

# Empirical Evaluation of Animated Agents In a Multi-Modal E-Retail Application

Helen McBreen, Mervyn Jack

Centre for Communication Interface Research,  
The University of Edinburgh,  
80 South Bridge, EH1 1HN, Scotland, UK  
+44 131 650 2779

Helen.McBreen@ccir.ed.ac.uk, Mervyn.Jack@ccir.ed.ac.uk

## Abstract

This paper presents the results of an empirical evaluation of the effectiveness and user acceptability of a selection of animated agents in a multi-modal electronic retail application. Male and female versions of six animated agent technologies were repeatedly evaluated in the role of an interactive conversational sales assistant. The experiment involved participants eavesdropping on spoken dialogues between a 'customer' and each of the animated assistants. Participants then completed usability questionnaires and took part in a debriefing interview designed to elicit information relating to the agents' voice, aspects of personality, appearance, facial expressions and gestures.

## Introduction

This paper presents the results of an experiment aimed at assessing usability attributes of twelve personified agents in the context of a multi-modal electronic retail application by having participants 'eavesdrop' in turn on brief dialogues between a customer (represented by a disembodied voice) and each of the twelve agents. The retail application (Figure 1) was created in the style of MUESLI (Wynd and Churcher 1999). The main window was a 3D view of a living room complete with furniture. Immediately above was a row of fabric and wallpaper samples that could be selected in order to 'decorate' the walls, sofa, chairs and curtains. The dialogues were designed to illustrate the 'customer' conversing with the agent to select colours and patterns in order to decorate the room.

The twelve agents differed in terms of their gender and visual appearance. Male and female versions of six different agent types (T) were created and are described in more detail here:

**T1:** Disembodied voice.

**T2:** 2D graphically animated head, appearing to the left of the 3D room. This agent had lip movement synchronised to the speech output. The agent blinked and smiled occasionally at appropriate times during the course of the conversation with the customer.

**T3:** 3D graphically animated head, appearing to the left of the 3D room. This agent had synchronised lip movement, blinking, and because it nodded and turning slightly its 3D appearance was evident.

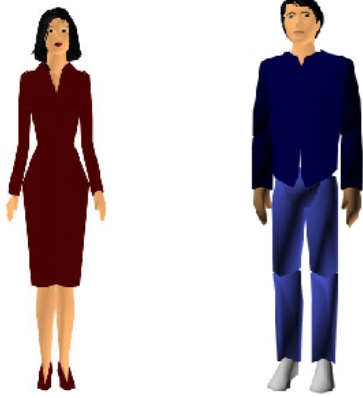
**T4:** 2D graphically animated full-bodied agent using the heads of T2, appearing outside the 3D room.

**T5:** 3D graphically animated full-bodied agent using the heads of T3, appearing outside the 3D room.

**T6:** 3D graphically animated full-bodied agent identical to T5, appearing inside the 3D room.



Figure 1: The E-Retail Application Interface with Animated Agent



**Figure 2: Female and Male Embodied Agents (used for T4, T5 and T6)**

All the agents that appeared on the screen (T2-T6) were provided with facial expression based on four categories defined by Cassell (Cassell et al. 1998). The 2D and 3D animated heads directed gaze toward the customer at all times as the dialogue turns were short. Head nods were used to add emphasis although this was more evident in the 3D head and embodied agents and eyebrow raising was also included at appropriate pitch accents. Further non-verbal facial feedback displays were included for all the agents allowing the agents to look toward the customer during pauses and when asking questions. The agents also looked toward the customers at the end of a turn. The 2D and 3D heads and 2D embodied agents maintained mutual gaze with the customer. However the 3D embodied agents were more mobile, they turned to look at the changes being made in the 3D room, but always returned to look at the customer at the end of the utterance. Gesturing was included in the design of the 2D and 3D embodied agents. Specifically, propositional and spontaneous gestures including iconic, metamorphic and deictic gestures were used (Cassell et al. 1998).

## Experimental Procedure

The experiment used a repeated measures, balanced order design in which 36 participants, balanced for gender, 'eavesdropped' on a short dialogue between a customer and each of the agents in turn (McBreen et al. 2000). Following each dialogue participants completed a 7 point (agree-disagree continuum). Likert format usability questionnaires with a maximum of 22 items, were taken as the dependent variables for the purposes of analysis (Likert 1932). The independent variables were agent gender and type. A post experiment interview was also conducted to obtain detailed qualitative data on users' responses to the agents.

Twelve similar but not identical dialogue scripts were created. One male voice was used as the voice for all the

male agents, and one female voice was used as the voice of all the female agents. An example of the dialogue between the customer and the assistant is presented in Figure 3.

|           |  |
|-----------|--|
| Assistant | <i>Hello</i>   |
| Customer  | <b>Hi, can I see a selection of curtain materials please?</b>          |
| Assistant | <i>Would you like to choose one of these fabrics or see some more?</i> |
| Customer  | <b>Show me the Camellia Stripe</b>                                     |
| Assistant | <i>Do you like it?</i>   |
| Customer  | <b>Not really, change it to Wesley Dual.</b>                           |
| Assistant | <i>OK</i>  |
| Customer  | <b>Show me some matching sofa fabrics please.</b>                      |
| Assistant | <i>Certainly, here you are</i>   |
| Customer  | <b>Show me some more</b>   |

**Figure 3: Section of the Dialogue**

## Results

The attitude questionnaire addressed five main issues regarding the agent: voice, personality, appearance, facial expressions and gesturing.

### Agents' Voice

The female voice was significantly ( $F = 7.37, df = 1.0, p < 0.01$ ) preferred to the male voice (mean female = 5.28, mean male = 4.84). Qualitative comments indicated that both voices were clear. However there were comments which suggested that the female voice was more friendly and natural. The male was perceived to be slightly monotonous. There was a greater preference by female participants for the female agents ( $F = 8.645, df = 1.0, p < 0.01$ ). The male participants however liked the male and female voices equally. T-Tests indicated that the male voice of T4 and T6 ( $p < 0.01$ ) were significantly less preferred to their female counterparts, suggesting that the male embodied characters may be less popular.

The voice of the male agents were perceived to be significantly less natural than the female ( $F = 13.53, df = 1.0, p < 0.01$ ), with the female participants preferring the female voice. Statistical results showed that the female voice was less annoying than the male voice, ( $F = 10.96, df = 1.0, p < 0.05$ ).

The questionnaires also showed significant differences between agent types with respect to the naturalness of the voices: an indication that the voice of the agent may effect participants' attitudes towards the appearance of the agent. The voice of T5 was the least annoying, in fact T-Tests confirmed that the voices of T5 were significantly preferred to T3 and T4.

## Agents' Personality

**Politeness:** Participants felt that all agents were polite, but the quantitative results showed some significant agent type differences, ( $F = 5.17$ ,  $df = 5.0$ ,  $p < 0.01$ ). T5 was significantly more polite than T2, T3 and T4, but it had similar mean scores to T1 and T6. The disembodied voices and the 3D embodied characters were thought to be more polite. This suggests that gesturing may play an important role in participants' perceptions of politeness. An interaction between participant gender and agent gender, ( $F = 9.15$ ,  $df = 1.0$ ,  $p < 0.01$ ), showed that female participants thought that female agents were more polite than male agents, and male participants thought that male agents were more polite than female agents.

**Friendliness:** There were significant differences between agent types for perception of friendliness, ( $F = 3.15$ ,  $df = 50$ ,  $p < 0.05$ ), and also an interaction between agent type and agent gender, ( $F = 2.74$ ,  $df = 5.0$ ,  $p < 0.05$ ). Mean scores show that in general T5 and T6 were deemed to be most friendly. In fact, T-Tests showed that T5 and T6 were significantly more friendly than T2 and T3, all at  $p < 0.01$ , yet another indication that gesturing plays an important role in participants' perceptions of characteristics such as friendliness for embodied agents.

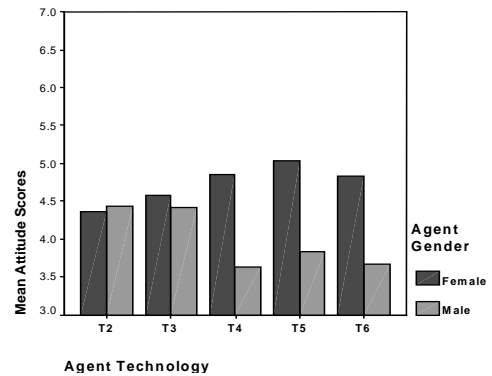
**Competence:** Female agents were perceived to be more competent than male agents. Because participants did not like the male voice as much as the female overall, they may have associated this with his competence (mean female = 5.657, mean male = 5.52). This was specifically the situation for T1, T4, T5 and T6, and significantly the case for T4. With respect to the talking heads (2D or 3D), participants thought the male and female agents were equally competent, but for the fully embodied agents participants felt that the female was more competent.

**Forcefulness:** T5 was significantly less forceful than all other technologies ( $p < 0.01$ ), and T4 was significantly more forceful than all other technologies, ( $p < 0.01$ ). Many participants said that agents who made suggestions were more helpful. Although the dialogue scripts were slightly different, the scripts of T4 did make specific suggestions about fabrics (e.g., "Would you like to try Stella Dual instead"). It seems feasible, based on the analysis of the qualitative feedback that participants could have felt that the assistant was indeed too forceful. The scripts for the other technologies did not make such specific selections, instead they asked if the customer would like to "try another", without saying the specific name of a fabric.

## Agents' Appearance (T2-T6 only)

The appearances of the female agents were preferred to the male agents, ( $F = 22.0$ ,  $df = 1.0$ ,  $p < 0.01$ ). More specifically, gender differences occurred significantly for agent types T4, T5 and T6 (fully-embodied agents), where the female appearance was significantly preferred to the male, all  $p < 0.01$ . Upon analysis of the qualitative

interview data, many participants thought that the male embodied agents used hand gestures that dominated the interface.



**Figure 4: Attitude to Agent Appearance**

The female agents were thought to be significantly more helpful (mean female = 4.617, mean male = 4.294). The mean score for T5 suggests that it was the most helpful of all agents.

Female agents were perceived to be more suitable for the Home Furnishings application than the male agents. A combination of poor attitude responses to the male voice and the exaggerated gestures of the male embodied agents were the probable cause of this gender bias. Female participants thought the female agents were more suitable for the application than the male agents. Male participants thought the male and female agents were equally suitable. T-Tests showed that the female embodied characters (T4, T5, T6) were more suitable than the male embodied characters, ( $p < 0.01$ ). Figure 4 shows these findings graphically.

## Facial Expressions (T2-T6 only)

Significantly the lip movements of T5 and T6 were less distracting than T2, T3 and T4, ( $F = 3.996$ ,  $df = 4.0$ ,  $p < 0.01$ ). In the interviews, many participants said that that the lip movements were distracting because they looked dubbed and that the lip movements of the talking heads were more noticeable but they looked artificial. It also became apparent that the female agent was rendered as if wearing lipstick, her lips were more noticeable and obvious.

Significant results for agent type differences ( $F = 3.692$ ,  $df = 4.0$ ,  $p < 0.05$ ), showed that the facial expressions of T3 and T5 appeared to be the most life-like. T-Tests showed that the facial expressions of T3 were significantly more life-like than T2, T4 and T6, ( $p < 0.01$ ). Even though the face of T5 was smaller, making it more difficult for participants to evaluate, it still had a mean score that was similar to that of T3, a talking head, where the facial expressions could be clearly seen. T2 had the most distracting lip movement, significantly less so than the embodied characters,  $p < 0.01$ .

Participants thought that the female T2 had the most noticeable smile; her lip colour seemed to attract attention. In fact, again showing that the facial expressions of T2 were noticeable, it was found that they were unhelpful in comparison to other agent types, namely T3 and T5,  $p < 0.05$ . It may be concluded from this that less obvious and more natural expressions could promote a sense of helpfulness in animated agents. It was found that the facial expressions of the agents were friendly, and that none were less friendly than others. T5 had the highest mean score (5.07).

### **Gestures (T4-T6 only)**

Participants preferred the female gestures to the male gestures (mean female = 3.898, mean male = 3.694) and also thought they were less exaggerated than the male gestures. The gesturing of T4 was less exaggerated than T5 and T6,  $p < 0.05$ . T4 only had pointing gestures, and no spontaneous gestures. It seems that deictic gestures may be more useful in e-retail interfaces. Many participants felt that the males' gestures were exaggerated which may explain why participants did not like the male appearance. The results did show that for T4 and T6, this gender divide was significant,  $p < 0.01$ , but for T5 both male and female agents had similar mean scores. There were significant results for technology when asked if the gestures made the assistants appear life-like, with T5 being more life-like than T6, and significantly more so than T4,  $p < 0.01$ .

The gestures contributed to the perceived friendliness of the agents. The agents with more gestures (T5 and T6) were significantly more friendly. Qualitatively it was found that participants wanted the agents to be friendly, a feature that can be promoted by the use of gestures. However, the point at which gesturing becomes annoying and undermines the perception of friendliness remains to be investigated. As regards gesturing effecting the helpfulness of the assistants, analysis showed that T5 was most helpful. Besides the fact that male T6 was poorly accepted because of his large gestures, T5 and T6 were still rated strongly and were thought to be more helpful than T4. T5 was significantly more helpful than T4,  $p < 0.01$ .

## **Discussion**

Half of the participant population preferred the female agents. This fact is related to participants' dislike of the male voice and the poor perception of the embodied male agents' exaggerated gestures. What is of particular interest is the fact that on many occasions the female participants significantly preferred the female agents, suggesting that female participants may prefer to interact with agents of their own gender. This bias was not apparent for male participants.

In previous experiments (McBreen et al. 2000) where different voices were used, the preference was for the male voice, with results suggesting that fluent conversational

voices with intonation should be used. Such voices were therefore used in this simulation. The female voice was acceptable and was 'more friendly', however the male voice seemed to be 'monotonous'. A finding that clearly emerged from this experiment was that it is important to select the voices of the agents carefully, as possible cross-modal effects may occur: if the voice is not favoured it could effect user perceptions of the visual display of the agents.

In general there was a preference for the female agents over male agents as regards personality. The embodied characters were thought to be more friendly, helpful and even polite. This raises interesting research questions about the perception of embodied agents' personality and the movement of their bodies.

The dialogue between the 'customer' and the 'assistant' was by design, sophisticated and included system initiative. The agent was assumed to be capable of making intelligent suggestions to assist the user in decorating the room. It can be seen from the results that competence and forcefulness are in some respect reflected in the speech output from the agent. It is for this reason that it is important to design dialogues where the agent can make suggestions, without being too forceful. Forcefulness can be off-putting for the participant, and in an interactive situation such an agent could undermine the participant's perception of its ability as an assistant.

In general, participants did think that the agents enhanced the service. Despite the poor acceptability of the male embodied agents, it was obvious from the results that participants had a preference to interact with embodied agents as opposed to talking heads. The inclusion of a full body (with non-verbal behaviour associated with such) can indeed enhance the interaction between the user and the agent. However, one third of this participant sample did not like to see the animated agent in the interface, for this reason interactive systems should cater for this by perhaps allowing users to turn off the visual display as required. However, if the non-verbal behaviour can be developed enough to provide essential information to the user, the user may prefer to look at the agent more. In some cases, participants felt that the 3D embodied agent distracted them from the task.

The experiment also investigated the relative desirability of an embodied 3D agent appearing inside the 3D world. For this type of shopping application, especially given that participants were asked to 'eavesdrop' on the dialogues between the customer and the user, the results are not conclusive about whether or not it would be more desirable for 3D agents to appear outside, rather than inside the 3D world. Half of the participants stated a preference to see the agent inside the 3D environment, saying that the agent in the room was 'more complete', 'natural' and 'they added to the realism'. Those who preferred to see the agent outside the 3D world commented that for this application having the agent inside the world distracted their attention from the task and they could not see what was happening in the 3D environment. This was especially a problem for the male

agent (T6) as this agent was larger than its female counterpart and also had a more extreme range of gestures.

## Conclusion

A number of hypotheses were made before running this experiment. It was predicted that there would be a dichotomy in the participant group about the use of animated agents in e-retail applications. Encouragingly, two thirds of the user group did prefer to see an animated character in the interface, as opposed to just hearing a disembodied voice. It was hypothesised that 3D talking heads would be preferred to 2D talking heads. This was indeed the case, because the facial expressions are more natural and users do have a desire to interact with more natural looking characters. It was also shown that facial expressions promote helpfulness and friendliness. Further, the facial expressions of a 3D talking head are more life-like than a 2D talking head.

It was also hypothesised that 3D embodied agents would be preferred to 2D embodied agents. This was indeed found to be the case. Although 2D deictic gestures are helpful, the integration of other non-verbal behaviour in a 3D embodiment can promote the acceptability of that agent. It was also hypothesised that 3D embodied agents in a 3D world would be more acceptable than 3D embodied agents outside a 3D world. For reasons outlined in the discussion, this in fact turned out not to be the case.

This research has provided the agents community with new facts about various aspects of the inclusion of animated agents in e-retail environments and several issues have been raised. Future work will aim to develop acceptable agents that will appear in a selection of *interactive* e-retail environments to investigate further non-verbal behaviour of such e-retail assistants and also their use in various retail applications.

The following summaries the main issues that evolved from this experiment to assist with the development of an acceptable agent to appear in interactive interfaces:

- If using a human voice, consider and choose it carefully. The voice should be fluent, conversational with intonation.
- Ensure the dimensions of the character are in proportional with the dimensions of the 3D environment in which it appears.
- When using animated humanoid talking heads or embodied agents, 3D agents are more appealing to the user than 2D agents.
- Carefully select appropriate gestures and facial expressions, too much could undermine the user's perception of the agent.
- When designing interactive agents, it may be appropriate for the user to personalise the agent with whom they are going to interact.

## Acknowledgements

The authors would like to acknowledge the financial support for this research from BT under its Strategic University Research Initiative, and the helpful planning of the experiment made by the research staff at BT Adastral Park. The work has benefited greatly from helpful discussions with colleagues at the Centre for Communication Interface Research, in particular Dr. John Foster.

## References

- Andre, E., and Rist, T. 1998. Personalising the User Interface: Projects on Life-like Characters at DFKI. In Proceedings of Workshop on Conversational Characters, 167-170, Tahoe City, California.
- Cassell, J.; Bickmore, T.; Billinghamurst, M.; Campbell, L.; Chang, K.; Vilhjalmsson, H.; and Yan, H. 1998. An Architecture for Embodied Conversational Characters. In Proceedings of Workshop on Conversational Characters, 21-30, Tahoe City, California.
- Cassell, J. 1998. *Embodied Conversational Agents*. MIT Press.
- King, J., and Ohya, J., 1996. The Representation of Agents: Anthropomorphism, Agency and Intelligence. In Proceedings of *CHI*.
- Lester, J., and Stone, B., 1997. Increasing Believability in Animated Pedagogical Agents. In Proceedings of Agents Conference, Marina del Rey, CA USA.
- Likert, R., 1932. *Some Applications of Behavioural Research*. Paris Press.
- McBreen, H.; Shade, P.; Jack, M.; Wyard, P.; 2000. Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications. In Proceedings of Autonomous Agents 2000, 39-45, Barcelona, Spain.
- McBreen, H., and Jack, M., 2000. Animated Conversational Agents in E-Commerce Enterprises. In Proceedings of Third International Workshop on Human-Computer Conversation, 112-117, Bellagio, Italy.
- Rust, J., and Golomok, S., 1989. *The Science of Modern Psychological Assessment*.
- Thorisson, K., 1996. Communicative Humanoids: Model of Psychosocial Dialogue Skills. Ph.D. Thesis MIT Media Laboratory.
- Wyard, P., and Churcher, G., 1999. The MUESLI Multimodal 3D Retail System. In Proceedings of ESCA Workshop on Interactive Dialogue Systems.