, Cambridge, MA, 1996.
9. Lai, J., Wood D. & Michael Considine. The effect of task conditions on the comprehensibility of synthetic speech, in *Proceedings of the ACM CHI 2000*, 321-328.
10. Laurel, B. Interface agents: Metaphors with character, in *The Art of Human-Computer Interface Design*, Laurel, B. (ed.), Addison-Wesley, Reading, MA, 1990, 355-365.
11. Maes, P. Agents that reduce work and information overload. *Communications of the ACM*, 37, 3 (July 1994), 31-40.
12. Massaro, D.W., Cohen, M. M., Beskow, J., & Cole, R.A. Developing and evaluating conversational agents (2000). In *Embodied Conversational Agents*, Cassell, J. et al. (eds.), Cambridge, MA: MIT Press.
13. McCrae, R. & Costa, P. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 1 (1987), 81-90.
14. Nass, C., Isbister, K. & Lee, E. Truth is beauty: Researching embodied conversational agents, in *Embodied Conversational Agents*, Cassell, J., Prevost, S., Sullivan, J., and Churchill, E. (eds.), MIT Press, Cambridge, MA, 2000, 374-402.
15. Nass, C., Steuer, J., & Tauber, E. Computers are social actors, in *Proceedings of CHI '94*, (Boston, MA, April 1994), 72-78.
16. Richard Catrambone, John Stasko, and Jun Xiao. "Anthropomorphic Agents as a User Interface Paradigm: Experimental Findings and a Framework for Research". To be appear in CogSci 2002.
17. Rickenberg, R. & Reeves, B. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces, in *Proceedings of CHI 2000*, (The Hague, Netherlands, April 2000), 329-336.
18. Shneiderman, B., Direct Manipulation Versus Agents: Paths to Predictable, Controllable, and Comprehensible Interfaces, in *Software Agents*, Bradshaw, J.M. (ed.), MIT Press, Cambridge, MA, 1997, 97-106.
19. Takeuchi, A. & Nagao, K. Communicative facial displays as a new conversation modality, in *Proceedings of INTERCHI '93*, (Amsterdam, April 1993), 187-193.
20. Takeuchi, A. & Nagao, K. Situated facial displays: Towards social interaction, in *Proceedings of CHI '95 Conference*, (Denver, CO, May 1995), 450-455.
21. Walker, J.H., Sproull, L., & Subramani, R. Using a human face in an interface, in *Proceedings of CHI '94*, (Boston, MA, April 1994), 85-91.