Evaluating Humanoid Synthetic Agents in E-Retail Applications

Helen M. McBreen and Mervyn A. Jack

***Abstract*—This paper presents three experiments designed to empirically evaluate humanoid synthetic agents in electronic retail applications. Firstly, human-like agents were evaluated in a single e-retail application, a home furnishings service. The second experiment explored application dependency effects by evaluating the same human-like agents in a different e-retail application, a personalized CD service. The third experiment evaluated the effectiveness of a range of humanoid cartoon-like agents. Participants eavesdropped on spoken dialogues between a "customer" and each of the agents, which played the role of conversational sales assistants. Results showed participants expected a high level of realistic human-like verbal and nonverbal communicative behavior from the human-like agents. Overall ratings of the agents showed no significant application dependency: Two different groups of participants rated the human-like agents in similar ways in a different application. Further results showed participants have a preference for three–dimensional (3-D) rather than two–dimensional (2-D) cartoon-like agents and have a desire to interact with fully embodied agents.**

I. Introduction

THE PRESENCE of synthetic humanoid agents in interfaces can result in an efficient, engaging, and social collaboration between humans and machines, as demonstrated by PPP-Persona [1], Rea [5], Steve [19], and Gandalf [24]. However, it cannot be assumed that the use of animated agents in interfaces guarantees successful human–computer interaction. There is a need for the establishment of objective and subjective measures of usability for embodied animated agents and the development of methodologies to support such evaluations. As Cassell [3] observes, empirical investigations of any kind of embodied interfaces are rare and the results so far have been equivocal. Dehn and van Mulken [6] reviewed several empirical studies of the usefulness of animated agents in interfaces. Although the evaluations used different methodologies, thereby, making it difficult to make comparisons and draw general conclusions, Dehn and van Mulken identified three categories of data important for the evaluation of interfaces using animated agents:

- the user's subjective experience;
- the user's behavior while interacting with the system;
- the outcome of the interaction as indicated by performance data.

Manuscript received December 18, 2000; revised April 10, 2001. This work was supported by British Telecommunications (BT) under its Strategic University Research Initiative.

The authors are with the Centre for Communication Interface Research (CCIR), University of Edinburgh, Edinburgh, U.K. (e-mail: helen@ccir.ed.ac.uk).

Publisher Item Identifier S 1083-4427(01)07725-6.

While the importance of the interactive and contextual aspects of the second and third categories must be recognized, there remain many important, outstanding issues with respect to the subjective experience of interfaces using embodied animated agents. Dehn and van Mulken drew attention to some important dimensions of users subjective experience that are commonly measured:

1) perceived intelligence;
2) believability;
3) likeability;
4) activity of the system;
5) degree of entertainment;
6) usefulness.

The research reported in the present paper extends this list of dimensions through three experiments using a wide range of synthetic humanoid agents.

In each experiment the interfaces were anthropomorphized by having the humanoid agents act as assistants in a retail context. In the first experiment, five human-like agents were selected to represent important points in the range of possible technologies that can be used to create conversational agents, vis-a-vis

1) video;
2) a three-dimensional (3-D) talking head;
3) a photorealistic image with quasidynamic facial expressions;
4) a still image;
5) a disembodied voice.

All the agent types (excluding the disembodied voice) were created from human photorealistic images. In the second experiment, the same five human-like agents were reevaluated in a contrasting retail environment, to investigate application dependency effects. The third experiment investigated humanoid cartoon-like agent types using the retail environment from the first experiment.

Video was included in the first experiment in order to allow the investigation of user expectations of human facial expressions in comparison to the visually less sophisticated expressions of the other technologies. Although the use of video may be impractical in interactive online applications, the MikeTalk project [7] has demonstrated that the illusion of video can be created successfully without recourse to the use of prerecorded segments. This particular technology may be a suitable substitute in the event of video being the preferred human-like agent type.

Parke and Waters [17], [27] pioneered work in the area of 3-D talking heads: "one of the more intriguing possibilities (for future research) is the construction of interactive face agents

capable of assisting and conversing with the user." Massaro [14] has been instrumental in investigating user perception of speech output from talking heads, using a sophisticated computerized 3-D talking head known as Baldi. This talking head can produce synthetic auditory and visible speech and has a highly developed simulation of a vocal tract that can be shown to the user during an interaction.

Thalmann [23] is focusing on facial communication in virtual environments, in particular facial cloning, real-time animation, and face feature tracking. The work draws heavily on the model of human facial expression developed by Ekman [20] to represent life-like, nonverbal communicative behavior in the face. Thalmann's work concentrates on speech animation and synchronization by extracting phonemes: "in virtual environments, realism not only includes the believable appearance and simulation of virtual worlds, but also implies the natural representation of the virtual humans and participants."

The first and second experiments also posed the question whether participants would prefer to be presented with a still image of an agent or just to hear a disembodied voice. Takeuchi and Nagao [21] showed that people do try to interpret facial displays, and conversing with a system that has facial displays is more successful than conversing with a system that lacks such displays. However, they also showed that the use of facial displays could interfere with the user's concentration. They argue that this is not necessarily a negative effect of anthropomorphized interfaces. On the contrary, they suggest that it shows that the user is appreciative of the human image that he or she tries to interpret. The inclusion of a quasidynamic image and a still image in the experiment reported here makes it possible to assess the effect of the image type on participants' attitudes.

The third experiment reported evaluated a selection of cartoon-like humanoid agents. Included in the cast of animated humanoids were two–dimensional (2-D) and 3-D heads and 2-D and 3-D embodied agents. Badler [2] has written about his vision of the future of virtual human animation in which new tools and techniques, such as motion capture, will provide animators with a greater scope for developing the necessary animations for the future. In addition, studying human movements can provide a better understanding of how animated agents should move. "Virtual humans should be alive, not just movable. They should have nice, virtual bodies, they should walk and not slide" [22].

The results of the experiments in this paper showed that retail applications offer an important and new environment for humanoid synthetic agents. Technological advances in speech recognition, speech synthesis and, more importantly, dialogue management have made it possible to use synthetic humanoids as interactive conversational agents to assist users in completing a task or retrieving information about a product or service. In a study of personified interfaces [26], it was suggested that the goal of human–computer interaction research should not necessarily be to give a computer a human face, but rather to determine when a face in an interface is appropriate. Not only does the research discussed in this paper evaluate various synthetic agent types, it also attempts to discover if there is scope to include such agents in e-retail applications.

Studies by Koda [10], Lester [12], and Walker *et al.* [26] showed that agents with strong visual presence and facial expression could be more engaging and motivating for the user. Expanding from this result, the empirical evaluation reported here shows that certain human-like agent types are liked more than others and the point at which the human-like faces are disliked is sensitive. Walker *et al.* [26] and Koda [10] also demonstrated that task performance was not negatively affected by the use of a face in the interface. The results from the experiments in this paper show that certain agent types can distract the user, not necessarily leading to poorer task performance, in agreement with the results of Walker *et al.* and Koda, but the distraction can reduce the user's attitude to the agent as an assistant and discourages further interactions with the agent.

The advantages of the presence of synthetic human-like or cartoon-like agents in interfaces are not fully realized. The long-term aim of the work reported here is to discover what, if any, besides being an alternative mode of communication, are the advantages of such synthetic humanoid agents in e-retail environments.

II. Synthetic Agent Implementation

A. Human-Like Agents (H)

The first two experiments compared user attitudes to a cast of five human-like synthetic agents types who appeared in two e-retail environments. Two contrasting applications were used to investigate application dependency effects. Male and female versions of five contrasting agents were created (see Fig 1).

1) H1—Video (H1): Using an auto-cue to read a script, a male and female person played the role of assistants and were recorded on video. The 2-min script was a dialogue between a "customer" and the assistant. The soundtracks from these .AVI script files were extracted and used as the speech output for all other human-like agents, e.g., female assistant voice soundtrack was used for the other four female human-like agents.

2) H2—3-D Talking Head (H2): Still images of the persons used to create H1 were mapped onto a wire-frame model of a human head, creating a 3-D talking head. Phoneme matching lip-synchronization was used to match the soundtrack taken from the video files.

3) H3—Photo with Facial Expressions (H3): Photorealistic still image of the male and female videos were taken. Using Adobe PhotoShop 4.0 varying frames of these images were produced. These frames were imported into Adobe Premier 5.1, edited, and run in a sequence to create an animation of the image with graphic lip movement and facial expression such as eyebrow raising and blinking.

4) H4—Still Image (H4): This was a static image of the male and female assistants who appeared in the video recordings.

5) H5—Disembodied Voice (H5): This was created using only the audio soundtrack files extracted from the recordings for H1. Including disembodied voices raised interesting issues about the need for anthropomorphic characters to visually appear in interactive e-retail interfaces.

It was predicted that the conversational applications would be received positively. This prediction was made based on the literature mentioned previously where users can benefit from the presence of a conversational agent in the interface. As the same male voice was used for all male agents, it was predicted

that the attitude to the voices of all the male agents would be the same, and as the same female voice was used for all female agents, the attitude to the voices of all the female agents would be the same. In addition to this as the male and female agents of the same agent type were the same by design (same verbal and nonverbal feedback), it was predicted that agents of the same type would be rated similarly.

B. Cartoon-Like Agents (C)

An incremental scale of humanoid cartoon-like agents was created primarily using 3-D Studio Max. Different voices to those used for the human-like agent creation (one male, one female) were used to create the voice soundtracks for the cartoon-like agents. This change was due to negative comments about the voices used for the human-like agents, which were deemed less suitable for extension into the soundtrack voices for the cartoon-like agents.

1) C1—Disembodied Voice (C1): This agent was included as a control between the two casts of agents (human-like and cartoon-like).

2) C2—2-D Head (C2): Two-dimensional male and female animated heads with lip-synchronization were created. See Fig. 2.

3) C3—3-D Head (C3): Three-dimensional male and female versions of the 2-D heads described above were created. The agents could nod at appropriate times during the conversation and turn to look at changes in the interfaces.

4) C4—2-D Embodiment (C4): Two-dimensional male and female full-bodied versions of the 2-D male and female heads were created. See Fig. 2.

5) C5—3-D Embodiment (C5): Three-dimensional male and female full-bodied versions of the 3-D male and female heads were created. As with the 3-D heads, nodding and eyebrow raising was included during the conversation. Gesturing also added to the realism of the agents and the agents could also turn to look at changes in the interface.

6) C6—3-D Embodiment in 3-D Environment (C6): The 3-D full-bodied cartoon-like agents actually appeared inside the 3-D application environment instead of to the left of the main application window. The male and female versions were identical to C5 and differed only in their position on the interface. Here it will be possible to investigate user attitudes toward a 3-D embodied agent in a 3-D virtual environment.

As with the first two experiments, it was predicted that the conversational application would be received positively. In addition, as the same male voice was used for all male agents, it was predicted that the attitude to the voices of all the male agents would be the same, and as the same female voice was used for all female agents, the attitude to the voices of all the female agents would be the same. In addition to this, as the male and female agents of the same agent type were the same by design (same verbal and nonverbal feedback), it was predicted that agents of the same type would be rated similarly.

C. Verbal and Nonverbal Behavior

1) Facial Movements: Facial movement and speech intonation was attributed to all the visible agents in all three experiments. Randomized blinking was introduced for all visible agents (except H4). Although facial expressions were recorded naturally when creating the videos (H1), they were introduced in a controlled manner for all other visible agents in all three experiments. When asking questions the agent raised the voice slightly, stressed the main word of the sentence, and raised its eyebrows (all visible agents, except H4). During an affirmation the agent nodded, raised its eyebrows, and blinked at the end of the sentence (nodding only for 3-D agents). Finally, mutual gaze with the user was maintained when the user was speaking and at the end of the agent's turn (all visible agents).

Fig. 1. Photorealistic images of male and female human-like agents.

Fig. 2. Two-dimensional heads and embodied cartoon-like agents.

2) Gesturing: Gesturing was introduced in a controlled manner for the 2-D and 3-D embodied agents. According to Cassell [3] there are three categories of gesturing:

1) emblematic;
2) propositional;
3) spontaneous.

Spontaneous gesturing constitutes the majority of gestures and was introduced into the embodied cartoon-like agents evaluated in the third experiment. In particular, the nature of the dialogue supported the inclusion of deictic and beat gesturing. The 2-D embodied agents only had deictic gesturing and pointed to the selection area at the top of the interface, thereby directing the users gaze to that point. The three-dimensionality of C5 and C6 afforded the use of beat gesturing. The nonverbal behavior for agents of the same type was the same.

III. Application Implementation

Two contrasting application interfaces were chosen. For consistency both interfaces contained the same three elements:

1) a window to display the synthetic agent (the same size in both interfaces);
2) a main application window;

3) a "product selection area" positioned above the application window.

To minimize visual effects, the visual elements of both application interfaces were designed to be consistent.

1) Application 1—Home Furnishings Service: The graphical user interface (GUI) was created in the style of MUeSLI [28]. The main window was a 3-D view of a living room complete, which was dynamically updated following changes requested by the customer (Fig. 3). Above this was a "selection area" containing fabric and wallpaper samples that could be selected in order to decorate the room. The synthetic agent was displayed in a window on the left-hand side of the interface. The dialogue illustrated a "customer" conversing with the agent to decorate the room. An identical dialogue was used for all agents, an extract of which can be examined in Table I.

2) Application 2—Personalized CD Service: In the CD service, the main application window was a 3-D view of the customer's personalized CD, updated after each addition. The selection area contained a row of tracks from one artist, which could be selected for inclusion on the personalized CD (Fig. 4). The synthetic agent was displayed in a separate window. The dialogue illustrated the "customer" conversing with the agent in order to select tracks of their choice (Table II). The customer, upon request, could listen to musical excerpts of the tracks that were available.

In all cases (except the disembodied voice agent type), standalone movies were created for the male and female versions of each of the agent types. The desktop video editor Adobe Premier 5.1 was used to manipulate a series of .JPG files that contained the images of the interface, for instance, the sequence of changes in the living room. These sequences of image changes were imported into Apple QuickTime and a movie file was created. The multimedia files, the QuickTime movie file of the interface changes, and the movie file of the synthetic agent were exported into Macromedia Director 6.5, to create a projector file, which timed exactly the audio and visual changes of the interface to the soundtrack.

IV. Experimental Procedure

The experimental procedure was essentially identical for all three experiments, although some necessary alterations were made to the design of the third experiment. Such improvements were based on feedback gathered in the first two experiments. Separate groups of 32 participants took part in the first and second experiments, distributed according to gender and age. Three age groups were used: 18–35, 36–49, and 50+. In the third experiment, 36 participants took part to accommodate for the additional agent type included in the cast. Participants first read a brief explanation of the purpose of the experiment after which they were also primed verbally by the experiment supervisor. They then viewed 2-minute videos (created using Macromedia Director 6.5 and presented in randomized order on a Pentium II PC), showing the dialogue between the "customer" and a synthetic agent.[1] A seven-point agree–disagree Likert [13] format questionnaire was used to retrieve quantitative data about users' attitudes to the agents. An example statement is shown in Table III.

In each experiment, the attributes of the agents were assessed by having participants "eavesdrop" on the dialogues between a customer (represented by a female disembodied voice) and each of the agents. The passive methodology used to assess the agents was extremely practical, as it avoided the complex technological

[1]For the first and second experiments, identical application dialogues were assigned to all agents. It transpired that participants felt the repetitive identical dialogues were monotonous, and therefore, in the third experiment, similar but not identical dialogues were assigned to each of the agents.

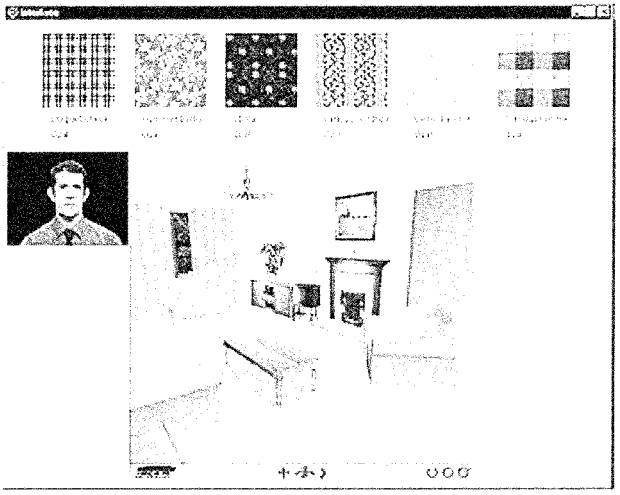

Fig. 3. Interface for application 1: Home furnishings.

TABLE I
DIALOGUE EXCERPT FROM APPLICATION 1

Customer	***I'd like to plan a make over for my sitting room***
Assistant	Good, what would you like to see first?
Customer	***Can you show me some green fabrics for the sofa?***
Assistant	Certainly, here's a selection. [swatches appear]

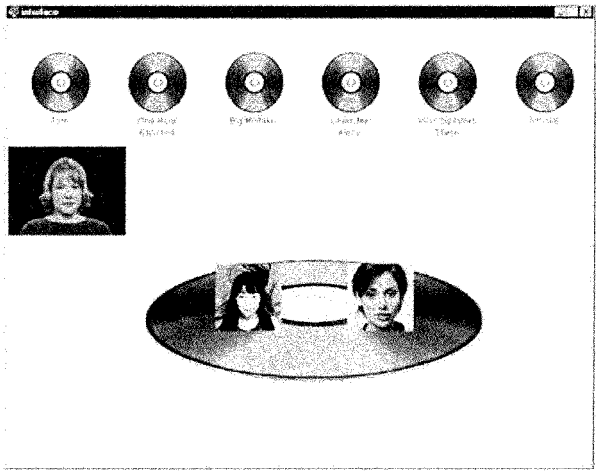

Fig. 4. Interface for application 2: CD service.

TABLE II
DIALOGUE EXCERPT FROM APPLICATION 2

Customer	***I want to create my own compilation CD.***
Assistant	Cool, what would you like for Track One?
Customer	***I want to start with something by Bjork? [selection of Bjork tracks appears]***
Assistant	Would you like to hear a track?

TABLE III
EXAMPLE OF LIKERT QUESTIONNAIRE ITEM

I liked the appearance of the assistant

strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
☐	☐	☐	☐	☐	☐	☐

issues involved in creating a fully functional interactive application with a range of agents, but it still allowed a full evaluation of each of them. Ideally, an interactive application would have provided more informative results; however, to evaluate such a substantial cast of agents this compromise was made.

At the end of the session participants took part in a short interview to provide them with the opportunity to expand on any underlying issues. The qualitative information focused on the participant's views of the application and the agent's functionality and behavior in the application. Following the interview, participants rated each sales assistant on a scale of 1 to 10 (ten being the best).

To summarize, experiment 1 evaluated the ten human-like synthetic agents in application 1; the home furnishings service. Experiment 2 evaluated the same ten human-like synthetic agents in application 2; the personalized CD service. Experiment 3 then progressed to assess a cast of cartoon-like humanoid agents in application 1; the home furnishings service. The results are presented below.

V. RESULTS

A. Human-Like Agents (Experiments 1 and 2)

The results from Experiments 1 and 2 have been grouped together to allow a comprehensive assessment and comparison of the human-like agents in the contrasting applications.

1) Overall Ratings: In order to obtain an overall rating of the ten synthetic human-like agents for both applications, the results of a ten-point rating scale completed during the post-experiment interviews in both experiments were analyzed. A $2\times5\times2$ repeated measures ANOVA was carried out taking agent gender, agent type, and application as the independent variables and the mean rating scores as the dependent variable. There was a highly significant main effect for agent type ($F = 61.48$, $df = 4.0$, $p < 0.01$) and a marginally significant effect for agent gender ($F = 4.26$, $df = 1.0$, $p < 0.05$). There was no main effect for application and no significant interactions. The effects of the between subject variables of participant age group and participant gender were also not significant. Table IV shows the mean ratings for each agent type (pooled for application and gender) with the results of pair-wise comparisons.

2) Attitude to Applications: Using the seven-point Likert questionnaire statements to express their opinion, participants were asked if they felt the service was a good idea. A $2 \times 5 \times 2$ ANOVA taking agent gender, agent type, and application as the independent variables showed highly significant effects for agent type, ($F = 4.11$, $df = 4.0$, $p < 0.01$). The participant attitudes did follow the general trend of the overall ratings, with videos (H1) being more popular than H2 ($p < 0.05$), H3 ($p < 0.01$), and H4 ($p < 0.05$). There was a marginally significant effect for agent gender ($F = 4.52$, $df = 1.0$, $p < 0.05$), with male agents given more positive ratings than female agents (mean female $= 5.30$, mean male $= 5.47$). No significant effect for application was found. There was a significant interaction between agent type and application ($F = 4.22$, $df = 4.0$, $p < 0.01$). The interaction between agent type and application suggested there were significantly different attitudes to the agents in both applications. In fact there was a much lower mean score for the still image (H4) in the CD Service, than the home furnishings service (Table V).

TABLE IV
MEAN RATINGS SCORES AND PAIR-WISE COMPARISONS FOR HUMAN-LIKE AGENT TYPE

Human-Like Agent Type	Mean Rating Score (max 10)	Type rated better than (all $p < 0.01$)
H1 (Video)	*7.25*	H2, H3, H4, H5
H5 (Voice only)	*5.85*	H2, H3, H4
H4 (Still image)	*4.70*	H2
H3 (Image w/ facial exp.)	*4.27*	H2
H2 (3D talking head)	*3.05*	-

TABLE V
MEAN SCORES FOR APPLICATION AND HUMAN-LIKE AGENT FOR QUESTIONNAIRE ITEM: "I THINK THIS SERVICE IS A GOOD IDEA"

Agent Type	H1	H2	H3	H4	H5
Mean Score (App 1)	*5.53*	*5.32*	*5.53*	*5.57*	*5.57*
Mean Score (App 2)	*5.73*	*5.13*	*5.15*	*5.03*	*5.30*

Analysis of the results showed that participants did not consider that they would find either service difficult to use if they were to use the applications interactively. A $2 \times 5 \times 2$ ANOVA taking agent gender, agent type, and application as the independent variables showed no significant effects. When asked if they would like to use the service themselves, participants' responses were in general positive, with no significant effect for agent type or application, and no significant interactions. The $2 \times 5 \times 2$ ANOVA showed highly significant effects for agent gender, ($F = 4.108$, $df = 1.0$, $p < 0.01$). Participants would have liked to use the service if the agent were male (mean female $= 4.92$, mean male $= 5.21$: max $= 7$). This result is discussed in more detail below, but it appears to be an effect of the negative response to the distinctive accent of the female voice.

3) Attitude to Voices: A $2 \times 5 \times 2$ ANOVA taking agent gender, agent type, and application as the independent variables showed (Table VI) highly significant effects for agent type ($F = 3.79$, $df = 4.0$, $p < 0.01$) when participants were asked about the clarity of the agent's voice.

These results suggest interesting crossmodal audiovisual perceptual effects. Even though the voices for the male and female agent types were identical for each gender, participants had varying attitudes toward the clarity of the voices for the different agent types. The videos (H1) and the disembodied voices (H5) were rated similarly with respect to clarity. Pair-wise comparisons show that the voices of H5 were significantly clearer than H2, H3, and H4, all at $p < 0.01$. Also, the voices of H1 were significantly clearer than H2 and H4, both at $p < 0.01$.

A $2 \times 5 \times 2$ ANOVA analyzed responses as to whether participants liked the voice. The results showed (Table VII) effects for agent type, ($F = 3.77$, $df = 4.0$, $p < 0.01$).

There was also a significant effect for agent gender ($F = 7.26$, $df = 1.0$, $p < 0.01$). No effect for application was found. The female voice was not liked as much as the male (mean female $= 4.53$, mean male $= 4.89$: max $= 7$). Participants explained that the female voice had a more distinctive accent than the male voice, which did not appeal to many participants. There were no significant interactions.

4) Naturalness of Conversation: First, a $2 \times 5 \times 2$ ANOVA taking agent gender, agent type, and application as the independent variables showed highly significant effects for agent type, ($F = 22.55$, $df = 4.0$, $p < 0.01$) with respect to the naturalness of the conversation. There were also highly significant effects for agent gender ($F = 8.05$, $df = 1.0$, $p < 0.01$) and for application ($F = 37.65$, $df = 1.0$, $p < 0.01$). The results show (Table VIII) that the conversation with the videos (H1) was felt to be most natural and was significantly more natural than H2, H3, H4, and H5, all at $p < 0.01$. The conversation with the male agent was preferred to the female. The result can be explained by the negative attitude to the female voice (mean female $= 2.82$, mean male $= 3.14$: max $= 7$). The figures show that the conversation in both applications were below neutral; it was more natural in the home furnishings service than in the CD service (mean home furnishings $= 3.82$, mean CD service $= 2.16$: max $= 7$). This could be due to participants not being familiar with the musical tracks that were chosen, as mentioned by some participants in the interviews.

5) Attitude to Friendliness and Competence: The questionnaire enquired if there were any differences in attitude to the friendliness and competence of each of the human-like agent types. A $2 \times 5 \times 2$ ANOVA showed no significant effects for agent gender, agent type, or application. No interactions were found. Participants felt that all agents were equally competent (mean for agent type $= 4.97$).

However, a $2 \times 5 \times 2$ ANOVA showed (Table IX) significant differences with respect to the friendliness of the human-like agents ($F = 3.87$, $df = 4.0$, $p < 0.01$).

There was also a significant result for agent gender ($F = 11.2$, $df = 1.0$, $p < 0.01$). The male agents were significantly friendlier than the female (mean female $= 4.20$, mean male $= 4.74$: max $= 7$). A significant result for application ($F = 11.72$, $df = 1.0$, $p < 0.01$) showed that the agents in the CD service were friendlier (mean home furnishings $= 5.39$, mean male $= 5.45$: max $= 7$). These results can be explained by an interaction between gender and application ($F = 12.36$, $df = 1.0$, $p < 0.01$), showing that female agents in the CD service were significantly friendlier than those in the home furnishings service (Table X).

6) Attitude to Helpfulness: Participants were asked if they thought that seeing the assistants was helpful. Significant results for agent type, ($F = 15.86$, $df = 3.0$, $p < 0.01$) and an interaction between agent type and application emerged, ($F = 5.84$, $df = 3.0$, $p < 0.01$). In the home furnishings service, seeing the videos (H1) was significantly more helpful than H2, $p < 0.01$. In the second application seeing the videos was significantly more helpful than H2, H3, and H4, at all $p < 0.01$ (Table XI).

TABLE VI
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE ASSISTANT'S VOICE WAS NOT CLEAR ENOUGH"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.80*	*4.29*	*4.52*	*4.39*	*4.86*

TABLE VII
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "I LIKED THE ASSISTANT'S VOICE"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.91*	*4.59*	*4.69*	*4.68*	*4.82*

TABLE VIII
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "I FELT THE CONVERSATION WAS NATURAL"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.05*	*2.64*	*2.71*	*2.71*	*2.81*

TABLE IX
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "I FELT THE ASSISTANT WAS FRIENDLY"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.72*	*4.23*	*4.40*	*4.35*	*4.63*

TABLE X
MEAN SCORES FOR HUMAN-LIKE AGENT GENDER AND APPLICATION FOR QUESTIONNAIRE ITEM: "I FELT THE ASSISTANT WAS FRIENDLY"

Agent Gender	Female	Male
Mean Score (App 1)	*5.37*	*5.40*
Mean Score (App 2)	*5.50*	*5.42*

TABLE XI
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE AND APPLICATION FOR QUESTIONNAIRE ITEM: "I THOUGHT SEEING THE ASSISTANT WAS HELPFUL"

Agent Type	H1	H2	H3	H4	H5
Mean Score (App 1)	*4.66*	*3.78*	*4.47*	*4.35*	*NA*
Mean Score (App 2)	*4.89*	*2.84*	*3.28*	*3.59*	*NA*

TABLE XII
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE APPEARANCE OF THE ASSISTANT WAS UNSUITABLE FOR THE APPLICATION"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.50*	*3.82*	*4.18*	*4.07*	*NA*

7) Attitude to Appearance: When asked if the appearance of the assistant was suitable for the application a $2 \times 4 \times 2$ ANOVA showed (Table XII) differences for agent type ($F = 6.4$, $df = 3.0$, $p < 0.01$). The videos (H1) were significantly different to H2 ($p < 0.01$), H3, and H4 ($p < 0.05$). There was

TABLE XIII
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE AND APP. FOR QUESTIONNAIRE ITEM: "I LIKED THE APPEARANCE OF THE ASSISTANT"

Agent Type	H1	H2	H3	H4	H5
Mean Score (App 1)	*4.87*	*3.58*	*4.23*	*4.26*	*NA*
Mean Score (App 2)	*4.98*	*4.64*	*4.75*	*4.75*	*NA*

TABLE XIV
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "I THOUGHT THE ASSISTANT LOOKED NATURAL"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*4.93*	*4.08*	*4.49*	*4.51*	*NA*

also a significant result for agent gender ($F = 5.9$, $df = 1.0$, $p < 0.05$). The appearance of the male agents was more suitable (mean female $= 4.40$, mean male $= 4.62$, max $= 7$).

A $2\times4\times2$ ANOVA was used to analyze the results when participants were asked if they liked the appearance of the agents. The results showed (Table XIII) highly significant effects for agent type ($F = 12.8$, $df = 3.0$, $p < 0.01$) with respect to the appearance of the human-like agents. There was a significant interaction for agent type and application ($F = 4.59$, $df = 3.0$, $p < 0.01$). Following the overall trend, pair-wise comparisons showed that the videos (H1) had the most popular appearance, and $p < 0.01$.

There was also a highly significant effect for agent gender ($F = 11.98$, $df = 1.0$, $p < 0.01$) with the appearance of the male agents being preferred (mean female $= 4.35$, mean male $= 4.65$). The negative response to the female voice could possibly have impacted on the attitudes toward their appearance.

Participants were also asked if they thought the assistant looked natural. Significant results (Table XIV) emerged with respect to agent type ($F = 7.36$, $df = 3.0$, $p < 0.01$), with the videos thought to be more natural than the other agent types.

There was also an interaction between agent type and application ($F = 16.7$, $df = 3.0$, $p < 0.01$). Agents in the first application followed the overall trend (Table XV), with the video looking more natural. The CD service results did not follow this trend and a reason for this could have been still images of the artist and the agent (H4) appearing in the interface simultaneously, causing confusion. There could have been ambiguity, that is if the participants did think the agent was unnatural compared to the other agents, but not unnatural compared to the other visual stimuli in the interface.

8) Attitude to Facial Expressions: In the questionnaires participants were asked it they felt the speech matched the lip movement. A $2\times3\times2$ ANOVA taking agent gender, agent type, and application as the independent variables showed (Table XVI) highly significant effects for agent type ($F = 13.35$, $df = 2.0$, $p < 0.01$), indicating participants were aware the agents had different lip movement. There was no effect for agent gender. The means for all three agent types were below neutral, suggesting that participants did not think this lip synchronization was impressive, even for the video H1.

TABLE XV
MEAN SCORES FOR HUMAN-LIKE AGENT GENDER AND APPLICATION FOR QUESTIONNAIRE ITEM: "I THOUGHT THE ASSISTANT LOOKED NATURAL"

Agent Type	H1	H2	H3	H4	H5
Mean Score (App 1)	*4.82*	*2.92*	*3.82*	*3.83*	*NA*
Mean Score (App 2)	*3.50*	*3.95*	*3.64*	*4.78*	*NA*

TABLE XVI
MEAN SCORES FOR HUMAN-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "I LIKED THE APPEARANCE OF THE ASSISTANT"

Agent Type	H1	H2	H3	H4	H5
Mean Score	*2.79*	*2.00*	*1.84*	*NA*	*NA*

TABLE XVII
MEAN RATINGS SCORES FOR CARTOON-LIKE AGENT TYPE

Cartoon-Like Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Voice only)	*6.25*	*5.97*
C2 (2D Head)	*5.89*	*5.86*
C3 (3D Head)	*5.92*	*6.12*
C4 (2D Embodied)	*6.27*	*5.73*
C5 (3D Embodied)	*6.33*	*6.22*
C6 (3D Embodied in room)	*6.52*	*5.44*

B. Cartoon-Like Agents (Experiment 3)

The questionnaires for the third experiment addressed attitudes toward the application, the agents' voice, personality, appearance, facial expressions, and in addition, the gesturing of the embodied agents gesturing was also assessed. Based on the results from the human-like agent experiments, additional questionnaire items addressing further the agents personality and facial expressions were included in the cartoon-like experiment. Due to the fact that experiment 2 showed no overall application dependency, the participants (36 in total) in this third experiment witnessed the cartoon-like agents in one application only, i.e., the home furnishings service (application 1).

1) Overall Ratings: Participants were given the opportunity to rate each agent on a scale from 1 to 10 and the mean scores give an indication of participant preferences for the agents. A 2×6 repeated measures ANOVA taking agent gender and agent type as the independent variables and the mean rating scores as the dependent variable showed (Table XVII) significant results for agent gender ($F = 5.606$, $df = 1.0$, $p < 0.05$). The female agents were preferred to male agents (mean female $= 6.20$, mean male $= 5.89$: max $= 10$). In fact, an interaction between agent type and agent gender ($F = 5.134$, $df = 5.0$, $p < 0.01$), showed that agent types C1, C2, C3, C4, and C5 of both genders were rated similarly, but there was a very significant difference between female and male versions of C6 ($p < 0.01$). The female 3-D embodied agent (C6) that appeared in the 3-D environment was rated significantly higher than the male counterpart.

2) Attitude to Voices: A 2×6 ANOVA taking agent gender and agent type as the independent variables showed (Table XVIII) a significant effect for agent gender ($F = 7.37$, $df = 1.0$, $p < 0.01$) when participants were asked if they like

TABLE XVIII
MEAN SCORES FOR CARTOON-LIKE AGENT GENDER AND PARTICIPANT GENDER FOR QUESTIONNAIRE ITEM: "I LIKED THE ASSISTANT'S VOICE"

Participant Gender	Female	Male
Mean Score - Female Agents	5.70	4.87
Mean Score - Male Agents	4.77	4.90

TABLE XIX
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE ASSISTANT'S VOICE WAS ANNOYING"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	4.93	4.96	4.72	4.64	5.19	4.96

the voices of the assistants. The female voice was preferred to the male voice (mean female $= 5.28$, mean male $= 4.84$: max $= 7$). There was a greater preference from female participants for the voices of the female agents ($F = 8.645$, $df = 1.0$, $p < 0.01$). The male participants, however, liked the male and female voices equally.

Participants were asked if the voice of the assistant was natural and a 2×6 ANOVA taking agent gender and agent type showed that the voice of the male agents was thought to be significantly less natural ($F = 13.53$, $df = 1.0$, $p < 0.01$) than the female agents (mean female $= 4.93$; mean male $= 4.44$).

Another questionnaire item asked participants if they thought the voice of the assistant was annoying. Statistical results showed significant results for agent gender ($F = 10.96$, $df = 1.0$, $p < 0.05$). Participants felt the female voice was also less annoying than the male voice (mean female $= 5.15$; mean male $= 4.65$). This questionnaire item also showed significant differences between agent types ($F = 10.96$, $df = 1.0$, $p < 0.05$). Specifically, the results showed (Table XIX) that the voices of the 3-D embodied agents (C5 and C6) were significantly less annoying than all the other embodied agents (C2, C3, C4: all $p < 0.05$). As the same voice was used for all agent types of the same gender, this result highlights interesting effects about the perception of the voice based on the actual agent type, i.e., agent appearance.

3) Attitude to Politeness: Participants felt that all agents were polite, but the results from the 2×6 ANOVA showed (Table XX) significant differences for agent type ($F = 5.17$, $df = 5.0$, $p < 0.01$). C5 was significantly more polite than C2, C3, and C4 , all at $p < 0.05$. Overall, the disembodied voices (C1) and the 3-D fully embodied characters were thought to be more polite than the 2-D or 3-D heads. This result suggests 3-D embodied agents (i.e., C5) could 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 (Table XXI).

4) Attitude to Friendliness: A 2×6 ANOVA taking agent gender and agent type as the independent variables showed (Table XXII) significant differences for agent type ($F = 3.15$, $df = 5.0$, $p < 0.05$). No significant result for agent gender or any interactions emerged. Mean scores show that C5 and C6 were deemed to be most friendly. In fact, t-tests show that C5 and C6 were significantly more friendly than C2 and C3, all at $p < 0.01$. This result shows that fully embodied agents may play a more important role than 2-D or 3-D heads in participants' perception of agent friendliness.

TABLE XX
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE ASSISTANT WAS POLITE"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	5.78	5.59	5.50	5.58	5.83	5.72

TABLE XXI
MEAN SCORES FOR CARTOON-LIKE AGENT GENDER AND PARTICIPANT GENDER FOR QUESTIONNAIRE ITEM: "THE ASSISTANT WAS POLITE"

Participant Gender	Female	Male
Mean Score - Female Agents	5.89	5.39
Mean Score - Male Agents	5.71	5.68

TABLE XXII
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE ASSISTANT WAS FRIENDLY"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	5.48	5.18	5.19	5.39	5.63	5.65

TABLE XXIII
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE AND GENDER FOR QUESTIONNAIRE ITEM: "THE ASSISTANT WAS COMPETENT"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score – Female Agent	5.80	5.56	5.47	5.77	5.67	5.67
Mean Score – Female Agent	5.50	5.61	5.53	5.38	5.58	5.50

5) Attitude to Competence: A 2×6 ANOVA showed (Table XXIII) a significant effect for agent gender with respect to the competence of the agents. No effect for agent type emerged. Female agents were more competent ($F = 6.81$, $df = 1.0$, $p < 0.05$) than the male agents (mean female $= 5.66$, mean male $= 5.52$: max $= 7$). This was specifically the situation for C1, C4, C5, and C6, and, significantly, the case for C4 ($p < 0.05$). This result may be explained by the poor perception of the male voice and, in addition, the poorer acceptance of the male embodied agent's gesturing.

6) Attitude to Helpfulness: Participants were asked if they thought that seeing the assistants was helpful. The result showed that being able to see the female agents was thought to be significantly more helpful than seeing the male agents (mean female $= 4.62$, mean male $= 4.29$: max $= 7$). No significant effect for agent type emerged, and results indicate it was helpful to see all the agent types (Table XXIV).

7) Attitude to Appearance: A 2×5 ANOVA showed significant results for agent gender ($F = 16.5$, $df = 1.0$, $p < 0.01$).

TABLE XXIV
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "BEING ABLE TO SEE THE ASSISTANT WAS HELPFUL"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*4.48*	*4.54*	*4.25*	*4.66*	*4.33*

TABLE XXV
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE AND GENDER FOR QUESTIONNAIRE ITEM: "THE APPEARANCE OF THE ASSISTANT WAS UNSUITABLE"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score – Female Agent	*NA*	*4.86*	*4.89*	*4.72*	*5.16*	*4.91*
Mean Score – Female Agent	*NA*	*4.72*	*4.62*	*4.27*	*4.16*	*4.16*

The results showed (Table XXV) that appearance of the female agents were more suitable for the application than the male agents (mean female $= 4.91$; mean male $= 4.38$). An interaction between agent gender and agent type ($F = 3.04$, $df = 4.0$, $p < 0.05$) was also evident. T-tests showed that the male 3-D embodied agents (C5 and C6) were less suitable for the application than the female counterparts, both at $p < 0.01$. The qualitative findings indicated that participants perceived the male 3-D embodied gestures to be more dominating than the female 3-D embodied gestures causing this gender difference.

These results were reiterated when participants were asked if they liked the appearance of the assistants. A 2×5 ANOVA showed (Fig. 5) that 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 C4, C5, and C6 (2-D and 3-D fully-embodied agents), where the female appearance was significantly preferred to the male, all $p < 0.01$. Again, upon analysis of the qualitative interview data, many participants perceived the male embodied agents' hand as dominating and in turn distracting.

8) Attitude to Facial Expressions: Based on feedback from the previous experiments, additional questionnaire items pertaining to the agents facial expressions were included during the analysis of the cartoon-like agents. A 2×5 ANOVA showed (Table XXVI) significantly that the lip movements of C5 and C6 were less distracting than C2, C3, and C4 ($F = 3.996$, $df = 4.0$, $p < 0.01$). In the interviews, many participants said that the lip movement was distracting because it looked dubbed and that the lip movement of the 2-D and 3-D talking heads was more noticeable and looked artificial. No significant results for agent gender emerged.

A 2×5 ANOVA showed significant results for agent type ($F = 3.692$, $df = 4.0$, $p < 0.05$) with respect to the lifelikeness of the agents' facial expressions (Table XXVII). The facial expressions of C3 and C5 (the 3-D head and 3-D fully-embodied agent) appeared to be the most lifelike. Again, no significant effect for gender emerged. T-tests showed that the facial expressions of C3 were significantly more lifelike than C2, C4, and C6 ($p < 0.01$). Even though the face of C5 was smaller, making it more difficult for participants to judge them, it still had a mean score that was similar to that of C3, which was a talking head, where the facial expressions could be clearly seen.

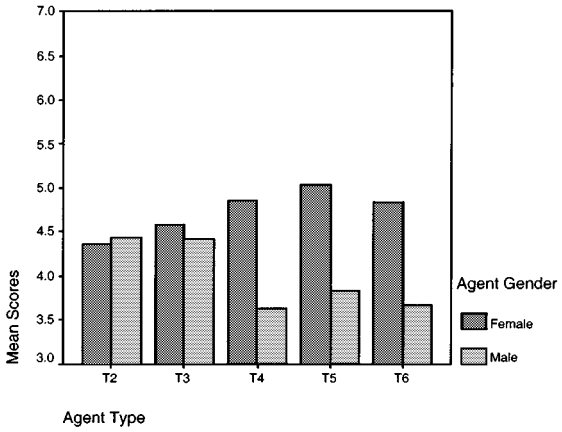

Fig. 5. Mean scores for cartoon-like agent type and gender for questionnaire item: "I liked the appearance of the assistant."

TABLE XXVI
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE LIP MOVEMENT WAS DISTRACTING"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*3.89*	*4.14*	*4.39*	*4.67*	*4.79*

TABLE XXVII
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR ITEM: "THE FACIAL EXPRESSIONS MADE THE ASSISTANT APPEAR LIFELIKE"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*3.89*	*4.48*	*3.68*	*4.12*	*3.80*

TABLE XXVIII
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE ASSISTANT'S GESTURES WERE EXAGGERATED"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*NA*	*NA*	*3.93*	*3.34*	*2.89*

9) Attitude to Gesturing: A 2×3 ANOVA taking agent gender and agent type showed that participants significantly ($F = 8.29$, $df = 1.0$, $p < 0.01$) preferred the female gestures to the male gestures (mean female $= 3.89$, mean male $= 3.69$: max $= 7$) and also thought the female gestures were significantly ($F = 11.7$, $df = 1.0$, $p < 0.01$) less exaggerated than the male agents (mean female $= 3.898$, mean male $= 3.694$: max $= 7$).

Another 2×3 ANOVA showed a significant effect for agent type ($F = 12.74$, $df = 1.0$, $p < 0.01$) when asked if the agents' gesturing was exaggerated (Table XXVIII). The results showed that the gesturing of all the agents was thought to be exaggerated (all below neutral) but, in particular, the 2-D embodied agents (C4) were less exaggerated than both of the 3-D embodied agents, all at $p < 0.05$.

Another 2×3 ANOVA produced significant results for agent type ($F = 3.78$, $df = 2.0$, $p < 0.05$) when asked if the gestures made the assistants appear lifelike (Table XXIX). In fact, C5

TABLE XXIX
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE GESTURES MADE THE ASSISTANT APPEAR LIFELIKE"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*NA*	*NA*	*3.23*	*3.79*	*3.59*

TABLE XXX
MEAN SCORES FOR CARTOON-LIKE AGENT TYPE FOR QUESTIONNAIRE ITEM: "THE GESTURES MADE THE ASSISTANT APPEAR FRIENDLY"

Agent Type	C1	C2	C3	C4	C5	C6
Mean Score	*NA*	*NA*	*NA*	*4.85*	*5.23*	*5.15*

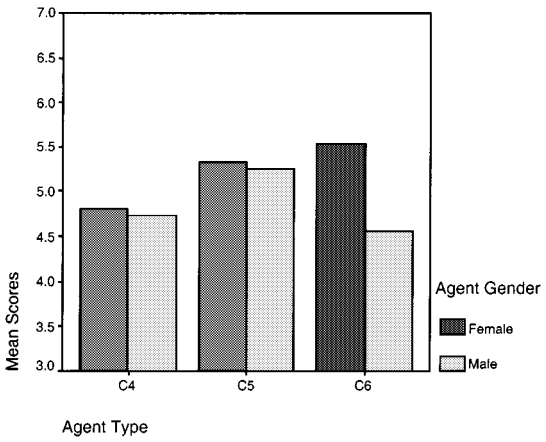

Fig. 6. Mean scores for cartoon-like agent type and gender for questionnaire item: "The gestures made the assistant appear friendly."

was perceived as being more lifelike than C6, and significantly more so than C4, $p < 0.01$.

Another 2×3 ANOVA produced significant results for agent type ($F = 3.60$, $df = 2.0$, $p < 0.05$) when asked if the gestures made the assistants appear friendly (Table XXX). In fact, C5 and C6 were significantly friendlier than C4, $p < 0.01$. Even though results showed the gesturing of C5 and C6 to be exaggerated, the gesturing did promote the lifelikeness and friendliness of the agents.

Finally participants were asked if the agents' gesturing appeared unhelpful. A 2×3 ANOVA produced significant results for agent gender ($F = 5.75$, $df = 1.0$, $p < 0.05$), agent type ($F = 15.78$, $df = 2.0$, $p < 0.01$), and an interaction between agent type and agent gender was also evident ($F = 3.85$, $df = 2.0$, $p < 0.05$). The gesturing of the 3-D embodied agents was perceived as being more helpful than the 2-D fully embodied agent, but this was significantly the case for female agents who appeared inside the 3-D space (Fig. 6).

VI. DISCUSSION

A. Human-Like Agents

The claim that the conversational interfaces would be liked was supported and the participant groups received the home furnishings application and CD service application positively. No application dependency issues emerged during the two evaluations of the human-like agents in the applications. The human-like agents were rated similarly for both applications. Between applications, one significant difference occurred. The still image (H4) was preferred in the home furnishings service. It is suggested that the presence of a still image of the artist in the CD service interface (Fig. 4), in addition to the still image agent, could have caused confusion for the participants, resulting in the difference between the applications for this agent type.

A number of issues emerged with respect to the voices of the human-like agents. Quantitative data indicated the female voice was not liked. The qualitative data reiterated this and the female voice received a number of negative comments. One participant commented that the "female voice was annoying and seemed disinterested." The voices of the agents need to be selected carefully and a clear fluent voice is desirable.

The appearance of the agents can also impact on the perception of the clarity of the agents' voice. The appearance of some agents was not liked (H2, H3, and H4), but also, these agents' voices were perceived to be less clear, despite the fact the voice output from agents of the same gender was identical. In addition, the conversation with the videos (H1) was thought to be more natural than the conversation with all other human-like agents, including the disembodied voice, despite the fact that the conversation between the agent and the customer was identical. This result suggests that conversations with human-like videos may be more natural than the other human-like agent types for e-retail applications.

In both applications, there was significant preference for the videos and disembodied voices over the other agents in the cast. Participants had a preference to interact with agents that exhibited human-like facial expressions and nuances during the conversation to complement the human-like appearance of the agents. This is consistent with experimental findings by Reeves and Nass [18], who discovered participants prefer to interact with agents who have consistent personalities, which reflects more human-like behavior. When the participants did not see such human-like behavior they preferred not to see a visual display of the agent, but preferred a disembodied voice. The popularity of the disembodied voices raises interesting issues about the need for synthetic human-like agents in e-retail interfaces. In addition to ensuring the agents' behavior corresponds to the visual display, the application task should afford a visually displayed agent. Some participants did comment that the image of the assistant distracted them from changes that were being made in the interface. It is therefore necessary for interface designers to assess services and applications carefully to establish if a personified agent is an actual enhancement and not a distraction.

Signals of friendliness and politeness were given high priority by participants for enhancing the retail applications. A number of participants from both sets of experiments suggested using cartoon agents as the assistants if it was not possible to use video images of humans. It may be more appropriate to have animated characters in the interface, rather than trying to make the agents completely human-like, raising user expectations above the capabilities of the agent.

B. Cartoon-Like Agents

The third experiment was largely encouraging for the use of animated agents in e-retail interfaces. Participants enjoyed the conversational capabilities of the application and two-thirds (24 from 36) preferred to see an agent in the interface. It was found that the 3-D agents were preferred to 2-D agents and that the 2-D and 3-D fully embodied agents were preferred to 2-D and 3-D heads. It must be remembered that one third of this participant group 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 if necessary. However, if the nonverbal behavior is developed to provide essential information to the user, the user may then prefer to look at the agent.

The results showed the general preference for the female voice, the opposite finding to the human-like agent experiments. Previous findings suggested that fluent conversational voices should be used, and it was attempted to use such voices for the cartoon-like agents. The female voice was "more friendly." The male voice seemed "monotonous." The results show it is important to select the voices of the agents carefully.

It was also claimed that attitudes to male and female agents of the same agent type would be similar. In fact, the results showed that the female agents gesturing was preferred; this can be explained by the largely negative attitude to the gesturing of the male embodied agent (C6) who appeared in the 3-D world. Overall, the competence and helpfulness of the female agents was thought to be better. Again, this can be explained by the poor perception of the male gestures, in combination with the poorer perception of the male voice. Differences between the agent types also emerged, and results showed that the fully embodied agents were thought to be more friendly, helpful, and polite.

Interesting research issues about the perception of embodied agents' personality and the movement of the animated body have been introduced. It was shown that a 3-D agent appearing in a 3-D world was less desirable than the 3-D agent appearing outside the world. There are two explanations to suggest that participants had a preference for the 3-D embodied agent appearing outside the 3-D world. First, participants felt that the agents distracted them from the changes that were happening in the interface; the agents tended to block the view of the 3-D living room space. It is important to then remember that not all 3-D environments can be suitably inhabited with 3-D animated agents. Depending on the application task, it may be more appropriate for the 3-D agent to appear outside the world or, in fact, it may be better to use a disembodied voice if the task demands too much attention from the user. Alternatively, the agent should have adequate mobility so as not to block the users' view. Second, the male 3-D embodied agent in the 3-D world was poorly accepted because the gesturing was perceived as being dominating and larger than its female counterpart. For this reason, the results show that embodied agents appearing in the 3-D environment were not accepted. Of the participants who did like the 3-D embodied agents appearing in the room, they said it was "more complete," "natural," and "it added to the realism." It is possible that if the dimensions of the agent are in proportion and the gesturing is less exaggerated, the agent may in fact be acceptable in the 3-D environment regardless of the task.

VII. Conclusion

This paper has provided the agents community with new facts about various aspects of the inclusion of human-like and cartoon-like agents in e-retail environments.

- If using a human voice, consider and choose it carefully. The voice should be friendly and conversational with intonation.
- If using an agent created from a photorealistic image of a human, make sure the agent is lifelike and has appropriate human-like nonverbal communicative behavior.
- Select human-like gestures and facial expressions to complement the agent's human-like appearance; exaggerated gesturing could undermine the users' perception of the agent. When using cartoon-like humanoid talking heads or humanoid embodied agents, 3-D agents can be more appealing to the user than 2-D agents. Gesturing can promote the users' perception of friendliness, politeness, and lifelikeness.
- Ensure the dimensions of the agent are in proportion with the dimensions of the 3-D environment in which it appears, especially if the agent appears inside the 3-D virtual environment.
- Examine the application task carefully and assess if the agent will distract the user from the task when it appears in the 3-D environment.

Acknowledgment

The authors would like to thank their colleagues at the University of Edinburgh, particularly Dr. J. Foster, for their input and suggestions and the research staff at BT Adastral Park for their helpful planning of the experiment.

References

[1] E. André and T. Rist, "Personalizing the user interface: Projects on life-like characters at DFKI," in *Proc. 3rd Workshop Conversational Characters*, Oct. 1998, pp. 167–170.

[2] N. Badler, M. Palmer, and R. Bindiganavale, "Animation control for real-time virtual humans," *Commun. ACM*, vol. 42, no. 8, pp. 65–73, Aug. 1999.

[3] J. Cassell *et al.*, *Embodied Conversational Agents*. Cambridge, MA: MIT Press, 2000.

[4] J. Cassell and K. Thorisson, "The power of a nod and a glance: Envelope versus emotional feedback in animated conversational agents," *J. Appl. Intell.*, vol. 13, no. 3, pp. 519–538, 1999.

[5] J. Cassell *et al.*, "An architecture for embodied conversational characters," in *Proc. 3rd Workshop Conversational Characters*, Oct. 1998, pp. 21–30.

[6] D. Dehn and S. van Mulken, "The impact of animated interface research: A review of empirical research," *Int. J. Human-Comput. Stud.*, vol. 52, no. 1, pp. 1–22, Jan. 2000.

[7] T. Ezzat and T. Poggio, "MikeTalk: A talking facial display based on morphing visemes," in *Proc. Comput. Animation Conf.*, June 1998, pp. 96–102.

[8] A. Guye-Vuillieme *et al.*, "Non-verbal communication interface for collaborative virtual environments," *Virtual Reality J.*, vol. 4, pp. 49–59, 1999.

[9] J. King and J. Ohya, "The representation of agents: Anthropomorphism, agency, and intelligence," in *Proc. CHI Conf. Companion Human Factors Comput. Syst.: Common Ground*, Apr. 1996.
[10] T. Koda, "A study on the effects of personification of software agents," M.Sc. dissertation, Mass. Inst. Technol., Cambridge, 1996.
[11] J. Lester and B. Stone, "Increasing believability in animated pedagogical agents," in *Proc. 1st Int. Conf. Autonom. Agents*, Feb. 1997, pp. 16–21.
[12] J. Lester *et al.*, "The persona effect: Affective impact of animated pedagogical agents," in *Proc. Human Factors Comput. Syst.*, Mar. 1997, pp. 359–366.
[13] R. Likert, "A technique for the measurement of attitudes," *Archives Psychol.*, vol. 140, p. 55, 1932.
[14] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.
[15] H. McBreen and M. Jack, "Animated conversational agents in E-commerce applications," in *Proc. 3rd Workshop Human-Comput. Conversation*, July 2000, pp. 112–117.
[16] H. McBreen *et al.*, "Experimental assessment of the effectiveness of synthetic personae for multimodal E-retail applications," in *Proc. 4th Int. Conf. Autonom. Agents*, June 2000, pp. 39–45.
[17] F. Parke and K. Waters, *Computer Facial Animation*. Cambridge, MA: A. K. Peters, 1996.
[18] B. Reeves and C. Nass, *The Media Equation*. Stanford, CA: Stanford Univ. Press, 1996.
[19] J. Rickel and L. Johnson, "Task oriented dialogs with animated agents in virtual reality," in *Proc. 3rd Workshop Embodied Conversational Agents*, 1998, pp. 39–46.
[20] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge, MA: Cambridge Univ. Press, 1982.
[21] A. Takeuchi and K. Nagao, "Communicative facial displays as a new conversational modality," in *Proc. Human Factors Comput. Syst.: InterCHI*, Apr. 1993, pp. 187–193.
[22] D. Thalmann, "Autonomous virtual humans in virtual environments," in *Proc. 4th Int. Conf. Autonom. Agents*, June 2000, Tutorial 6.
[23] N. Magnenat-Thalmann *et al.*, "Face to virtual face," *Proc. IEEE.*, vol. 86, May 1998.
[24] K. Thorisson, "Communicative humanoids: Model of psychosocial dialogue skills," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1996.
[25] S. van Mulken, E. André, and J. Muller, "The persona effect: How substantial is it?," in *Proc. Human Comput. Interaction Conf.*, Berlin, Germany, 1998, pp. 53–66.
[26] H. Walker, L. Sproull, and R. Subramani, "Using a human face in an interface," in *Proc. Human Factors Comput. Syst.*, Apr. 1994, pp. 85–91.
[27] K. Waters and L. Levergood, "DECface: A system for synthetic face applications," in *Multimedia Tools Applicat.*. Boston, MA: Kluwer, 1995, vol. 1, pp. 349–366.
[28] P. Wyard and G. Churcher, "The MUeSLI multimodal 3-D retail system," in *Proc. ESCA Workshop Interactive Dialogue Syst.*, Sept. 1999.

Helen M. McBreen received the B.E. degree with honors in electronic engineering from the National University of Ireland, Dublin, in 1998.

In 1998, she joined the Centre for Communication Interface Research (CCIR), University of Edinburgh, Edinburgh, U.K., to pursue her Ph.D. thesis, where she is investigating the use of embodied conversational agents in e-retail applications. Her research interests include speech technology, the design of autonomous animated agents, and the development of socially intelligent agents.

Mervyn A. Jack received the B.Sc. degree in electronic engineering and the M.Sc. degree in digital techniques, both from Heriot–Watt University, Edinburgh, U.K., and the Ph.D. degree in engineering from the University of Edinburgh, in 1971, 1975, and 1978, respectively.

He is Professor of Electronic Systems, University of Edinburgh. He leads a multidisciplinary team of 20 researchers at the Centre for Communication Interface Research (CCIR), University of Edinburgh, investigating usability engineering of eCommerce services. His main research interests are dialogue engineering and virtual reality systems design for advanced eCommerce and consumer applications.