

Evaluating 3D Embodied Conversational Agents In Contrasting VRML Retail Applications

Helen McBreen, James Anderson, Mervyn Jack

Centre for Communication Interface Research,

University of Edinburgh,

80, South Bridge, EH1 1HN

+44 131 650 2779

{Helen.McBreen, James.Anderson, Mervyn.Jack}@ccir.ed.ac.uk}

ABSTRACT

This paper discusses the results of an empirical evaluation assessing 3D male and female embodied conversational agents that were formally dressed and casually dressed in three interactive VRML retail environments – a cinema box-office, a travel agency and a bank. Participants completed tasks in each application by conversing with the agents after which they completed questionnaires regarding their attitude toward the retail applications, the agent's behaviour in the retail application and the agent's voice, personality and appearance.

Results showed that participants enjoyed speaking to the agents in all three applications. The agents were rated similarly in all three applications and the claim that embodied agents have a role to play as assistants in retail applications is supported. However, qualitative results suggested that participants found it difficult to trust the agents in the banking application. Participants also expressed a desire for agents in the cinema application to be casually dressed and those agents in the banking application to be formally dressed.

1. INTRODUCTION

Retail applications are good domains for assessing user perceptions toward embodied agents as the agents can play the role of a conversational retail assistant. They also offer many conversational possibilities between agents and users. Applications such as banking, making travel arrangements and booking tickets are all common, but contrasting telephone activities. With the expansion of Internet technologies (speech and graphics), and the growing evidence [1, 2, 4, 7, 8] that embodiment can enhance interfaces, in the future tasks in retail applications on the Internet may successfully be completed in collaboration with an embodied agent.

A number of predictions were made prior to the experiment. The first was that participants would believe that ECA's have a role to play as assistants in the retail applications. This prediction was made based on the results of previous experiments, where customers passively viewed conversational agents in retail spaces

[7,8] and indicated a desire to actually converse with them in an interactive situation. A second prediction was that participants would enjoy speaking to the agents equally in all three applications. This prediction was made based on the fact that the agents were designed to offer the same enhancement in each application, i.e. assisting the user with their tasks. Thirdly it was hypothesised that the stereotypes created (formal and casual) would be better suited to different application environments. In general assistants in cinema box offices dress casually and those in banks more formally. It was predicted that the situation in the virtual environments would mirror these real life scenarios.

A further prediction was that male and female participants would respond similarly to the agents, and the responses would be similar for both the male and female agents, i.e. no differences would emerge based solely on the gender of the agents. A final prediction was that the perception of the agents' personalities would be identical within applications as the agents (male and female; formal and casual) were built on the same platform and had the same verbal and non-verbal behaviour. Nass et al [9] showed that participants are responsive to gender in the interface, and that it is possible for gender stereotypes to emerge based on the content of the speech output from an interface. To prevent such gender differences occurring the system described in this paper ensured that both the male and female agents had the same output utterances and these utterances were as neutral as possible, with no intentional inclusion of utterances that were regarded as being more masculine than feminine or vice versa.

2. SYSTEM DESIGN

The architecture is based on a client-server system. Using a Nuance™ speech recogniser, the users speech input is captured on the client PC. A Java-based dialogue manager, which is run from another PC networked to the server, controls the direction of the dialogue as the user completes a task. The code to describe the virtual retail applications is stored on the server PC.

2.1 Creating the VRML Environments

Virtual Reality Modelling Language (VRML) is an excellent software tool for the creation of interactive simulations that incorporate animation, and real-time user participation. The 3D retail worlds were created using VRML97, the international standard file format for describing interactive 3D multimedia on the Internet. The three applications (cinema, travel agency and bank) were identical at the core, they were identical in dimension, and the assistants appeared in identical positions in the environments. To distinguish the three application environments different colours were used to describe the scenes. Posters and information relating to the individual applications were visible in the individual applications. Figure 2 illustrates the appearance of the environments with the embodied agents.

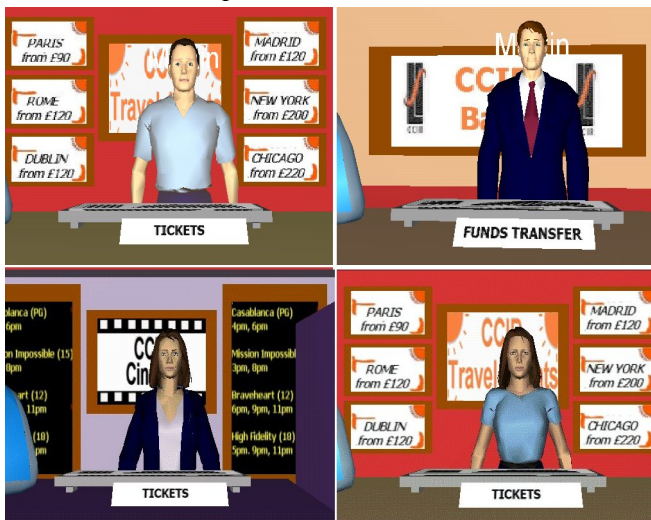


Figure 2: Images of Agents in Applications

(top: left: Casual Male (Travel); right: Formal Male (Bank);

bottom right: Casual Female (Travel); left: Formal Female (Cinema))

2.2 Creating the Embodied Agents

2.2.1 Agents' Appearance

MetaCreations Poser 4.0, a character animation software tool was used to create male and female animated agents. Once the agents were created in Poser 4.0 it was necessary to export the agent files to 3D Studio Max. Using optimisation tools in this software package the file was reduced in size and altered to allow it to be exported to VRML97. The VRML code was finally altered to fit into the H-Anim specification. The H-Anim spec specifies a standard way of representing humanoids in VRML97. "The standard allows humanoids created using authoring tools from one vendor to be animated using tools from another." [12].

2.2.2 Agents' Gesturing

Typically an H-Anim file contains a set of joint nodes that are arranged to form a hierarchy. Each joint node can

contain other joint nodes and may also contain a segment node, which describes the body part associated with that joint. Using this specification it was possible to obtain access to the joints of the agent and alter the joint angles to create gestures and mouth movements. Four gestures were created for the embodied agents: nodding, waving, shrugging and typing.

2.2.3 Agents' Voices

One male and one female voice recorded the necessary output prompts for the male and female agents respectively. When an entire output utterance was to be constructed, the dialogue manager called on the file containing all the necessary prompts. The relevant prompts were concatenated in a particular order to produce a plausible output sentence.

3. EXPERIMENT PROCEDURE

This experiment aimed to assess the functionality of 3D male and female embodied agents, as assistants in VRML retail application environments. Two types of male and female agents were assessed. The first was a smartly dressed formal agent, and the second a casual informally dressed agent. In order to evaluate the two agent types in the three VRML retail applications, the real-time experimental platform system, capable of face-to-face conversation between the user and the agent was used.

The dependent variables in the experiment were the responses to the individual items in (1) the application questionnaire, (2) the assistant questionnaire, and (3) the application comparisons. The independent variables were embodied agent gender, embodied agent appearance (formal, casual), and the VRML retail applications (cinema, travel, bank). These were treated as within-subject variables in a repeated measures design. The presentation of the agents to the participants was randomised within the applications and the presentation of the three applications was balanced amongst the participants. Four similar tasks created for each application were randomised amongst the agents. A group of 36 participants took part in the experiment, distributed evenly according to gender and age group (age 18-35, 36-49, 50+). Effects of between-subject variables of age and gender were investigated. One experimental supervisor was used for all participants.

The experimental procedure required participants first of all to read an information sheet regarding the application task. For instance, if the participant was going to see the cinema application first, they were told they would have to converse with an automated shop assistant to buy tickets to see a movie. In all cases the participants were asked to carefully observe the assistant and the service, which appeared on the PC screen in front of them. They were also told that they might be asked for security number

information, which was presented to them before the interaction began. After the conversation participants were asked to fill out a questionnaire relating to the assistant. The questionnaire items were 7-point Likert [6] attitude questionnaire statements presented as that shown in Table 1. Within the questionnaire, statements were balanced for polarity (equal number of positively and negatively worded stimulus statements).

I liked the appearance of the assistant.						
Strongly Agree	Agree	Slightly Agree	Neutral	Slightly Disagree	Disagree	Strongly Disagree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Table 1: Example of a Likert Questionnaire Item

When participants had seen all four agents in an application they filled out a short attitude questionnaire (again 7-point attitude statements) relating to the application. After the participants interacted with all the agents in all three applications they completed a questionnaire stating their preferences among the applications. The participants then took part in a closing interview designed to elicit further information about the agents, which also gave participants the opportunity to make suggestions for improvements to the system.

4. RESULTS

4.1 Attitude to Applications

4.1.1 Quantitative Analysis

The mean rating scores from the 10-point (low-high) application rating scale show a largely positive response to all three applications. A 3 x 1 repeated measures ANOVA taking experimental application as the independent variable showed no significant effects for certain applications. The cinema was rated the highest, followed by the travel agency and thirdly the bank. No interactions between applications and participant age or gender emerged.

	Mean Rating Score
Cinema	6.56
Travel Agency	6.46
Bank	6.12

Table 2: Mean Rating Score for Application (max 10)

The 7-point Likert questionnaire that was used to retrieve information about the participant's attitude toward the applications showed that the participants felt that the applications were not difficult to use (Item 2). No significant differences or interactions emerged. Participants also felt the services were good ideas (Item 3), with no significant difference between applications. The results showed that participants felt the applications were equally convenient (Item 4).

Questionnaire Item	Mean Score Cinema	Mean Score Travel	Mean Score Bank
1. I would use this service myself.	4.92	4.62	4.42
2. I felt the service was difficult to use.	5.05	5.19	5.05
3. I do not think this service is a good idea.	5.28	5.12	5.12
4. I think this service is convenient.	5.23	5.19	5.25

Table 3: Statements for Application Questionnaire

4.1.2 Application Comparisons

After participants had interacted with all four agents in the three applications they were asked to make a selection as to which application they preferred overall. Participants were also given the option to make a selection, which indicated they liked all three applications equally or didn't like any of the applications. This questionnaire gave strong indications that the cinema application was the preferred choice *in comparison* to the other applications.

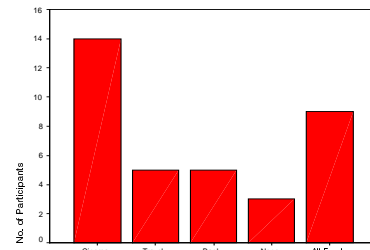


Figure 3: Application Preferences

40% of participants preferred the cinema application, 14% of participants preferred the travel agency and 14% preferred the banking application. 8% did not like any of the applications and 25% of the participant sample liked all applications equally. A chi-square test showed that the cinema application was significantly preferred to the other two applications, ($p < 0.05$).

4.1.3 Qualitative Comments about the Applications

The qualitative comments help explain the strong preference for the cinema application over the travel and banking applications. Of the 14 participants who stated a preference for the cinema application, 11 provided comments for their choice. In comparison to the other applications, the participants were mainly concerned with issues relating to trust and security of payments in the banking application. One participant commented that he "has not enough confidence in the technology yet". The issue of confidence was raised again when comments were made about the cinema in comparison to the travel agency. One participant stated that he would "rather miss a movie, than a flight" in the event of an error being made

by the system. Another participant described the information for the banking and travel interactions as being more “critical”, with users becoming more anxious if something goes wrong. Overall the cinema application was preferred because it “seemed easier to use” and the straightforward transaction seemed “simpler”. One participant commented that the system was not that different from the telephone booking services already in place, but this experience was an improvement because of the feeling of “dealing with someone face to face”.

4.2 Attitude to Assistants

A series of 2 x 2 x 3 repeated measure ANOVAs taking agent gender, agent type and application as the within-subject independent variables and participant age and participant gender as the between-subject independent variables were conducted to analyse participants’ attitudes to the questionnaire items relating to the embodied agents as assistants. The questionnaire addressed three key issues: agents’ voice, personality and appearance.

4.2.1 Attitude to Voice & Conversation

The voices of both the male and female agents were rated similarly with no significant differences evident when participants were asked if they liked the voices. There were no significant differences found for agent type or agent gender in any application. Participants were also asked if they liked speaking to the assistant, and the results showed that participants did like speaking to the assistants in all the applications, with no significant differences for agent type or gender. Overall, participants felt the agents understood them during the course of the interaction and no significant effect, for application, agent gender and agent type were found, suggesting that the recogniser worked equally well in all three applications.

Participants were asked if they thought the assistants’ voices were annoying. A significant effect was found for the agent gender ($F = 5.52, df = 1.0, p < 0.05$). Post-hoc t-tests showed that this was caused by a significant difference between the male and female formal agents, ($p < 0.05$), which indicated that participants felt the voice of the formal female agent was less annoying than the voice of the formal male agent. Extending from this highly significant effects were found for agent gender when participants were asked if the assistant spoke naturally, ($F = 19.6, df = 1.0, p < 0.01$). Overall, participants perceived the female voice to be significantly more natural than the male voice. It became apparent that the concatenated nature of the speech output was less natural for the male voice than for the female voice.

4.2.2 Attitude to Personality

All the agents were perceived as being equally friendly. None of the agents were perceived as being bossy and all

were perceived as being equally competent. In addition all four agents were perceived as being sociable, cheerful, and agreeable. Participants were asked if the assistants were trustworthy. Although just approaching significance, the mean results did show that the assistants in the bank scored less than the assistants in the other applications.

	Mean Score (max 7)
Cinema	5.15
Travel Agency	5.24
Bank	4.94

Table 4: Mean Scores for Trustworthiness

4.2.3 Attitude to Appearance

Participants were asked if they liked the appearance of the assistants. Results showed (Figure 4) that they significantly preferred the formal agents to the casual agents in the banking application, ($p < 0.01$).

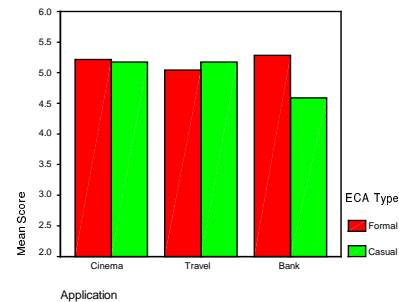


Figure 4: Attitude to Appearance

Participants were asked if the assistants were dressed appropriately for the applications. Significant results, illustrated in Figure 5, showed that participants felt the agents in the cinema application should be dressed informally and the agents should be dressed formally in the banking application, ($F = 15.65, df = 2.0, p < 0.01$).

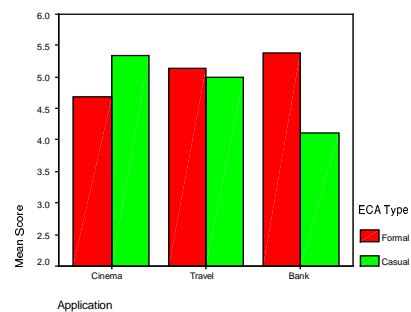


Figure 5: Attitude to Appropriateness of Assistants Dress

5. DISCUSSION

Five predictions were made before the experiment began, four of which were supported. The first stated that the participants would respond positively to the embodied agents. The results support this claim and 3D embodied conversational agents have a role to play as assistants in VRML retail application environments. This can be said given the positive responses recorded in the quantitative

questionnaires and the qualitative comments. It is important to establish that ECA's are welcomed in retail domains especially with increased numbers of websites considering anthropomorphised interfaces.

Despite the fact that the participants enjoyed speaking to the agents in all three applications, supporting the second prediction, the cinema application was more popular in comparison to the other applications. Participants felt it was more entertaining than the travel agency or banking applications. Although ECA's were welcomed in the three retail applications in this experiment, it is important to consider carefully the seriousness and entertaining nature of the application task and be aware that ECA's may be more effective in less serious applications. Nevertheless, the responses to the use of ECA's in these more serious applications may be improved if users' confidence in the system can be increased and the trustworthiness in the agent can be firmly established. The results supported a further claim, that casually dressed agents are more suitable in virtual cinemas, and formally dressed agents are more suitable in virtual banking applications.

The fourth claim addressed the possible emergence of any gender differences within or between applications and this claim was largely supported and no gender differences were found. In the experiment, participants liked both the male and female voices for both types of agent, but did feel the male voice was more annoying and less natural than the female. The interviews indicated that this difference was caused due to the concatenated nature of the output utterances from the agents, in particular the male agents. As mentioned a male and female person was selected to record the output prompts to be used as the voice output for the agents. The relevant prompts were played in a particular order to produce a plausible output sentence. It was significantly felt that the concatenation of the male utterances was not as natural as the female voice output. In the development of complete applications and in event of recorded speech being used, great care must be taken with concatenated recordings to ensure that the transitions between the prompts in an utterance sound natural, with appropriate intonation.

All four agents were perceived as having similar personalities, and were thought to be polite, friendly, competent, cheerful, sociable and agreeable; all traits important for assistants in retail spaces. However extending from this, the trustworthiness of the agents differed between the applications. This did not then support the claim that all the agents would be perceived as having similar personalities between the applications. The qualitative results showed that participants were less likely to trust the agents to complete tasks correctly particularly in the banking application.

Research is suggesting that the development of ECA's in all domains (entertainment, educational, retail etc.) will be dictated not only by technological advances but more so by advances in the understanding and creation of the social interaction between the agent and user. With respect to retail applications the establishment of trust between the user and the agent is of great concern. Further research, using the experimental platform described in this paper will be used to manipulate the 3D interface, aiming to increase user confidence in the system and the trustworthiness of the agent.

6. REFERENCES

- [1] E. André, and T. Rist. 'Personalising the User Interface: Projects on Life-like Characters at DFKI'. In Proc. *3rd Workshop on Conversational Characters*, pp. 167-170, October 1998.
- [2] G. Ball and J. Breese. 'Emotion and Personality in a Conversation Agent'. In *Embodied Conversational Agents*, ed. J. Cassell, J. Sullivan, S. Prevost, E. Churchill. Cambridge, Mass., London, MIT Press, 2000. ISBN 0-262-03278-3.
- [3] T. Bickmore and J. Cassell. 'How about this weather? Social Dialogue with Embodied Conversational Agents'. In Proc. *Socially Intelligent Agents: The Human in the Loop*. AAAI Fall Symposium, pp. 4-9, 2000. ISBN 1-57735-127-4.
- [4] J. Cassell, J. Sullivan, S. Prevost and E. Churchill. *Embodied Conversational Agents*. Cambridge, Mass., London, MIT Press, 2000. ISBN 0-262-03278-3.
- [5] K. Dautenhahn ed. 'Socially Intelligent Agents: The Human in the Loop'. To appear in *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*.
- [6] R. Likert. 'A Technique for the Measurement of Attitudes'. In *Archives of Psychology* 140, p.55, 1932.
- [7] H. McBreen and M. Jack. 'Empirical Evaluation of Animated Agents In a Multi-Modal Retail Application'. In Proc. *AAAI Fall Symposium: Socially Intelligent Agents - The Human in the Loop*, pp. 122-126, November 2000. ISBN 1-57735-127-4.
- [8] H. McBreen, P. Shade, M. Jack and P. Wyard. 'Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications'. In Proc. *4th International Conference on Autonomous Agents*, pp. 39-45, ACM Press, June 2000. ISBN 1-581-13230-1.
- [9] C. Nass, K. Isbister and E. Lee. 'Truth is Beauty: Researching Embodied Conversational Agents'. In *Embodied Conversational Agents*, ed. J. Cassell, J. Sullivan, S. Prevost, E. Churchill. Cambridge, Mass., London, MIT Press, 2000. ISBN 0-262-03278-3.
- [10] B. Reeves and C. Nass. *The Media Equation*. Stanford University, California. CSLI Publications, 1996. ISBN 1-575-86053-8.
- [11] <http://www.deepmatrix.com>
- [12] <http://www.h-anim.org>