Enhancing Human-Computer Interaction with Embodied Conversational Agents

Mary Ellen Foster

Robotics and Embedded Systems Group
Department of Informatics, Technische Universität München
Boltzmannstraße 3, 85748 Garching bei München, Germany
foster@in.tum.de

Abstract. We survey recent research in which the impact of an embodied conversational agent on human-computer interaction has been assessed through a human evaluation. In some cases, the evaluation involved comparing different versions of the agent against itself in the context of a full interactive system; in others, it measured the effect on user perception of spoken output of specific aspects of the embodied agent's behaviour. In almost all of the studies, an embodied agent that displays appropriate non-verbal behaviour was found to enhance the interaction.

1 Introduction

An Embodied Conversational Agent (ECA) is a computer interface that is represented as a human body, and that uses its face and body in a human-like way in conversation with the user; see [1] for an overview of the field. The main benefit of ECAs as an interface metaphor is that they allow users to interact with a computer in the most natural possible setting: face-to-face conversation. Reeves and Nass [2] and others have demonstrated that, when computers produce social cues, users will respond socially, even if they are not conscious of this behaviour.

There is no longer any question that the production of language and its accompanying non-verbal behaviour are tightly linked [3–5]. The communicative functions of body language listed by Bickmore and Cassell [6] include conversation initiation and termination, turn-taking and interruption, content elaboration and emphasis, and feedback and error correction; non-verbal behaviours that can achieve these functions include gaze modification, facial expressions, hand gestures, and posture shifts, among others. Many of these behaviours have been implemented on embodied conversational agents to address many of these communicative functions.

In this paper, we investigate the impact of embodied agents on human-computer interaction by surveying a range of systems for which the impact of the agent has been directly evaluated. We first describe the two high-level techniques that have been used to evaluate embodied agents: system-level evaluation and evaluation of individual aspects. We then describe recent embodied-agent systems that have been evaluated using each of these techniques. At the end, we summarise the overall

findings of these studies and draw conclusions regarding the ways in which an embodied agent can enhance the interaction with a computer system.

2 Evaluating Embodied Agents

Embodied agents are generally evaluated by measuring user responses to the output, in or out of context; automated evaluation has not been widely used in this area. Evaluation can take two forms [7]: the behaviour as a whole can be evaluated in the context of a system, or the individual modalities and behavioural aspects can be assessed separately.

System-level evaluation of an ECA has three aspects: the fluency and efficiency of the user-ECA interaction, the subjective user experience, and the effectiveness of the application in achieving its goals. These criteria are similar to the common evaluation criteria for multimodal dialogue systems described by Dybkjær et al. [8]; however, the addition of the ECA adds another modality whose settings can be varied, and that can potentially have an effect on any of the outcome variables.

Evaluation of individual aspects—e.g., non-verbal behaviour, audio-visual speech, or personality and emotion—is necessary to measure whether the way those behaviours are implemented in the ECA is understood by users as intended, and generally uses the following pattern. First, a model is implemented, based on a combination of documented behaviour from the literature and direct observation of human behaviours. The implemented model is then used to generate materials, and a judgement study is performed to measure whether participants are able to perceive the intended content in the output. Perception studies provide a different perspective on ECAs than system-level evaluation: passive observers may receive a different impression of an ECA than users who are directly engaged in interacting with the agent [e.g., 9].

In the following two sections, we describe a number of recent evaluations of each type. In Section 3, we describe a range of evaluations that have looked at the impact of the embodied agent on the overall system: in each case, we first describe the system and the role of the agent within it, and then describe the design and results of the evaluation study. In Section 4, we then summarise studies that have investigated the impact of specific embodied-agent behaviours directly.

3 System-level evaluations

In a system-level evaluation of an embodied agent, participants generally interact with a complete system under one of a range of conditions; various subjective and objective measures are then used to compare the different versions of the system. The conditions that are used can vary: in some cases, the system is run with and without the embodied agent; in others, different versions of the agent implement different behaviours; while in still others, the behaviours of participants interacting with an agent are compared to those of participants interacting with other humans.

3.1 REA: Social Dialogue and Embodiment

REA [10] is an embodied agent that acts as a real estate salesperson, answering user questions about properties in a database and showing users around a virtual house. REA is able to sense the user through cameras and a microphone and produces output including speech with intonation, facial expressions, and gestures of a fully-articulated body. In addition to supporting task-based dialogue in the real-estate domain, REA is also able to engage in social dialogue with the user: "small talk" in which social goals are primary and the task goals are left in the background. The primary function of small talk is to build rapport and trust and to allow the interlocutors to establish a style of interaction. Body language plays a critical role in human-human social dialogue: behaviours such as leaning forward, nodding, smiling, and direct gaze all contribute to feelings of warmth and trust in the interaction.

To measure the impact of REA's social dialogue, Bickmore and Cassell [6] performed an experiment in which participants interacted with one of two possible versions of REA: one version that employed only task-based dialogue, and one that used social dialogue as well. As an additional factor, half of the participants interacted with the fully embodied REA agent, while the other half talked with REA over a telephone. Participants answered a range of questions measuring their subjective impressions of REA, their degree of trust, and the amount that they were willing to spend on an apartment after having used the system. There was a complex interaction between the conditions: one main result was that participants that interacted with REA over the telephone preferred the social dialogue style, while participants that used the embodied version of REA preferred the task-based dialogue. Bickmore and Cassell hypothesise that this is due the body language used by REA inadvertently projecting an unfriendly personality.

3.2 MagiCster: Consistent Affective Facial Displays

The MagiCster project had two main goals: to develop believable embodied conversational agents making use of synchronised gaze, facial expression, gesture, body posture, and speech, and to evaluate agents in laboratory conditions to determine the aspects of the agents that are important for a range of human-computer interactions. As part of this project, the Greta embodied agent [11] was developed; this agent is able to display a range of performative facial displays such as surprise and feeling sorry for the user, and to combine them dynamically as needed to produce more complex expressions.

Berry et al. [12] describe an evaluation of the impact of Greta on users' responses to health-education materials. This study used two messages about healthy eating, one positively framed (emphasising the positive effects of eating well) and one negatively framed (emphasising the negative effects of not doing so). Participants were presented with one of the two messages and asked to rate the utterance on a number of scales, and were also asked several memory questions about the content of the utterance. In general, participants viewing messages presented by Greta performed worse on the memory task than those using a textual or speech-only presentation of the

information; however, when Greta used facial displays that were consistent with the content of the speech, the performance impact was mitigated.

3.3 Gesturing Strategy and Agent Appearance

Buisine et al. [13] describe an experiment in which a range of multimodal gesturing strategies and agent appearances were compared in the context of a system that generated embodied technical explanations—for example, describing how to use a copy machine. The explanations contained many references to particular components of the machines, so there were several opportunities to make multimodal spatial references. Three different gesturing strategies were implemented for spatial references: redundancy (all relevant information is given in both speech and gesture), complementarity (half of the relevant information is given on each channel), and speech-only (all of the relevant information is given in speech, and the gestures have no semantic content). As another factor, agents with three different appearances were implemented: two male agents and one female.

In an evaluation, participants viewed explanations of each type produced by each of the agents; the dependent variables were users' subjective evaluations of the agents and their performance on a recall task. The presentations including speech only were judged to have lower quality than either of the other two types, particularly by the male participants; participants also reported more trust in the presentations that were judged to have higher quality. One of the male agents was perceived as significantly less likeable than the other two. The gender of the participant had a significant effect on their recall performance, but there was little effect of any of the experimental manipulations on this variable.

3.4 COMIC: Turn-Taking and Task Performance

The COMIC multimodal dialogue system adds a dialogue interface to a CAD-like application used in bathroom sales situations to help clients redesign their rooms. The input channels include speech recognition, along with handwriting and pen gestures provided either on a tablet display or with a mouse; the output combines synthesised speech, non-verbal behaviour of a talking head, deictic gestures using an on-screen pointer, and direct control of the underlying application. The nonverbal behaviour of the talking head performs several functions in COMIC. It gazes at objects on the screen as it describes them; it also uses facial expressions to convey the system state, for example using a "thinking" expression while processing input, looking confused if the user input was not understood, and looking happy when it was understood.

White et al. [14] describe an experiment designed to measure whether the non-verbal behaviours of the agent make a difference to users' interactions with the system. As a between-participants factor, the talking head was run in one of two modes: expressive, where all of the non-verbal behaviours of the head were enabled, and zombie, where all behaviours except lip movements were disabled. The results demonstrate that the expressive face mitigated the perception of slow system response and attracted the user's gaze more frequently. However, participants' performance on

a recall measure was significantly lower with the expressive head than it was with the zombie head, especially for the male participants (whose overall recall performance was also lower).

3.5 Mel: Engagement in the Interaction

Mel [15] is a robot designed to resemble a penguin wearing glasses who acts as a host for visitors to a research lab. Its capabilities include participating in spoken dialogue, locating and tracking the position and gaze direction of the visitor, interpreting and responding to nodding behaviour during a conversation, and pointing and looking at objects in the environment. Mel leads the user through a demonstration of a research prototype, giving instructions on its use and describing how it works and how it can be useful. A specific goal for Mel was to maintain user engagement in the interaction. The level of engagement was monitored by tracking the verbal and non-verbal behaviour of the user, while a set of recipes based on observed human-human interactions were used to maintain engagement—for example, by looking at the user's face or repeating a prompt when the user attention wavers.

A study was conducted comparing the experience of two groups of users interacting with Mel. For one group, Mel employed its full range of non-verbal behaviour (mover mode); for the other group, the only motions were movements of the lips to accompany the speech (talker mode). These conditions are similar to the expressive and zombie conditions used in the evaluation of the COMIC system described above. The responses on a post-interaction questionnaire indicate that female participants reported more engagement in the interaction, regardless of condition; that participants in the talker condition found the robot more reliable; and that participants in the mover condition found the robot motions more appropriate. When the recordings of the interaction were analysed, it was found that participants in the mover condition had significantly longer interactions, coordinated their gaze with the robot more frequently, and looked back at the robot significantly more often.

3.6 GAMBLE: Deceptive Body Language

GAMBLE [9] is a scenario in which two human users play a dice game against an embodied agent. To win in this game, it is essential both to lie to your fellow players and to detect when they are lying to you. For the evaluation described by Rehm and André [9], two different versions of the agent were implemented: one that shows clues to deception in its facial expressions, and one that does not. In an initial study where the expressions were presented out of context, participants were able to detect these deceptive expressions; this study is summarised in Section 4. However, in the game context, there was no difference between participants' responses to the two versions of the agent: they caught the agent lying about 73% of the time for both versions, and incorrectly accused it of lying when it was telling the truth 54% of the time for both. Rehm and André propose that the lack of effect in this case is due to the fact that participants were concentrating on the task, rather than the details of the agent's expression.

In another analysis of the data from these interactions [16], it was found that participants looked at the agent approximately as often as they looked at the other human player when they were speaking to it. However, when addressed by the agent, participants looked at it much more often than they looked at the other human player when addressed by them.

3.7 MIT FitTrack: Relational Agents

The MIT FitTrack system [17] was designed to investigate the role of relational agents—embodied agents that are intended to build and maintain long-term social-emotional relationships with users. The system used an embodied agent to play the role of an exercise advisor who discussed physical activity with the user and encouraged them to become more physically active. The system output combined synthesised speech and synchronised non-verbal behaviour, using many of the same components as the REA system described in Section 3.1. It employed a range of behaviours in an effort to build and maintain a relationship with the user, including small talk, addressing the user as a friend, and displaying empathy and humour.

The effectiveness of the system was evaluated via a longitudinal study in which participants used the system daily for a month. Participants used the system in one of two modes: relational, in which the system employed its full range of relational behaviours in an effort to build a working relationship with the participant, and non-relational, in which the relational behaviours were all disabled. The outcomes were measured through a set of subjective questionnaires, and also by measuring the participants' exercise level before, during, and after the study. The participants in the relational condition displayed higher levels of trust and liking for the agent than did those using the non-relational version. Both groups of participants that used the FitTrack increased their physical activity more than a control group who did not use the system, but there was no significant difference between the two FitTrack groups on this measure.

3.8 iCat Home Companion: Social Intelligence

De Ruyter et al. [18] studied a home dialogue system that used the iCat user-interface robot to program a DVD recorder and participate in an online auction. The iCat is a platform for studying social human-robot interfaces: it is equipped with servos that can control many parts of the face to produce expressions, and is also able to produce lip-synchronised synthesised speech. A range of socially intelligent behaviours were implemented for the iCat: displaying attentive listening behaviour, responding to non-verbal cues of the user, and using expressive facial displays where appropriate. In the evaluation, the socially intelligent system was compared against a socially neutral version that used no facial displays apart for lip-synchronisation and that did not respond to or exhibit social cues. The results of the study indicated that participants rated the socially intelligent version as more social than the neutral version, and also expressed more satisfaction with the DVD player when they used the social iCat.

3.9 Evaluation via Physiological Measures

Prendinger and colleagues have performed several studies in which the impact of embodied agents on an interaction is measured through physiological measures of the participants. In [19], they used a web-based system that displayed pictures of a Tokyo apartment and described them to a user. The system had three different presentation modes: an embodied agent that used synthesised speech and simple left/right deictic gestures, combined speech and incrementally-displayed text (with the same content in both), and speech only; the same synthetic voice was used for all presentation modes. They used eye-tracking to compare participants' responses to descriptions presented in the three modes. When the agent was present, participants followed its verbal and non-verbal navigation directions and mostly looked at the agent's face rather than other parts; with written text, participants looked at the box nearly twice as often as they did at the agent. On a questionnaire, participants rated the voice-only presentation to be significantly more useful for the task than either of the other presentation types.

In another study, Prendinger et al. [20] implemented a mathematical quiz game that included an embodied agent acting as quizmaster. The game included delays designed to frustrate the user (who was told that the game was in a development stage and could have bugs). The agent was able to produce a range of affective feedback, including apologising for delays in the system (with apologetic body language), looking happy when an answer was correct and sad when it was wrong, and using more polite language. In an experiment, participants used the system in one of two modes: with all of the affective feedback enabled, and with it all disabled. The results indicated that the affective agent reduced the stress of participants as measured by galvanic skin response, and also led participants to experience the quiz as less difficult.

4 Evaluation of Individual Aspects

In the previous section, we listed a number of studies where the impact of an expressive agent is evaluated in the context of an entire system. Other studies have looked at specific aspects of agents in isolation: for these studies, the general strategy is to test whether participants are able to perceive the intended prosodic or affective content of spoken output based on the body language of the agent that accompanies the utterance.

In a series of studies, Swerts and Krahmer [21] have investigated the influence of facial displays on users' perception of the prosody of an utterance. They found that congruent speech prosody and visual cues (nodding and eyebrow raises) are preferred to conflicting cues on these two channels; that correct facial displays enhance participants' ability to perceive stress in speech; and that the upper part of the face and the left side are the most relevant for perceiving the intended prosody.

Foster [22] has investigated the influence on facial displays of the intended user-model evaluation in the context of the COMIC multimodal dialogue system; this system is described in detail in Section 3.4. The studies used the RUTH talking head

[23] to compare different methods of using data from a single-speaker corpus to select facial displays based on the intended user-model evaluation and other contextual factors. The results of these experiments demonstrate that participants are able to identify the intended user-model evaluation based on the motions of the talking head, and that they prefer outputs where the user model expressed in speech matches the facial displays.

Rehm and André [9] investigated the impact of facial cues to deception on users' responses to an agent, using the Greta talking head developed in the MagiCster project (Section 3.2). First, a set of deceptive facial expressions were created by "masking" negative emotions (anger, fear, disgust, sadness) with a smile. For the experiment, two versions of a number of videos were created in which Greta presented movie reviews: in one version, true smiles were used to accompany positive sentences such as I liked the happy ending, while the other version used expressions in which a smile was used to mask disgust; the same synthesised speech was used for both versions. Participants were shown a series of videos, half presented by each version of the agent. The version using deceptive facial displays was judged to be less reliable, less trustworthy, less convincing, less credible, and less certain about what it was saying. Note that, as described in Section 3.6, these facial displays did not make any difference in a task-based context.

Marsi and van Rooden [24] implemented selected facial signals of uncertainty on the RUTH talking head in the context of a multimodal question-answering system, and created videos including just the eyebrow motions, just the rigid head motions, or both. In an experimental study, participants were asked to use a five-point scale to rate the uncertainty level for a series of videos. Participants generally rated the videos intended to express certainty as being more certain than those intended to express uncertainty; however, when the eyebrow movements are investigated specifically, the brow movements designed to be uncertain were actually judged to be displaying certainty.

5 Summary

In general, the results of the studies described here indicate that an embodied agent can improve user satisfaction and engagement with a computer system, and in some cases (e.g., with the iCat companion) can even improve users' opinion of the objects being described by the system. Agent body language can also influence users' perception both of the prosody of synthesised speech and of various affective aspects of the speech.

However, not all of results were entirely positive: some agent implementations did not make any difference (e.g., the deceptive body language in the dice game), while others actually had an effect counter to what was intended (e.g., the inadvertently unfriendly body language of REA, the penalty on task performance for male participants with the expressive COMIC head, and the eyebrow motions that were judged to indicate certainty rather than uncertainty). In many cases, there was a differential effect of gender on participants' responses to the agent: female users

generally seem to respond more positively to expressive embodied agents than do male users.

In summary, then, these studies demonstrate adding an expressive embodied interface agent to a computer system can often have a positive effect on users' interactions with that system. However, they also show that it is vital to test any particular agent implementation to ensure that it is having the intended effect on the target user group.

6 Acknowledgements

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST" (FP6-003747-IP), http://www.euprojects-jast.net/.

References

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E., eds.: Embodied Conversational Agents. MIT Press (2000)
2. Reeves, B., Nass, C.: The Media Equation: How people treat computers, television, and new media like real people and places. Cambridge University Press (1996)
3. Bavelas, J. B., Chovil, N.: Visible acts of meaning: An integrated message model of language in face-to-face dialogue. Journal of Language and Social Psychology 19(2) (2000) 163–194. doi:10.1177/0261927X00019002001
4. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press (2004)
5. McNeill, D., ed.: Language and Gesture: Window into Thought and Action. Cambridge University Press (2000)
6. Bickmore, T., Cassell, J.: Social dialogue with embodied conversational agents. In van Kuppevelt, J., Dybkjær, L., Bernsen, N. O., eds., Advances in Natural, Multimodal Dialogue Systems. Kluwer, New York (2005)
7. Ruttkay, Z., André, E., Johnson, W. L., Pelachaud, C., eds.: Evaluating Embodied Conversational Agents, number 04121 in Dagstuhl Seminar Proceedings (2006)
8. Dybkjær, L., Bernsen, N. O., Minker, W.: Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43(1–2) (2004) 33–54. doi:10.1016/j.specom.2004.02.001
9. Rehm, M., André, E.: Catch me if you can – exploring lying agents in social settings. In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (2005), 937–944
10. Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., Yan, H.: Human conversation as a system framework: Designing embodied conversational agents. In [1], 29–63
11. Poggi, I., Pelachaud, C.: Performative facial expressions in animated faces. In [1], 154–188
12. Berry, D. C., Butler, L., de Rosis, F., Laaksolathi, J., Pelachaud, C., Steedman, M.: Final evaluation report. Deliverable 4.6, MagiCster project (2004)
13. Buisine, S., Abrilian, S., Martin, J.-C.: Evaluation of individual multimodal behaviour of 2D embodied agents in presentation tasks. In [25], 217–238. doi:10.1007/1-4020-2730-3 8
14. White, M., Foster, M. E., Oberlander, J., Brown, A.: Using facial feedback to enhance turn-taking in a multimodal dialogue system. In Proceedings of HCI International 2005 Thematic Session on Universal Access in Human-Computer Interaction (2005)

15. Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1–2) (2005) 140–164. doi:10.1016/j.artint.2005.03.005
16. Rehm, M., André, E.: Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent. In Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA 2005) (2005), 241–252. doi:10.1007/11550617 21
17. Bickmore, T. W., Picard, R. W.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction 12(2) (2005) 293–327. doi:10.1145/1067860.1067867
18. de Ruyter, B., Saini, P., Markopoulos, P., van Breemen, A.: Assessing the effects of building social intelligence in a robotic interface for the home. Interacting with Computers 17(5) (2005) 522–541. doi:10.1016/j.intcom.2005.03.003
19. Prendinger, H., Ma, C., Yingzi, J., Nakasone, A., Ishizuka, M.: Understanding the effect of life-like interface agents through users' eye movements. In Proceedings of the 7th international conference on Multimodal interfaces (ICMI 2005) (2005), 108–115. doi:10.1145/1088463.1088484
20. Prendinger, H., Mori, J., Ishizuka, M.: Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. International Journal of Human-Computer Studies 62(2) (2005) 231–245. doi:10.1016/j.ijhcs.2004.11.009
21. Swerts, M., Krahmer, E.: On the perception of audiovisual cues to prominence (In Press)
22. Foster, M. E.: Evaluating the impact of variation in the generation of multimodal object descriptions. Ph.D. thesis, School of Informatics, University of Edinburgh (2007). In submission
23. DeCarlo, D., Stone, M., Revilla, C., Venditti, J. J.: Specifying and animating facial signals for discourse in embodied conversational agents. Computer Animation and Virtual Worlds 15(1) (2004) 27–38. doi:10.1002/cav.5
24. Marsi, E., van Rooden, F.: Expressing uncertainty with a talking head. In Proceedings of the Workshop on Multimodal Generation (MOG 2007) (2007)
25. Pelachaud, C., Ruttkay, Z., eds.: From Brows to Trust: Evaluating Embodied Conversational Agents. Springer (2004). doi:10.1007/1-4020-2730-3