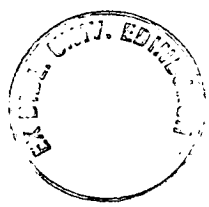
Embodied Conversational Agents

Extending the Persona Metaphor to Virtual Retail Applications

Helen McBreen

Ph.D. Thesis

2002



Declaration of Originality

February 28th, 2002

This thesis is submitted in partial fulfillment for the degree of Doctor of Philosophy. I declare that it has been composed by myself, and the work described is my own research.

'What you see here is an Andralad of my making, moulded for the first time by the amazing vital agent we call electricity. This gives to my creation the blending, the softness, the illusion of life.'

'An Andralad?'

'Yes', said the professor, 'a human-imitation, if you prefer that phrase.'

L'Eve Future (Villiers de I'Isle Adam, 1960)

Acknowledgements

Many people, from many disciplines, have influenced this research. I give particular thanks to researchers, both past and present, at the Centre for Communication Interface Research (CCIR). Dr. Steve Love and Dr. Iain Nairn particularly influenced early research. Technical contributions from Dr. James Anderson, statistical advice from Dr. Fergus McInnes, psychology explanations from Ms. Stella Topalidou and enthusiastic conversations about human animation with Mrs. Rachel Whitby, enriched later work. Important contributions were also made during experiment completion and design from Ms. Shamuna Ali, Mr. Nicholas Anderson, Mr. Craig Douglas, Mr. Craig Douther and Mr. Paul Shade.

Throughout the duration of this thesis Dr. John Foster engaged in helpful commentary, and endless discussions on key research topics and related academic publications. My sincere appreciation is also expressed to my supervisors, Dr. Iain McKay and also to Professor Mervyn Jack for his tireless support and encouragement to complete innovative research and to also present this research to an international community.

Researchers at other universities have been influential through their writings and advice, including researchers at the Gesture and Narrative Language Group at Massachusetts Institute of Technology (MIT) Media Lab; the German Research Centre for Artificial Intelligence (DFKI), in particular Dr. Elisabeth André; Dr. Kerstin Dautenhahn at the Adaptive Systems Research Group, University of Hertfordshire and finally Dr. Lewis Johnson at the Information Sciences Institute, University of Southern California.

Doctoral students in the Division of Informatics at The University of Edinburgh have been excellent conversationalists and have helped shape this work. Thanks is also given to researchers, students and friends, from various departments at The University of

Edinburgh and also to previous colleagues from the Faculty of Engineering at University College Dublin.

Finally, I thank my family, in particular my sister Sarah, who did all this before and of course to my parents for inspiring me, through their enthusiasm, to pursue academic research.

Abstract

Engineering computer interfaces to communicate through the modality of speech serves to bridge the communicative gap between computers and their human users. Adding non-verbal performances to these spoken language interfaces, through the creation of embodied conversational agents, initiates dialogues where users' innate communicative capabilities are used for the benefit of a more engaging and effective interaction. Anthropomorphising the interface with lifelike behaviour animates the communicative process and recent research suggests that the extension of this persona metaphor into retail applications will provide personalised real-time interactions for users, improving on-line relationships.

This thesis is a contribution to the emerging and innovative area of embodied conversational agents. It is undertaken to advance knowledge about the effectiveness of these agents in electronic retail applications. To date, few empirical studies have documented user perceptions of the agents and this has thus become the primary research objective of this thesis. The interdisciplinary investigation aims to determine how embodied conversational agents should be physically represented in retail interfaces. The research involves the undertaking of a series of progressive empirical evaluations. Firstly, a retail interface template was created, where variations of the persona metaphor were evaluated using participatory observation techniques. Following this an interactive spoken language system, inhabited with embodied conversational agents was designed and implemented to serve as an experiment platform from which to evaluate the perceived and expected behaviour of agents in contrasting retail applications. This interactive system was then used to determine the effectiveness of multi-modal interface

features designed to improve user perceptions of the trustworthiness of the agents in applications where users are asked to make financial transactions. The conclusions drawn from this body of research support the deployment of embodied conversational agents in virtual retail applications. Design guidelines and interface development strategies for the most effective deployment of these agents are presented.

Table of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Publications

McBreen, H., and Jack, M. (2001). Evaluating Humanoid Synthetic Agents in E-Retail Applications. In *IEEE Transactions on Systems, Man and Cybernetics - Special Issue on Socially Intelligent Agents: The Human in the Loop*, ed. K. Dautenhahn.

McBreen, H. (2001). Embodied Conversational Agents in ECommerce Applications. In K.Dautenhahn, A. Bond, D. Canamero & B. Edmonds (eds.) *Socially Intelligent Agents - Creating Relationships with Computers and Robots.* Kluwer Publications.

McBreen, H., Anderson, J. & Jack, M. (2001). Evaluating 3D Embodied Conversational Agents in Contrasting VRML Retail Applications. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents*, 5th International Conference on Autonomous Agents, 83-8.

McBreen, H., and Jack, M. (2000). Empirical Evaluation of Animated Agents In a Multi-Modal Retail Application. Proceedings of the *AAAI Fall Symposium: Socially Intelligent Agents – The Human in the Loop*, Nov. 122-126. ISBN 1-57735-127-4.

McBreen, H., and Jack, M. (2000). Animated Conversational Agents in ECommerce Applications'. Proceedings of the *3rd Workshop on Human-Computer Conversation*, 112-117.

McBreen, H., Shade, P., Jack, M., and Wyard, P. (2000). Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications. Proceedings of the *4th International Conference on Autonomous Agents*, June, 39-45, ACM Press. ISBN 1-581-13230-1.

Contents

Chapter 1

1.1 Introduction

Speech expresses our thoughts and notions. When complemented with non-verbal behaviour, speech animates our presence in an effort to communicate more effectively. New possibilities for human computer interaction are created when verbal and non-verbal performances are introduced to computer interfaces. Embodied conversational agents (ECA) are created to introduce this verbal and non-verbal behaviour to the interface. These agents are reactive to user's speech input and capable of verbal and non-verbal output. They may serve in a vast range of interface domains, including interactive educational interfaces, games environments and with increased computational developments and the ubiquity of Internet technologies, their functionality may be extended to electronic retailing domains.

The spirit of this thesis is a contribution to the emerging and innovative area of embodied conversational agents. It is offered as a hopeful advancement for the deployment of communicative entities, such as embodied conversational agents in electronic retail interfaces. To date, few empirical studies have documented user perceptions of embodied conversational agents and from those that have it is difficult to draw general conclusions, as will be discussed in Chapter 2. This interdisciplinary investigation describes a search for the appropriate humanoid embodiments that can best enhance the computer interfaces. To conduct such research, progressive usability evaluations were completed to investigate users' perceptions of embodied conversational agents, to determine whether ECA are engaging and to explore novel retail environments in which these agents may reside. This chapter introduces the objectives of the research, presents an overview of the thesis and defines the necessary terms and definitions used throughout the text of the thesis.

1.2 Terms and Definitions

The following terms are thought to be the most important to define with respect to the work documented in this thesis. Research conducted in the area of human-computer interaction, and in particular with the creation and evaluation of embodied conversational agents is often best conducted in 'interdisciplinary' environments. Such

research groups combine the skills of software engineers, graphic designers, mathematicians, linguists and psychologists.

The term 'embodied conversational agents' coined by Cassell et al. (2000) refers broadly to animated characters, human-like and cartoon-like, animate and inanimate, that often appear in interfaces. The agents are endowed with conversational capabilities primarily through speech recognition software, natural language processing and speech output generation. Humanoid agents are also often referred to in related literature as virtual humans, virtual characters or virtual agents. These agents are thought to 'anthropomorphise' the interface by bringing lifelike or human-like qualities to the interaction. Such lifelike interactions use 'multi-modal communication', which implies the use of all or a combination of oral, auditory and visual signals during a human-computer interaction. The use of such modes can initiate the growth of a social interaction between the agent and the user and therefore the creation or illusion of the existence of a 'socially intelligent agent'. Such agents are by definition, software or robotic entities that behave socially. The creation of 'social intelligence' is geared toward creating human-style intelligence in communicative agents in order to better support interactions between humans and computers.

1.3 Thesis Outline

An exploration of the research motivations behind the investigations of communicating with embodied conversational agents is provided in Chapter 2. Opening with a historical analysis of early attempts at creating talking machines, the timeline moves forward to a presentation of contemporary technological capabilities, encompassing speech generation and speech recognition software. The persona metaphor is introduced and the advantages of visually anthropomorphising an interface are discussed. To explain these advantages it is necessary to understand theories that support face-to-face interaction, and background research is provided to assist with this understanding. The chapter then introduces the technologies available to actually create embodied conversational agents and highlights the success of such communicative entities in educational domains. Following this, related research into virtual reality, virtual digital narratives and the virtual environments in which embodied agents could inhabit is presented, with particular focus on the extension of the persona metaphor into retail domains. This

literature review then turns to the state-of-the-art with respect to intelligent agents, focusing in particular on the current research into the creation of socially responsive embodied agents and emotionally adaptive agents. From this literature review the need for empirical evaluations of humanoid embodied conversational agents is confirmed.

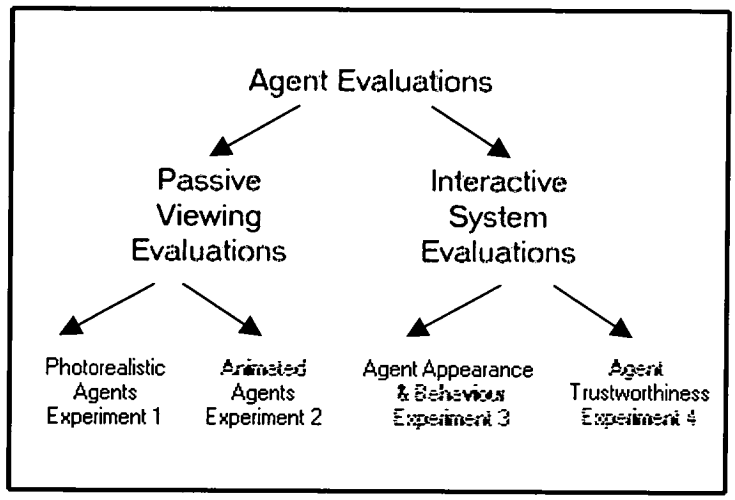


Figure 1.1 Map of Evaluations

The thesis describes four individual but progressive empirical evaluations, whereby the functionality and representation of embodied conversational agents were assessed (Figure 1.1). Chapter 3 provides a detailed discussion of the experiment techniques used to conduct these usability evaluations. The evaluation methods employed throughout the thesis are explained, followed by a description of the information retrieval strategies and statistical computation methods used. The four chapters that follow present the four empirical evaluations respectively. The aims of the first two experiments were to explore the effectiveness and user acceptability of the humanoid photo-realistic and humanoid animated agents in multi-modal electronic retail scenarios. The experiments were steered toward the long-term aim of the thesis, which is to produce design guidelines for the creation of effective embodied agents to be used in intelligent multimedia and multi-modal applications involving automatic speech recognition and speech generation technologies. Figure 1.2 illustrates that embodied conversational agents can be realised and represented in a variety of formats and, as will be explained clearly in Chapter 2, little research is available to describe which agent types best match user expectations.

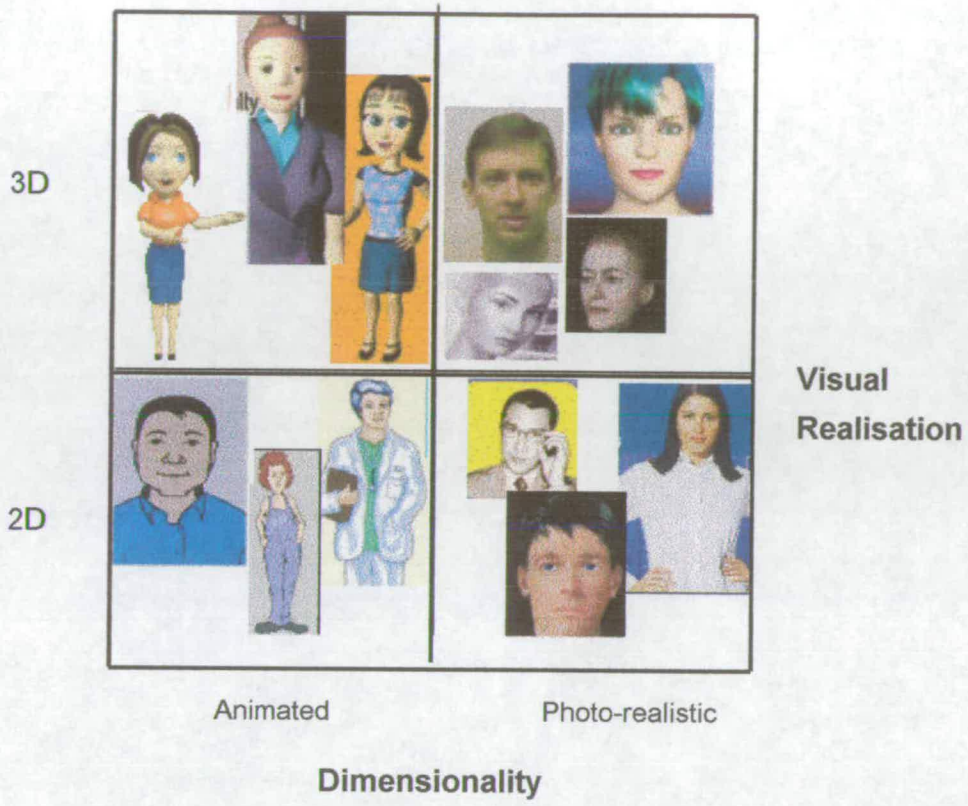


Figure 1.2 Array of Humanoid Embodied Conversational Agents

In Chapter 4 a discussion of the retail interface template that was designed and implemented in order to evaluate by observation a range of humanoid embodied agents is presented. After a technical description of the design used to create the retail template, experiment data from the first passive viewing evaluation is presented. This experiment specifically evaluated a range of humanoid *photo-realistic* agents. Consideration was paid to the possible technologies that may be used to create *photo-realistic* agents and in addition the gender of the agent was represented as part of a repeated measures experiment design. Using the same retail interface template a second passive viewing evaluation was conducted, this time specifically focusing on humanoid *animated* agents. The experiment, presented in Chapter 5, again used a repeated measures method to assess user perceptions of different representations of male and female animated agents who appeared, as did the photo-realistic agents, as virtual assistants in retail applications.

Using the results from the first two experiments the most popular embodied agent type was selected and used for the creation of an interactive system in which the agent appeared as a functional conversational agent (Chapter 6) in a variety of contrasting

retail applications. After a detailed technical description of the construction of the interactive system, two empirical evaluations are presented. In the first (Chapter 6), participants were invited to converse with the embodied agents and through these user evaluations, design guidelines and development strategies for the successful deployment of ECA in retail interfaces are presented. Following this, the fourth and final evaluation in the progression of empirical studies, explores the effectiveness of multi-modal interface features added to the retail interfaces as a mechanism to inspire greater confidence in the applications and thus boost user perceptions of trust in the agents, particularly in applications where the user must perform personal financial transactions. A thorough discussion of the research findings from all four evaluations is presented in Chapter 8. This is followed by a presentation of the impact, or contribution that this work has made for the use of embodied conversational agents in future applications.

The thesis identifies, for the first time, key design features for embodied conversational agent in terms of the visual representation of the agent in retail applications. The research indicates that conversational interfaces inhabited with agents, in particular those represented as human-like video agents and 3D fully-embodied agents are significantly preferred in a variety of retail interfaces. The thesis then contributes evidence in support of multi-modal (text and speech) interfaces inhabited with 3D fully-embodied agents as means to improve user attitudes to retail environments in which users are required to complete financial transactions.

Chapter 2

Motivations To Research Communication with Embodied Conversational Agents in Virtual Retail Applications

2.1 Introduction

A major research goal of the interdisciplinary field of human computer interaction (HCI) is the design of new interactive schemes that better fit with and exploit the communicative characteristics of humans, in order to reduce the gap between humans and machines. Such interactive schemes must be robustly designed to support the complex communication process. A particularly salient mode of communication currently being explored in the design of new schemes is the use of spoken language interfaces through which users communicate with their computers via that most natural modality of human communication - speech. According to Zue and Cole, conversing with a machine represents the ultimate challenge to our understanding of the production and perception processes involved in human speech communication (Zue & Cole, 1995). The idea of creating a talking machine was born, however, long before the digital era.

> The earliest speaking machines were perceived as the heretical works of magicians and thus as attempts to defy God. In the thirteenth century the philosopher Albertus Magnus is said to have created a head that could talk, only to see it destroyed by St. Thomas Aquinas, a former student of his, as an abomination. The English scientist-monk Roger Bacon seems to have produced one as well. That fakes were appearing in Europe in the late sixteenth and early seventeenth centuries is shown by Miguel de Cervantes's description of a head that spoke to Don Quixote -- with the help of a tube that led to the floor below. Like Magnus, this fictitious inventor also feared the judgement of religious authorities, though in his case he took it upon himself to destroy the heresy. By the eighteenth century science had started to shed its connection to magic, and the problem of artificial speech was taken up by inventors of a more mechanical bent. (Lindsay, 1997).

As the time line in Figure 2.1 illustrates, the effort to create talking machines using a more scientific approach dates back to the 18th century and the challenge of creating a communicative device has altered little since then.

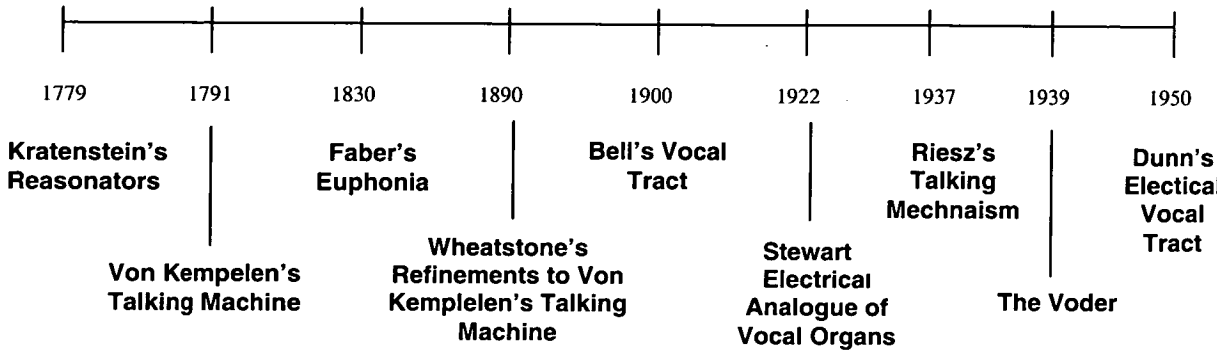


Figure 2.1 A Time Line of Early Talking Machines

A brief description of some of the early attempts at creating talking machines is presented below and using the work of Rubin and Vatikiotis-Bateson (1998), images of the devices are also provided. The earliest talking machines were complex devices dominated by intricate mechanical workings. For instance, in 1779 Kratzenstein successfully constructed a series of acoustic resonators patterned after the human vocal tract. By varying the resonators, limited speech in the form of the five vowel sounds was produced. Figure 2.2 illustrates the resonators that were operated by blowing through a reed.

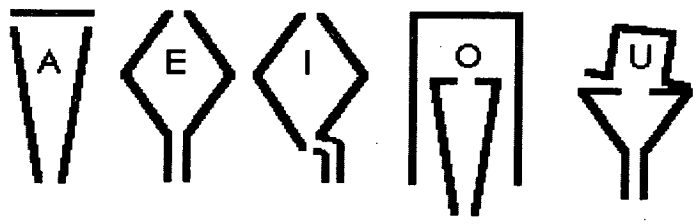


Figure 2.2 Kratzenstein Resonators

Parallel with Kratzenstein efforts, von Kempelen constructed a machine that could generate connected speech. A bellows supplied air to a reed that stimulated a resonator to produce sounds. Von Kempelen aimed to produce continuous speech and through skilled handling and control of the device, the machine was essentially "played" like an instrument to produce distinguishable consonants. In the late 1800's Sir Charles Wheatstone improved some operational aspects of the device, but from that point the machines received very little attention (Figure 2.3).

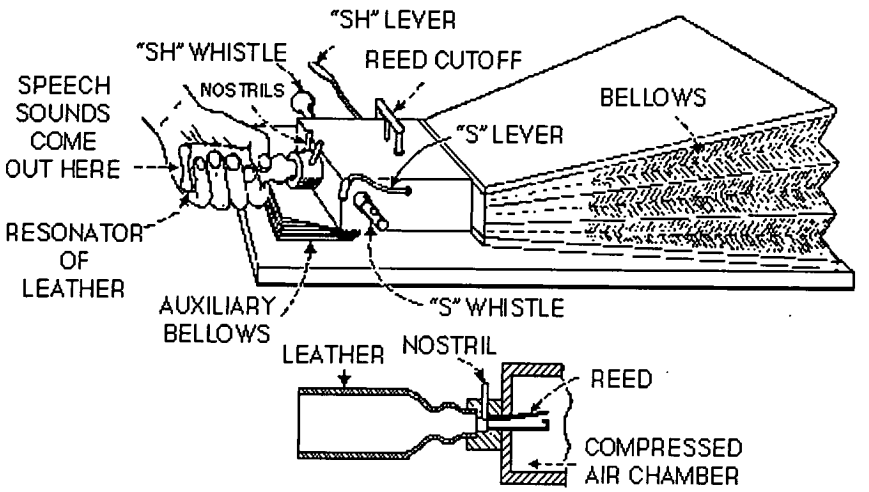


Figure 2.3 Wheatstone's Version of Von Kempelen's Machine

Despite the lack of interest in talking machines, Alexander and Melville Graham-Bell continued on the quest to create the ultimate talking device and produced a physical working model of the human vocal organs, making a cast from a human skull and moulding the lips, tongue, palate, teeth, pharynx and velum. This very complicated device was successful at producing simple utterances, but like its forbearers received little attention after its construction.

In the 1830's Faber's Talking Machine, known as The Euphonia was constructed. In describing the machine Lindsay (1997) states:

> [It is] a speech synthesizer variously known as the Euphonia and the Amazing Talking Machine. By pumping air with the bellows and manipulating a series of plates, chambers, and other apparatus (including an artificial tongue), the operator could make it speak any European language. A German immigrant named Joseph Faber spent seventeen years perfecting the Euphonia, only to find when he was finished that few people cared.

An image of the Euphonia is presented in Figure 2.4. Interestingly, the static face that rests on the side of the machine indicates the first visual personification of a talking device, in a sense the ultimate point of departure for this thesis.

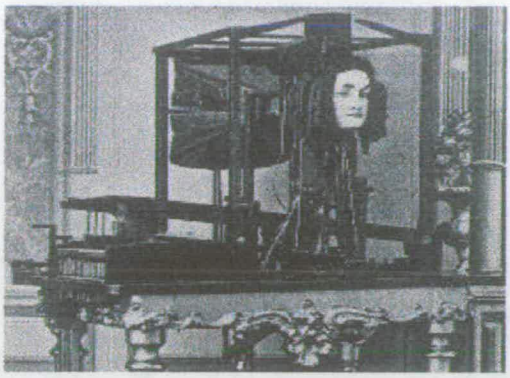


Figure 2.4 The Euphonia

Riesz's 1937 talking machine, again similar in construction to a musical instrument, could produce relatively articulate speech and in fact the complex design allowed the control of every movable part of the human vocal tract (Figure 2.5). It may have been the complexity involved in operating these cumbersome mechanical devices that explains the lack of interest in them.

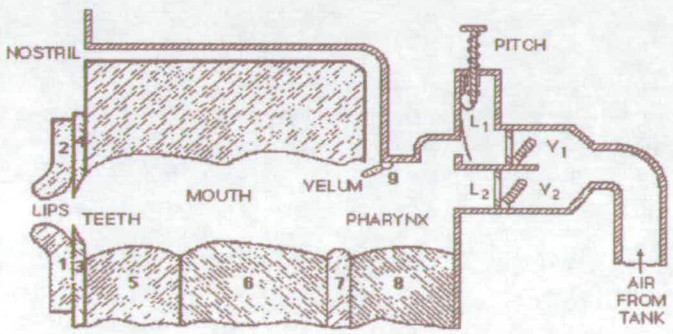


Figure 2.5 Riesz's Talking Machine

Stewart made the first transition to an electrical device in 1922 (Figure 2.6). Although his construction was crude it does illustrate the turning point from mechanical devices to their electrical counterparts. An electrical buzzer simulated a combination of inductive and capacitive resonators creating resonance to produce two distinct electrical analogue frequencies. The variation of capacitance, resistance and inductance led to the production of a number of vowel sounds.

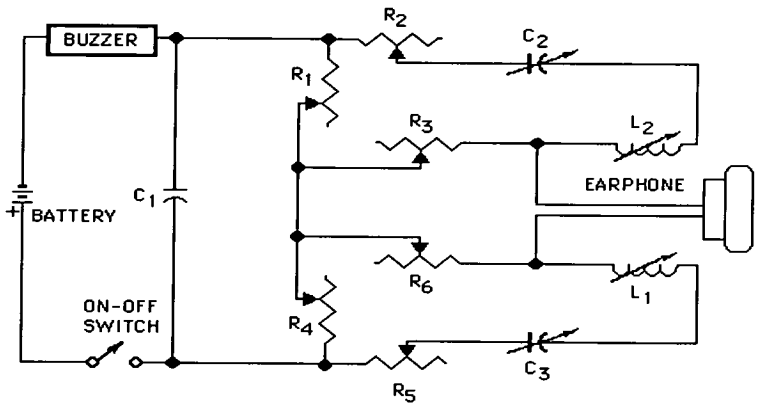


Figure 2.6 Stewart's Electrical Circuit

As with the mechanical predecessors, the next logical transition was to produce an electrical device that could output connected speech. This was the success of Dudley, Riesz and Watkins when they created the Voder in 1939 (Figure 2.7). The resonators were connected to a finger keyboard and through a buzz oscillator the keys stimulated different filters arranged in parallel to produce sounds to form a speech frequency range. This device also utilised a foot pedal to control the pitch of the buzz oscillator. Rather like playing a piano, the Voder could only be correctly used by experienced operators.

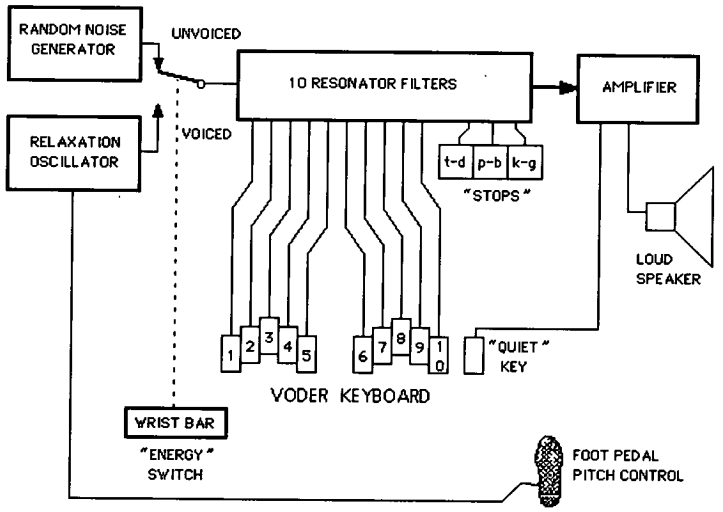


Figure 2.7 The Voder

In 1950, The Haskins Laboratories produced the Pattern Playback Machine which is described as a sonagraph working in reverse, transforming frequency and altering

intensity over time as depicted on a paper roll, to produce intelligible speech sounds (Figure 2.8).

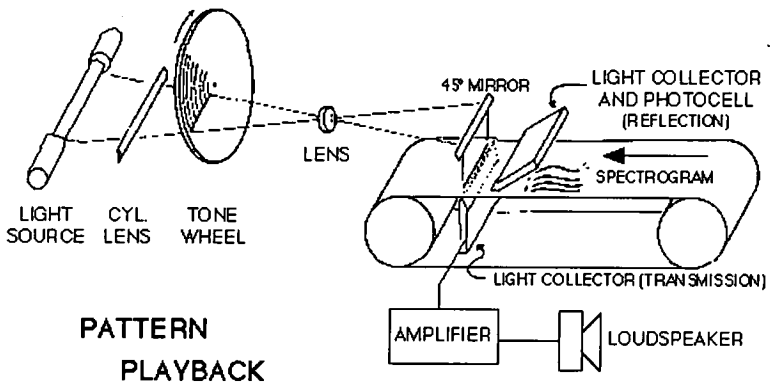


Figure 2.8 Haskins' Pattern Playback Machine

Also in the 1950's, Dunn's Electrical Vocal Tract was constructed. This device used transmission line theory and by modelling the vocal tract as a transmission line, it was possible to generate vowel resonance (Figure 2.9).

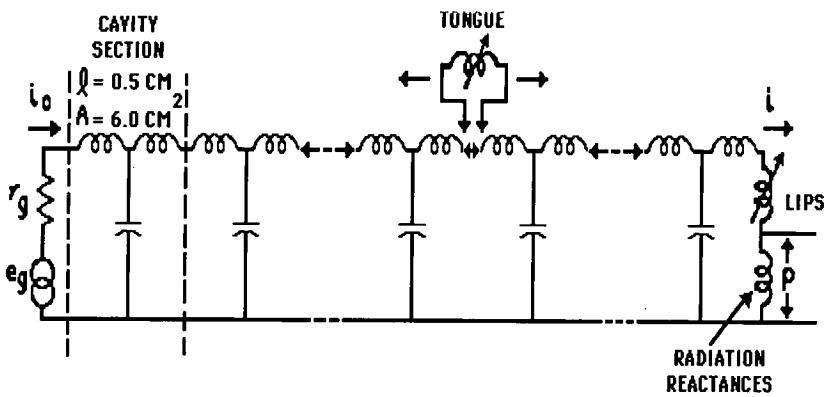


Figure 2.9 Dunn's Electrical Vocal Tract

From this point, models of the human vocal tract could be made with electrical circuits and the mechanical talking devices became redundant. However, it was the simulation of electrical circuits using computers and the conversion from analogue to digital signals that truly initiated a revolution in the advances being made in speech technology research.

2.2 Contemporary Speech Systems

According to Cohen and Oviatt (1995) the birth of the computer era was dominated by the vision of futurists who dreamed of the conversational computer. Since the 1950's steady progress has been made with regard to speech recognition and it is such progress that will eventually enable the creation of computers with Turing test capabilities, machines that can conduct a fluent conversation thus making them indistinguishable from humans. To encourage this persona metaphor, visual personification of the interface has become popular and it is the concern of this thesis to assess the effectiveness of this persona metaphor through empirical research in order to help bridge the cognitive space between users and their computers.

Speech is a natural mode of communication, consequently it is expected that spoken language systems will allow people to complete innovative tasks more easily and more conveniently. This is supported by Flanagan (1995), who states that "because the sensory modalities are highly learned and natural, we seek to endow machines with the ability to communicate in these terms". Although the research challenge has not altered, and the focus still lies with creating the ultimate talking machine, much of the emphasis now is the integration of continuous speech recognition software, combined with speech synthesis technology in order to produce an effective, flexible and natural communicative speech interface.

Bristow (1986) explains that prior to the electronic age the possibilities of creating a device that could recognise speech were impossible, as the simplest attempt at recognition required electronic equipment, to capture the utterance with a microphone and then to analyse the signal. In 1947, the Spectrogram became the first device that demonstrated the capability of producing a graphical representation of a continuous spectrum of speech. The device brought with it considerable optimism for the future of human machine interaction using voice. After 1950 efforts continued to produce an automatic recognition device. This came in the form of a Stenosonograph, produced by Dreyfus-Gaf. Using electronic filtering, the input to the device was guided through six band-pass filters and used to deflect an oscilloscope beam in different directions, giving each speech sound a unique position (Bristow, 1986). This led to the production of the first automatic speech recogniser, which was created in 1952 by Bell Labs. It used the basic principle of storing information representing various speech sounds for comparison

with the (unknown) speech input. The input signal was filtered above and below 900Hz, and the frequency of energy in both bands was measured. The graph of these two parameters was plotted against time and compared with similar stored profiles. The closest stored word template to the input was selected as the input utterance.

Once the silicon revolution began it became feasible for any algorithm to be stored on an integrated chip. Bristow (1986) reports that in the 1970's the algorithms for speech synthesis were mature enough to be stored on silicon chips and it was during this decade that some of the first speech synthesisers became commercially available. However, the same advance was not made with respect to speech recognisers for another decade due to the fact that the process of speech recognition requires lengthy calculations for comparing models and the first large scale integrated circuits (LSI) capable of such processing only became available in the 1980's.

Although a detailed discussion of the computations involved in speech recognition is beyond the scope of this thesis, a brief description is provided. Most automatic speech recognition (ASR) approaches involve speech pattern matching (SPM). This procedure is based on the principle that for a machine to recognise speech it must have prior knowledge about the signal and how words manifest themselves in the signal. To be efficient, SPM models information from the input speech patterns. A speech signal will first go through a pre-processing stage, which converts the speech signal into a series of vectors, which are then modelled and stored. Finally when an unknown utterance is captured it is then compared with the stored models. The pre-processing stage is necessary to convert the signal to a more explicit form, but it is the modelling stage that is far more complex.

Simple single word models do exist, but there are many variabilities in speech and as two spoken word signals are rarely identical, the modelling must take into account these variabilities. This is possible using stochastic modelling processes, where the variabilities of time and frequency in speech signals are accommodated. Markov models are responsible for dealing with the variabilities of time, as it is rare that a speaker can repeatedly produce speech patterns of the same length. Each vector in the speech pattern is regarded as the output from a Markov process, and each of these vectors is associated with a state in a Markov chain. The transitions between states accommodate the evolution of the speech pattern over time, recognising by means of mathematical

probabilistic functions that speakers may repeat or eliminate parts of an utterance, resulting in missing vectors from the pre-processed speech signal.

Variabilities can also occur within the states of the Markov chain. For instance variations in the produced pitch of a word or phoneme may occur. A process known as hidden Markov modelling (HMM) stochastically models or assigns probabilities to each of the transitions and accommodates for variabilities in the frequency of the speech pattern. Because of this probabilistic modelling there is no longer a direct and unique relationship between each state and the vector in the speech pattern, therefore the modelling process is described as 'hidden' (Bristow, 1986).

As mentioned, the complexities of the development of speech recognisers arise from a number of speaker variabilities. Pronunciation differences will be evident in all languages, as will accent differences and the use of dialects. The speaker's emotional state can also significantly vary their speech pattern; and finally changes in the environment in which the speaker is situated can also cause acoustic variabilites (Zue & Cole, 1995). To deal with these complexities in developing interactive speech systems Bernsen et al. (1998) suggests designing interactive speech systems to focus on real users, real-time and real tasks. In doing so the differences between and within user's speech can be accounted for, leading to more robust and efficient communicative systems. Such systems may thus demonstrate Turing test capabilities, where the user is engaged with the computer in a believable, human-like situation. As testament to the development of such spoken language systems, there is now an annual award, called the Loebner prize, awarded to the interface that best fulfils those Turing test capabilities.

Bernsen et al (1998) propose that through speech interaction theory "a complete and applied theory of spoken human-machine interaction would rigorously support efficient interactive speech system development from initial requirements capture through to the test and maintenance phases". Using this interactive speech theory, the elements to describe the structure, contents and dynamics of the spoken human computer interaction from the point of view of the interactive speech system can be fully understood. On the one hand users need to have a pleasant and natural conversation; on the other hand the theory should support good computational properties and system development. The Danish Dialogue Project (Bernsen, Dybkjær & Dybkjær, 1998) is one of the most comprehensive conceptual models for task-oriented applications, emphasising that context and interactional control must be given attention and the features of dialogue

including flexible turn-taking, back channel feedback, repair mechanisms, anaphora, and ellipsis must be considered. As mentioned, speech variabilities are also responsible for significant differences in speech patterns and these too must be considered. Prosodic performances, enriched input, co-operativity and initiative of the system, and the influence it has on the user's behaviour are all topics that need considerable attention when creating interactive spoken dialogue applications. By modelling each of these elements, the process of speech recognition, converting an acoustic signal captured by a microphone to a set of words, can be fully implemented. The capabilities of speaker independent continuous speech recognition reflect the mathematical and computational developments that have been made in order to understand speech patterns making it possible for a variety of users to interact with computer systems, successfully demonstrating, as Zue & Cole state (1995), that speech recognition is now much more robust, portable, adaptable and dynamic. They report very low word error rates (0.3%) for speaker independent digit recognition and they report high perplexity, large vocabulary, speaker independent spontaneous speech services such as Air Travel Services to have error rates of 3%. The world leaders in developing commercial speech based applications include Nuance Communications, SpeechWorks and Phillips, are the driving forces behind the successful deployment of recognition systems in real world applications. The interactive spoken language interface applications that were constructed for the purpose of the evaluations in this thesis (Chapter 6 and Chapter 7) were created with recognition software commercially available from industry leader, Nuance Communications. The Nuance recognition software offers accuracy at over 96% and was used effectively in this thesis for computer-mediated spoken language interaction.

2.3 The Persona Metaphor

It is the developments in ASR technology, making it more robust and portable that have made it possible for speech technology to become pervasive on desktop computers, mobile telephones and other portable devices. According to Granström (1999) the production of spoken language interfaces for these devices has largely ignored the addition of a visual modality and explorations of multi-modal systems where the voice is given a face have been neglected. In response to attempting to develop the ultimate

communication device between humans and computers the efforts of researchers have recently extended spoken dialogue systems to multi-modal interface applications and research has begun to coalesce around the development of embodied conversational agents. The concept of creating anthropomorphic embodiments on the interface is being pursued in order to give speech a visual identity. It is the unanswered questions with respect to this persona metaphor that are the primary concerns of this thesis and in order to engage users in effective performances with humanoid entities, the thesis aims to form conclusions about ECA through a series of progressive evaluations using observation and interaction techniques.

To understand what an ECA may potentially offer to a user it is necessary to define the goal of software agent research, which is to create systems that can engage and help all types of users in a variety of tasks by having goal directions and by being proactive in their environment. Expanding this definition further, an autonomous agent can be described as an agent in pursuit of its own agenda, where it ultimately has the authority to act on the user's behalf, for the benefit of the user or the system as a whole (Bradshaw, 1997). Adhering to this definition, embodied conversational agents can then be defined as graphical embodiments capable of understanding speech input and generating speech output in order to converse with a user to complete tasks efficiently in particular environments for the benefit of the user.

For Laurel (1990) embodied conversational agents are metaphors with character and it is through this metaphor that attention is specifically drawn to the qualities that form the essential nature of any software agent: responsiveness, competence, accessibility, and the capacity to perform actions on the user's behalf. It is not only through the agent's conversational capabilities that the user is invited to engage and interact with the agent, but also through the agent's embodiment. This can further promote an engaging and efficient interaction through the manipulation of verbal and non-verbal behaviour. Although it is generally understood that speech is the main carrier in face-to-face communication (Sacks, 1992), the visibility of facial expressions (Ekman, 1975), gaze direction (Argyle, 1976) and other non-verbal communicative behaviour regulates conversation in the most natural way possible. Through the process of visual anthropomorphism, users have the ability to exploit the communicative strategies with which they are psychologically adept and they can thus relate to and communicate with the agents comfortably. Nevertheless, Norman (1997) explains that anthropomorphism

is not the same as relating to other people, but it is rather the application of a particular metaphor with all its concomitant selectivity. The metaphor allows for some qualities of the embodiment to be suppressed and others to be emphasised, for use in a particular context. For instance, in this thesis the physical representation of the agents is varied (e.g. 2D head, 3D head). By varying the dimensionality in this way, suppressing aspects of the agent's appearance, it is possible to document varying reactions and attitudes to the agent metaphor.

2.4 Face-to-Face Interaction

When designing conversational agents that use the persona metaphor, designers must not only be concerned with the linguistic and conversational ability of the agent, but if they are to create a believable agent on the screen they must also be concerned with issues of facial expression, gaze direction and other visual behaviours that regulate conversation. For this reason a greater understanding of the complexity of face-to-face interaction is necessary. It is through face-to-face interaction that 'grounding' can be achieved, which is inherent in any communication process (Clark and Brennan, 1990). Grounding can be defined as the understanding of mutual knowledge, mutual belief and also the comfortability of shared information, which is also referred to as mutual attitudes or 'assumptions'. The face, which is the primary tool of non-verbal behaviour, promotes the establishment of a well-grounded conversation. Goffman (Kendon, 1988), an advocate of promoting the now widely accepted theory of face-to-face interaction, explains that the art of conversation involves much more than just spoken words, and that through gestures and gazes, the "external signs of orientation and involvement" contribute significantly to the promotion of a grounded interaction. Through his tireless efforts Goffman firmly established what is frequently called 'interaction order' as a separate unit or field of sociology and in the multidisciplinary field in which this thesis is situated Goffman's sociological viewpoint is used to further the understanding and importance of grounding in conversation.

The theory of interaction order is founded on some essential requirements for an effective face-to-face interaction. Firstly, it is necessary to establish a two-way capability to send and receive information. Non-verbal behaviour of the face and body combined with verbal output can ensure smooth turn-taking, which is the process of

organising the exchange of information between communicating parties in an interaction (Thorisson, 1996) and ensures effective, regulated transition between the utterances of the participants. It is essential in negotiation and clarification (Whittaker & Walker, 1991). Goffman's interaction order also proposes the need for senders and receivers to display signals in confirmation that reception of the information has taken place. This can occur in conversation through the process of back-channel feedback (Yngne, 1970), which includes displays by the listener that the information from the talker has been received. Such signals include nodding and saying "um-hm". To avoid conversational breakdown Goffman also states that it is necessary to send and receive information during interruptions, enabling conversations to be restored effectively. This occurs through combined use of appropriate back-channel feedback and turn taking. It is only through understanding the complexities of face-to-face interaction that it is possible to attempt to create communicative humanoids by modelling the various aspects essential to the grounding process.

2.5 Creating an Agent

Thorisson (1996) created the first embodied conversational agent that for the most part, albeit in a narrow application, could communicate with a human user while being sensitive to the essential requirements of interaction order as set out by Goffman.

> Face-to-face interaction between people is generally effortless and effective. We exchange glances, take turns speaking and make facial and manual gestures to achieve to goals of the dialogue. Endowing computers with such interaction style marks the beginning of a new era in our relationship with machines, one that relies on communication, social convention and dialogue skills (Kendon, 1988).

Thorisson's research examined multi-modal interaction with an embodied agent holistically, and he successfully constructed a computer system where an agent became believable by mimicking the activities of another human. The construction of the ECA was based on a computation model of psychosocial dialogue skills, which included a model of the essential requirements for effective face-to-face interaction. Although extensive hardware was used (Figure 2.10) to create the agent, making it unsuitable for the types of web-based applications which are the concern of this thesis, it did

demonstrate that it was possible for the ECA to engage effectively in conversation with a user and display realistic verbal and non-verbal behaviour to regulate the face-to-face interaction.

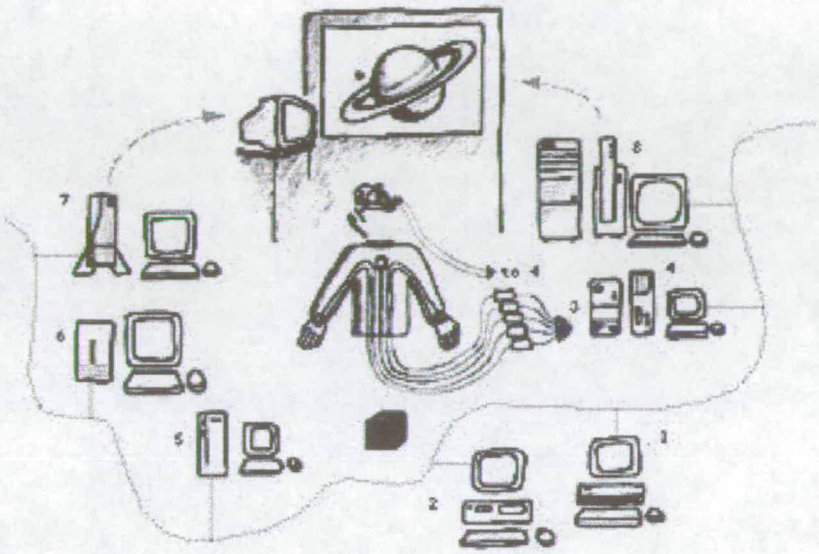


Figure 2.10 Diagram of Hardware Used to Implement 'Gandalf'

The embodied conversational agent named 'Gandalf', illustrated in Figure 2.11 uses speech, gesture and gaze, and is also sensitive to the user's speech, gestures and gaze to sustain believable interactions with the user. The research demonstrated that the face-to-virtual-face interactions obey the rules of human speakers and listeners, however to position generic agents in a variety of application domains further investigation would be required and a reduction in hardware is also imperative, especially for web-based applications.

Figure 2.11 Image of 'Gandalf'

Thorisson demonstrated the importance of adhering to the rules of discourse structure, and showed that the 'Gandalf' agent was capable of displaying appropriate facial expressions during conversation. As Ekman and Friesen (1969) have determined, facial expressions play an essential role in regulating interaction. In fact, facial gesturing plays an essential role when displaying affect, or the expression of emotion. During conversation the expression of emotion is crucial in order to punctuate or emphasise speech. Ekman and Friesen (1978) showed that through mouth and eyebrow movements the generation of facial expressions can impact heavily on the perception of the content of speech. Developing facial expressions and gestures has become hugely important within the research field of face-to-virtual-face conversation, with many researchers concentrating on the autonomous generation and transition between the primary emotional facial expressive states: joy, sadness, anger, fear, disgust and surprise (Massaro, 1998; Magnenat-Thalmann, Kalra & Escher, 1998).

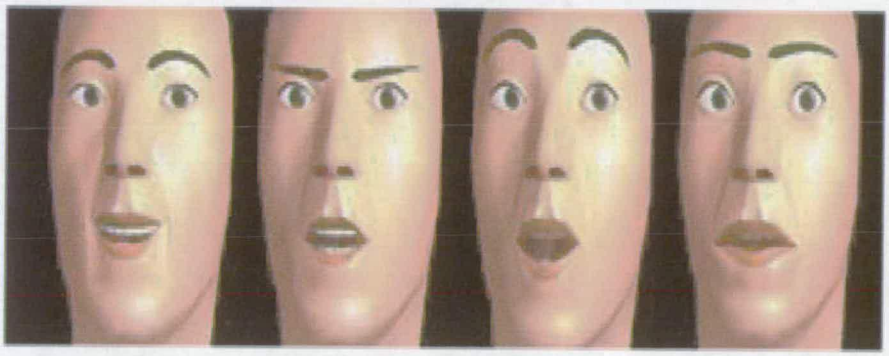


Figure 2.12 Displays of Facial Expressions (Happy, Angry, Surprise, Fear)

Parke & Waters (1996) pioneered research on facial animation, exploring the biological and physiological attributes of the face and thus developed mechanisms to create computational models of the face, which enable the production of lifelike facial expressions. The work by Parke has enabled computer interfaces to be personified visually and the construction of the facial models has made it possible to create expressive faces that can complement and thus enhance the content of the speech output. Other researchers are concentrating on systems to automatically generate the facial expressions during conversations with users. The particular focus of research emphasises the correct matching of phonemes with visemes. This is both necessary and essential for realistic lip synchronisation (Magnenat-Thalmann, Kalra & Escher, 1998). Masarro (1998) similarly demonstrated, using an animated head known as 'Baldi' (Figure 2.12), that mismatches between the synthetic speech phonemes and the facial

visemes can occur, suggesting that it is not sufficient to simply have intelligible speech with animated heads, but that in fact synthetic agents must also demonstrate corresponding intelligible facial movement. The phenomenon of this mismatching between speech and visual content is known as the McGurk effect (1999), where the speech of one syllable, combined with the viseme of another, is perceived as a different syllable altogether (Figure 2.13). In a similar vein, Chapter 4 of this thesis describes empirical research, which analysed user attitudes to different levels of facial expression and lip-synchronisation from agents realised as humanoid photo-realistic heads in retail applications.

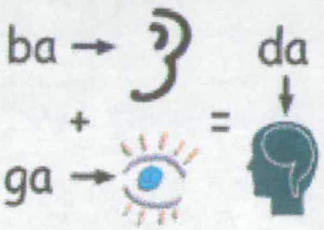


Figure 2.13 McGurk Effect

Parallel to this, the understanding and creation of gesture synthesis and recognition is also immensely important for the creation of believable synthetic agents. What Massaro (1998) has researched with respect to facial expressions, McNeill (1997) is doing for gestures and is working on a comprehensive approach to understanding human discourse through speech, gesture and gaze. Through ethnological techniques, McNeill serves to define the meaning of gesture as being a co-operative of speech. It is through studies like this that computational gesture models, such as that by Kipp (2001), can be constructed in order to generate autonomous expressive gestures for embodied agents. This gesturing can serve to enhance the interaction between the user and the agents. McNeill states the "speech and gesture together can be conceptualised as bringing thinking into existence as modes of cognitive being". On the continuum of progressive empirical evaluations described in this thesis, Chapter 5 documents prior research findings regarding user attitudes to ECA that were represented both with, and without, gesturing and the impact that gesturing may have in retail applications is explored.

2.6 Early Applications for ECA

The progression of empirical research described in this thesis moves from observations of ECA toward interactions with ECA in virtual retail applications and explores the types of applications in which ECA may be effective. There are reasons to believe that ECA may play an effective role as assistants in web-based retail applications. According to Rist (2001), for a company and its customers the presence of an ECA can serve to raise the company's web-based profile by offering 24-hour personalised service. The customised delivery of information and assistance can reduce costs, drive web transactions, encourage future visits and improve customer relationship management (CRM). As there is little empirical evidence yet available documenting information about the presence of ECA in web-based retail applications, this thesis addresses the issues through controlled experiments with potential users. Questions about the appearance of the agent in its role are answered. In addition the functionality of ECA in contrasting retail applications is also examined, addressing the issue of whether the presence of ECA which have been shown to be entertaining entities (André, 1998) are limited to web-based retail applications that have an underlying entertaining theme.

A brief description of the agents that have appeared in educational domains and why they were in fact successful is presented. It is suggested in this thesis that some of the mechanisms that promoted the effectiveness of ECA, for example gesturing in the form of pointing (deictics) may also be transferred to the benefit of virtual retail environments. Lester (1997) explored the effect of embodied agents during interactive teaching experiences and showed that synthetic agents can significantly improve the learning experience for users by engaging them in effective conversations that employed feedback techniques. The agent known as 'Cosmo' (Figure 2.14) was capable of emotive believability, which served to encourage the user very effectively throughout the interaction. In addition, through its deictic believability model, by pointing to relevant areas of the interface the user's attention was drawn to essential information necessary for the learning experience.

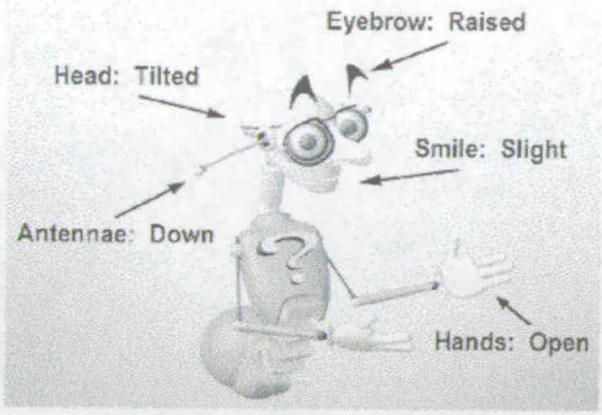


Figure 2.14 'Cosmo' Posture

Similarly using 'Baldi', Massaro (1998) demonstrated that the learning experience for the hearing impaired and for those learning a foreign language was significantly improved through the use of ECA. The 'Baldi' system incorporates speech recognition, natural language understanding, speech synthesis and facial animation technologies in the CSLU toolkit (Cole, 1999). It provides a comprehensive, powerful and flexible environment for building interactive language systems (Figure 2.15). The agent provided a focal point to which the users could direct their attention and through its sophisticated facial expressions and lip-synchronisation ability, sufficiently engaged the user thereby producing significant improvements in the learning experience.

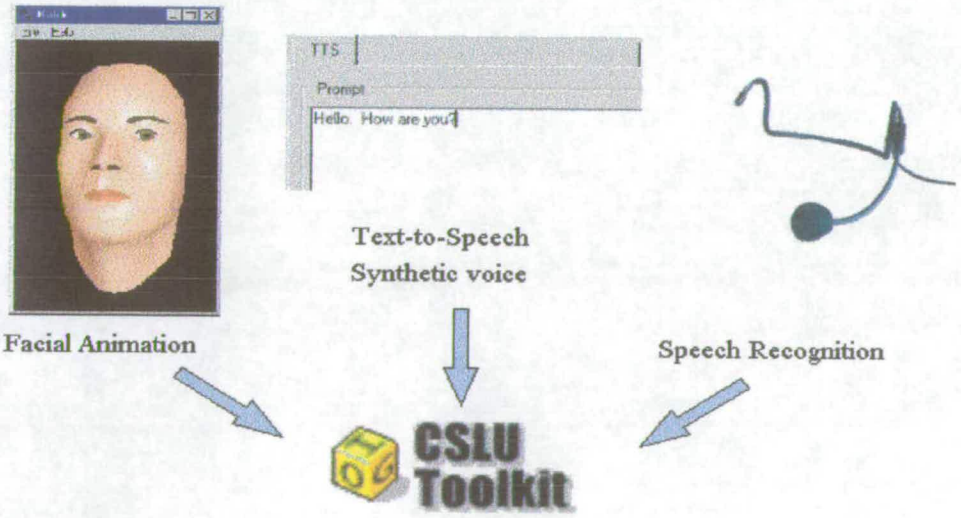


Figure 2.15 The CSLU Toolkit

Continuing in a learning domain, Rickel and Johnson (2000) used an embodied agent, 'Steve' (Figure 2.16) to assist procedural skills training in immersive virtual environments. Through conversation and gesture 'Steve' (Soar Training Expert for Virtual Environments) helps students to learn about new applications. The agent had the ability to move freely through the virtual environment explaining (through demonstration) how machinery functioned.

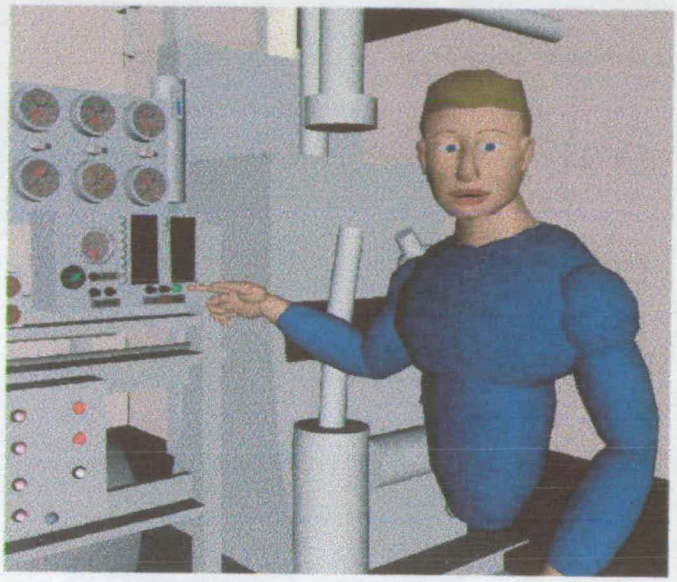


Figure 2.16 Illustration of 'Steve' in Virtual Training Environment

From the earlier work with 'Steve', a number of other guidebots have been created and tested. One example 'Adele' (Agent for Distributed Learning Environments), is a medical tutoring agent who aims to motivate students in a web-based learning environment. Students can take part in interactive exercises, receive effective feedback and also be evaluated and encouraged by the agent in order to learn the skills of patient diagnosis (Figure 2.17). Both 'Steve' and 'Adele' can interact with the student via speech and gesture, and as research on face-to-face communication suggests, the inclusion of non-verbal behaviour into these embodied agents seems to support the learning process (Johnson, 2000, Marsella, Gratch & Rickel, 2001).

Figure 2.17 Illustration of 'Adele'

For Person (2000) the inclusion of appropriate non-verbal behaviour in agents would seem to significantly improve interactions in virtual learning environments. AutoTutor (Figure 2.18) is an animated pedagogical agent that converses with a user to help students learn more about computers. Positive indications are evident that AutoTutor can improve the learning experience for students; however, this is primarily through its sophisticated speech technology and dialogue management techniques. Feedback from the students who use AutoTutor suggests that the introduction of non-verbal behaviour may improve the experience significantly, by providing the user with information and encouragement though feedback mechanisms.

Figure 2.18 The AutoTutor Agent

In support of the effectiveness of including non-verbal behaviour in animated agents a study by Cassell and Thorrison (1999) introduced the idea of providing different levels of feedback during an interaction. Using the 'Gandalf' agent, users were invited to

converse in order to learn more about the solar system. The agent was rated significantly higher with respect to helpfulness ($p < 0.01$) [1]; see Section 3.5) for an explanation of significance figures) when it displayed non-verbal feedback which was sensitive to discourse structure and the requirements of interaction order as per Goffman (Kendon, 1988), which are necessary to regulate the conversation. This type of feedback was termed 'envelope feedback'. In contrast to this when 'Gandalf' displayed just emotional feedback in the form of smiling and frowning, no significant differences for helpfulness emerged, suggesting that envelope feedback is much more effective in a conversation.

Embodied conversational agents have also been effective as presentation agents (André & Rist, 1998) in types of learning environment applications that make a transition from traditional knowledge-based presentation of information to automatically generated multimedia presentations enhanced with animated agents. In favour of the use of presentation agents within a multimedia presentation system, André argues (1999) that the user's attention can be effectively directed to relevant information through the agent's deictic ability. In addition, the agent can act as a guide through the presentation and contribute an expressive power to a system's presentation skills. The images in Figure 2.19 illustrate the effect that the agent may have on guiding users attention during the description of an electronic circuit using a pointing device.

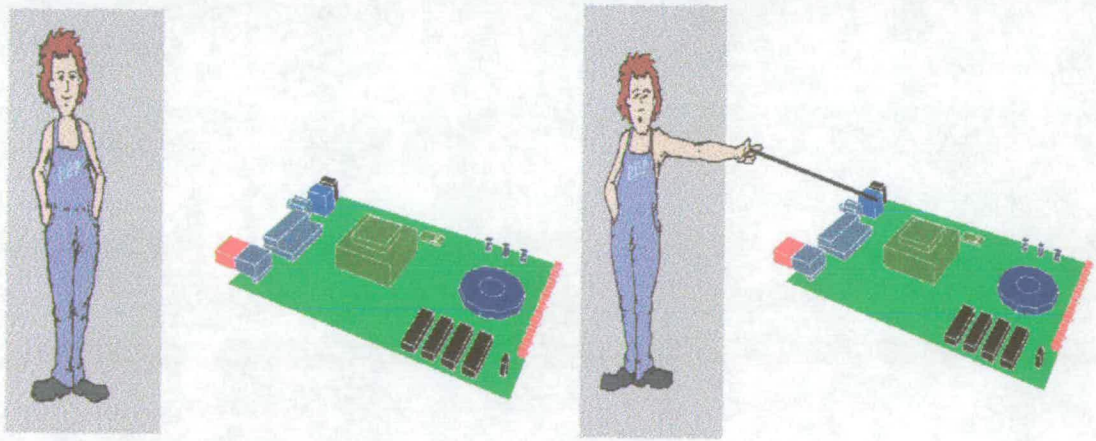


Figure 2.19 PPP-Persona Displaying Deictic Ability

---

[1] P is known as a significance figure, used to indicate the probability of a result occurring by chance. At $p < 0.01$ there is a 1% probability that the result occurred due to chance. More detailed information on significance figures is provided in Section 3.5.

2.7 ECA in Virtual Retail Applications

Research shows that embodied conversational agents appear to have a place as pedagogical aids in interactive learning environments and as stated previously this thesis extends the persona metaphor to interactive retail environments on a quest for evidence to support the use of ECA in domains other than educational scenarios. Allbeck (2001) explains that a character in a virtual world "must meet the expectations of the role it is playing" and this should be reflected in the agent's cognitive model. Accentuating the importance of role, Isbister (1998) found that the role of the agent must be clear to constrain the actions users will take in their corresponding roles. In an effort to analyse the role of ECA in retail applications, this thesis contributes empirical evidence regarding the impact of ECA in contrasting retail applications.

Embodied conversational agents appearing in retail applications are commonly called Bots, for example there are Guidebots (Johnson, 2000), Chatbots and ShopBots (Rist, 2001). Bots intend to promote the collaborative use of speech and gesture recognition in virtual environments in leading the senses to be stimulated to promote more successful and usable retail interfaces. However, the actual appearance of the agents needs to be considered. According to Allbeck (2001) the user perception of the agent's appearance plays a significant part in the perceived role of the agent and since the determination of the most appropriate representation of an ECA to enhance retail applications has been neglected in empirical studies to date, it is of concern in the investigation reported in Chapter 6.

There is evidence to suggest that there are an increased number of websites using ECA technology, primarily for the benefits outlined above. 'Ananova' was the first virtual newscaster and represented the introduction of embodied agents to the web (Figure 2.20). 'Ananova' presents news features to the user on a personalised basis and although the character is not yet capable of interacting with users through speech, its success demonstrates the possibilities for the deployment of agents in retail applications. 'Ananova' can now also appear on users' mobile WAP (Wireless Application Protocol) telephones and their portable palm top computers. Such agents provide the opportunity for companies to deliver information, help and assistance to meet the particular needs and preferences of customers. The success of 'Ananova' since its launch in 2000 is a testament to the use of embodied agents in interfaces and as Rist (2001) explains a

company's online visibility could be increased and the potential to develop branded characters may extend the corporate identity of a company's website, thus differentiating it from others.

Figure 2.20 Illustration of 'Ananova'

Agents with speech recognition capabilities can engage the user in conversational interaction and through the use of social cues, the on-line shopping experience can be enhanced. Additionally, the capacity for the agent to form user profiles increases the level of personalisation, which can be manipulated through dialogue (Bickmore & Cassell, 2000) and this familiarity may improve the perception of customer service, leading to stronger customer retention. As mentioned previously the advantages of ECA technology in a website include a possible reduction in costs for the company, as the agent will have the ability to respond instantly to customer queries, driving transactions and increasing sales. In support of this Forrester Research state that a typical customer call handled by a live agent costs about $33. The same call handled in a live chat-room environment is $7.80. But if a bot can find the answer from the company knowledge base and automatically deliver it to the caller on-screen, the cost is $1.17. This point is accentuated by Arafa et al (2000) who showed that the bots can drive transactions and increase sales by responding to queries instantly and online. Such desktop agents are being deployed as marketing tools or assistant presenters of virtual information. For example the Microsoft Agent® supports the presentation of interactive animated characters within the Windows environment. The interactive assistants can be used to guide, entertain, and enhance web pages or other applications. The agents manipulate the concept of the conversational interface in an attempt to promote natural human social communication. There are a variety of characters that the user can chose from, including some humanoid agents, as shown Figure 2.21.

Figure 2.21 Selection of Humanoid Microsoft Agents

2.8 Agent Technologies

The reasons for suggesting that ECA are attractive devices for retail applications have been explained previously. This thesis addresses the issue of how the agent is perceived in its role as an interactive assistant, but for the agent to actually appear on a website the supporting technology must be in place. Aside from how the agent might be displayed on a website, it is also necessary to consider how the agent is actually generated. Of course it depends on what type of agent is actually required and although the interactive agents assessed in this thesis were created primarily using Virtual Reality Modelling Language (VRML), there are a number of options available to the designer.

For the representation of the agent as a cartoon, traditional animation techniques can be used effectively by converting sketches of the character profile to electronic format. Bates (1991) recommends that for the creation of believable cartoons, the designer must focus heavily on the artistic perspective of animation. By documenting the relevant gestures the agent can then display them from a pre-defined list. Software packages such as Macromedia Director can be used effectively and as Rist (2001) explains, this package is successful as it draws on core concepts from filmmaking, including cast, stage and score. The cast is the collection of media objects (e.g. audio, video), the stage is the window for the display of the media contents and the score specifies the temporal and spatial display of the media objects on the stage. The score serves to turn a script into an interactive presentation. In this thesis Macromedia Director was used to create the interface presentations of humanoid photo-realistic and animated agents that were evaluated in Chapter 4 and Chapter 5. To create 3D versions of agents, the software package 3D Studio Max was used. Characters created using this package can be

examined in Chapter 5. The package is effective as it offers a rich set of operations for creating and manipulating wireframe models. Its versatility allows a designer to create, model, texture and render 3D animations in whatever context is most convenient for an application scene.

To create human-like faces, either video recording can be used, or motion-capture methods of a real person. Although motion capture generates effective fluid movements and realistic motion, promoting the believability of the agent, this type of gesture generation is not entirely suited for real-time interaction or web-based autonomous agent implementation. The most up-to-date method for generating faces is to use MPEG-4 Face Animation Parameters (FAP). The object-based multimedia standard allows different audio-visual objects to be encoded independently. The face is parameterised, where each parameter corresponds to a particular facial action.

Another international standard, this time to create full-bodied animation, is known as the Humanoid Animation Specification (H-Anim). The H-Anim specification was created in the light of increasing need to represent human beings in online virtual environments, either as avatars or agents and to achieve that goal, libraries of interchangeable humanoids needed to be created (Humanoid Animation, 2001). H-Anim specifies a standard way of representing humanoids and allows humanoids created using authoring tools from one software package to be animated using tools from another. H-Anim humanoids can be animated using key framing, inverse kinematics and performance animation systems. In Chapter 6 and 7 of this thesis the H-Anim specification was used to create a range of 3D ECA that appeared as assistants in a range of virtual retail environments.

There also exist a selection of player technologies available to actually host the agents on the web. One such technology is Synchronised Multimedia Integration Language (SMIL), recommended by the World Wide Web Consortium (Web Consortium, 2001). This XML-based language allows the designer to write interactive multimedia presentations, which can also be embedded in JavaScript. SMIL enables simple authoring of interactive audiovisual presentations and is used for 'rich media' and multimedia presentations, which integrate streaming audio and video with images, text or any other media type. It is an easy-to-learn HTML-like language, and can be written using a simple text-editor. Animated agents created using SMIL are more suitable to applications where the agent acts as a presenter of information. Although there is often

no extra installation cost and the animations can be streamed to the user, the large audio data volumes make this technology less suitable for conversational interaction.

The use of the Flash Macromedia Player also allows designers to arrange a set of animation clips in a non-linear structure. It is possible to use Flash editing tools effectively, combining FlashMovie and FlashScript, Shockwave and Director, the multimedia authoring tool, to combine graphics, sound, animation, text and audio to create character animation clips and audio output that can be transmitted in compact form to the client. As with SMIL this technology is effective for creating well-animated, believable embodied agents, but is also less suitable for actual conversational interaction.

Virtual Reality Modelling Language (VRML) is another technology that can be used to create believable agents that can be immersed in desktop environments. By embedding a VRML player into a website, using Java, JavaScript or C++ an embodied conversational agent can comfortably inhabit the environment. VRML, which is soon to be succeeded by the X3D language is a platform independent definition of 3D spaces on websites and is an ideal technology for creating 3D interactive agents that can inhabit 3D virtual environments. It is designed to combine the best features of virtual reality, networked visualisation and the global hypermedia environment of the Web. In association with the H-Anim Specification, VRML allows the creation of believable lifelike humanoid agents. This specification is a result of research into creating natural physical humanoid agents where Badler et al (2000) demonstrated that through the study of human movements a better understanding of how agents move can be developed.

Rist states that there is no single technology that designers must use, but it is necessary for the designer to select a technology that is best for the application and also for the client-side costs (Rist, 2001). Throughout the thesis a variety of technologies were used to create both the characters and the virtual worlds in which the agents appeared. Macromedia Director and 3D Studio Max were used to create the agents that are evaluated in Chapter 4 and Chapter 5. Chapter 6 marks a transition to VRML technology. This language was more suitable to create the 3D virtual agents and the virtual retail environments in which they appeared. VRML is accessible technology and "is advanced enough to be capture the excitement and imagination of developers, and yet it is simple enough to be both practical and accessible".

A user's mental representation of an environment can vary but the advantage of VRML is that the reality can be brought to the desktop such that a users' knowledge and understanding of a situation can in some ways be consistent with their perceptions of reality. Simulated and virtual realities stimulate the elements of perception and "common to these definitions and conceptions is that they are all concerned with the stimulation of human perceptual experience to create an impression of something which is not really there" (Carr & England, 1995). Virtual environments on the desktop now connect the user more effectively allowing them to be sensitive to three dimensions. By describing the content, the geometry and the dynamics of the world it is possible to take advantage of the remarkable human cognitive ability to make inferences from minimal information.

Virtual places and social worlds will soon be frequented by many more users, and Cuddihy and Walters (2000) state that with the availability of inexpensive 3D graphics cards the rise in power of desktop computers, displays of real-time interactive 3D graphics has become a reality for most computer users. Using VRML a sense of presence can be created for the user, which is the subjective experience of being in one place or environment, even when there are physically situated in another.

> Techniques for increasing a user's sense of presence are important because they enable the technology to draw the user into the virtual environment, suspend their disbelief and engage their attention. This is obviously of interest to the entertainment industry who in turn have interest in social virtual environments, but increased presence is likely to also benefit other kinds of social scenarios (Carr & England, 1995).

2.9 Socially Intelligent Agents

It is this sense of presence that has been at the forefront of recent research within the HCI field, which has fostered the creation of socially intelligent agents. According to Norman (1997) the foreseeable difficulties with creating agents are social and not technical. These new forms of automata are different, as they interact with people in human-like ways and so to advance the research much more investigation is needed into user perceptions of these social entities. Landmark research by Reeves and Nass (1996) showed that humans in general treat computers in a social manner, and this has

encouraged the development of social interfaces. However the existing research and development on embodied conversational agents was further fuelled by the desire to enhance these agents, improving interactions and therefore creating socially intelligent agents (Dautenhahn, 2001). The fact that users assign social cues to interfaces laid the foundations for a research programme that not only has an end goal of a conversational interface, embodied by an animated character, but also that of an agent that exhibits social awareness. To be effective these agents must be both emotive and adaptive to users' emotional states and it is thus through this social interaction that users can collaborate with agents to complete tasks, but furthering previous research on conversational characters, a relationship could develop between the agent and the user, which could serve to encourage future interactions. Though empirical research, Chapter 6 and Chapter 7 of this thesis present new findings about the importance of establishing social relationships with agents in contrasting applications, and move toward research strategies that improve the perceived trustworthiness of agents in applications, in order to encourage further social interactions leading to effective social relationships.

Socially intelligent agents research focuses heavily on an area within the field of HCI known as Affective Computing. As Picard (1997) states "the latest scientific findings indicate that emotions play an essential role in decision making, perception, learning and they influence the very mechanisms or rational thinking". To create truly intelligent computers, they must have the ability to display and recognize emotions and so it also becomes essential for embodied agents to display such emotions for them to be perceived as believable, socially intelligent agents. In a similar vein, Dautenhahn (2001) states that much of the work on creating socially intelligent agents "is strongly inspired by forms of the natural social intelligence characteristic of humans" and the research field of socially intelligent agents concentrates on supporting interactions among humans and encourages the development of models that explicitly show aspects of human-style intelligence. Prendinger (2001) explains that consideration for social dimensions in animated agents adds value as believability increases the illusion of life, which is often captured through reasoning about emotion and personality. Secondly, Prendinger (2001) states that consideration must be given to social dimensions, which play an important role in human conversation adding social robustness to the interaction.

Role-playing with embodied agents in interactive drama scenarios became an effective method to begin to measure the social expectations from users during interactions. To

create effective dramas the agents have be able to display autonomous emotive responses and they must also be reactive to users' varying emotive input (Marsella, Gratch and Rickel, 2000). Preliminary results from Prendinger (2001) indicate that animated agents with social competence and affective behaviour are considered to be interesting training companions. Another system, developed by Gratch and Marsella (2001), is being used to generate a computational model to support characters that act in virtual environments, but whose decisions form an underlying emotional current. They have demonstrated, although not yet through empirical evaluations, that such agents can significantly contribute to learning through experience by developing a social emotive conceptual model. The emotive model is developed using a Mission Rehearsal Environment where lieutenants can interact with civilians (i.e. autonomous embodied conversational agents) to defuse situations were there exists high anxiety and emotional states (Figure 2.22). The experience serves as a learning and training exercise for the lieutenants where they can experiment with various strategies and learn from the reactive behaviour from the ECA.

Figure 2.22 Autonomous Agents in Mission Rehearsal Environment

Emotions are needed in SIA technology to convey intentionality, to elicit emotions in others and to better understand communication. To elicit the emotions, it is necessary for the agent to have personality, which is an important factor in the creation of socially intelligent agents. For this reason it is important to perceive the personality of the agent and to ensure that it is the correct personality for the application domain in which the

agents appear (Cañamero, 2000). Magnenat-Thalmann concentrates on the evolution of the emotional state of the agent (Magnenat-Thalmann, 2000). The work defines and models the personality of an emotional autonomous agent to help the evolution of an emotional state through dialogue. The definition of personality is made in terms of the probability of a person being in a particular emotional state. For instance momentary emotions, such as a smile exist only for a few seconds, but moods are manifested due to the prolonged cumulative effect of momentary emotions (e.g. smiling may imply a happy mood). Personality, which is a larger subtle trait, can be represented by a matrix of the probability of the transition from one emotional state to another in a specific hierarchy of emotions and this can be displayed visually using the facial expressions defined earlier. The smooth transition between emotions and thus facial expressions can occur using probabilistic methods, maintaining the naturalness and believability of the agent. It is important to portray personality and emotion to promote interest and empathy in order to appear more natural (Trappl & Petta, 1997). But this will also depend on the personality of the user, the application domain and the context of the domain. In this thesis, the perceived personalities of the agents, who appear as assistants in the virtual retail application are investigated. The empirical evidence assists in forming conclusions about the expected personalities of agents who play the role of assistants. Understanding more about personality will lead to the development of a social interaction (Magnenat-Thalmann, 2001).

From an alternative perspective, the REA (Real Estate Agent) platform monitors the emotional state of the user in an attempt to create more affective dialogue (Bickmore & Cassell, 2000). The notion of small-talk between an embodied conversational agent and the user is serving as an important strategy to establish a comfortable social interface. Analysing the amount of information a user is willing to disclose can serve as an effective strategy to determine how comfortable the user is in an interaction. The agent then displays the capability to respond to the user appropriately. The research raises interesting issues about building a rapport with the agent in order to infer trust to the agent. This is essential in situations where the user may have to disclose personal financial information. The importance of small-talk can be explained as a ritualised way for people to move into conversation in what may be an otherwise awkward or confusing situation. Using non-verbal communicative behaviour to maintain the conversation, the agent through its verbal ability also uses small-talk as an exploratory function to establish its capabilities and credentials and also disclose its expertise by relating stories

of past successful problem solving. By showing appreciation of users' utterances it is possible to build solidarity with users. Preliminary research indicates that the use of small-talk during conversations between a user and an embodied conversational agent is an effective strategy to establish a helpful social relationship.

Figure 2.23 Embodied Agent 'REA' in Virtual Environment

Arafa & Mamdani (2000) state that "an animated figure, eliciting quasi-human capabilities may add an expressive dimension to the agent's communicative features, which can add to the effectiveness and personalisation of the interface and the interactive experience on the whole". The work is conducted in the area of communicative agents, in particular animated agents as web enhancements for retail sites. In the form of embodied conversational characters, or electronic personal sales assistants, Arafa et al. (2000) explore a new paradigm for eCommerce in the form of presentation and personalisation and explains that in the service-based electronic market the focus of attention falls on personalised service and more personal customer-related management. It is explained that the customer's affective state is partly responsible for the customer's mental state and behaviour in particular applications. To understand and adapt to the user's affective state, Arafa et al. (2000) propose forming a social relationship during an interaction. To utilise the persona metaphor the human anatomy is used and the agent is divided into head, heart and body, each with specific models to represent actions and behaviours designed to promote a social personalised interaction.

- Head: deals with perception, continuous memory, reasoning, memory
- Heart: maintains and manipulates the affect of emotion and personality
- Body: deals with behaviour, action execution and visual representation

The authors believe that for more widespread focus on considerations for individualized social behaviour, the persona metaphor could be effective. Arafa et al also investigate the topic of building customer loyalty in eCommerce applications and again propose the use of animated ECA as tools to form this essential loyalty aspect. Specifically they are beginning to investigate the perception of loyalty in ways in which the customers are encouraged to return to a service that is promoted by the presence of an ECA.

2.10 User Evaluations

Despite the progress being made with respect to speech technology, multimedia and computation models to essentially provide cognitive skills for ECA, there is little evidence to support their effectiveness in retail applications. According to Cohen and Oviatt (1995):

> An important but under appreciated requirement to the successful deployment of spoken language technology is the development of empirically validated guidelines for creating interfaces that incorporate spoken language.

Empirical research is necessary in order to assist designers with how the agent should be represented in the interface. For the development and widespread implementation of ECA, Rist (2001), Cassell (1999) and Pelachaud and Poggi (2001) recommend that much more empirical research needs to be conducted. Such evaluations will lead to a greater understanding and acceptance of ECA technology, thus providing the motivation to create effective agents. Pelachaud and Poggi (2001) explain that empirical research needs to be completed to understand more clearly the attitudes of the human user. Using empirical findings, models can be created to develop appropriate agents. It is the need for empirical evaluations that has motivated the evaluations of embodied conversational agents described in this thesis. As Dehn and Van Mulken (2000) explain it is difficult to draw comparisons and general conclusions between the existing empirical studies on

embodied conversational agents as the "empirical investigations on the effect of animated agents are still small in number and differ with regard to the measured effects". Laurel (1990) recommends rapid prototyping techniques to facilitate user testing and evaluation.

> If we can continue to gather feedback from individual users and inspiration from popular culture as a whole, then the notion of agents will evolve, as it should, in collaboration with the people from whose fantasies it arose.

Dehn and van Mulken (2000) attempt to highlight and explain the reasons for inconsistent findings between varying evaluations with regard to the user's subjective view of agents in interfaces and concludes that it is futile to attempt to draw general conclusions. Although Lester demonstrated using 'Cosmo' (Figure 2.14) that there exists a "Persona Effect" in an educational environment, and that the presence of a lifelike character in an interactive learning environment can have a strong positive effect on students' perception of their learning environment, it is not known if the persona metaphor can be extended effectively into the retail domain.

In another evaluation using the PPP-Persona (Figure 2.19) the question of presence was addressed. The presentation of technical information was significantly improved ($p < 0.01$) by the presence of an agent. In further support of a persona effect, Koda showed that personified faces can help users engage in a task (Koda, 1996). Studies by Walker et al. and Takeuchi and Nagao also concluded that having a face in the interface was more engaging (Walker, Sproull & Subramani, 1994; Takeuchi & Nagao, 1993).

In Koda's study participants were invited to play in a card game against agents who were represented as male and female, photo-realistic and caricature, animate and inanimate, embodied and disembodied. The results showed that, based on appearance, a photo-realistic humanoid image was perceived as being more intelligent than a caricature face and caricature dog, ($p < 0.01$). However, after playing poker with these agents participants then rated the perceived intelligence on the agent's display of competence during the interaction and less so on the appearance of the agent.

In a study by King and Ohya (1996) significant differences were found when participants were asked to observe and then rate a variety of anthropomorphic agents, which included the rating of intelligence of geometric shapes, caricatures and 3D photo-realistic heads.

Participants rated the human agents to be more intelligent than the caricature faces and the geometric shapes. The fact that the participants did not actually interact with the agents makes it difficult to draw further comparisons between the two studies and although it may seem trivial to state that humans will perceive humanoid representations as being more intelligent that geometric shapes, Koda's results show that it is important to assess agents with the context of a domain application, and not solely on isolated observation.

Koda also found that having an agent visible on the screen was significantly more engaging and comfortable ($p < 0.01$) than one without in the context of the poker game. However, as stated above, participants rated the intelligence of the agent on its display of competence and not appearance, therefore no significant differences for perceived intelligence were discovered between having a face on the screen and not having a face.

In contrast to the results by Koda, an empirical result by Sproull et al (1996) contradicts the finding that having a face in the interface is more engaging and comfortable, when it was discovered the participants in a career-counselling service considered a humanoid agent to be less attractive and friendly than a text-based version of the system. Again it is difficult to interpret the contrasting results due firstly to differences in the nature of the task, and also due to the fact that the participants in Sproull's study also had to consider the text-to-speech voice output of the agent, which may have influenced the rating. The few empirical evaluations described above tend to suggest that a persona effect may exist, but this is highly dependent on the application context and the representation of the agent in that particular application.

2.11 Summary

This chapter provided an overview of the history of speech technology research, dating back to the 18th century. A description of early talking machines was provided. The foundations of the digital era then promoted the development of speech recognition software and spoken language interfaces. The persona metaphor, or the exploration of anthropomorphising the interface was then discussed in the form of embodied conversational agents. The effectiveness of ECA in earlier applications such as the

educational domain was explored, followed by reasoning to consider that ECA technology may also be effective in retail domains.

A description of technologies used to create believable ECA was provided, together with the player technologies available that have the capabilities of hosting ECA in web-based applications. The merits of using VRML as a player technology were discussed and the implications of experiencing a sense of presence were introduced, followed by a discussion toward the establishment a social relationships. Following this, the call for empirical research was amplified by an attempt to draw conclusions between the few empirical studies on embodied conversational agents.

Through empirical research it will be possible to determine the retail applications in which embodied conversational agents are effective and for them to then sustain this effectiveness determine how they should be represented. Instead of simply adding to the existing empirical evaluations, this thesis consciously focuses on conducting structured progressive experiments, beginning with observational techniques and progressing to evaluations where participants were invited to engage with the agent interactively. The spectrum of experiments allows the results of each to be effectively compared. The work addresses both the representation of ECA and the effectiveness of the agents in retail domains.

Chapter 3

Description of Experiment Methods Used to Empirically Evaluate Embodied Conversational Agents in Retail Applications

3.1 Introduction

This thesis aims to produce a body of empirical evidence detailing the effectiveness of embodied conversational agents in retail applications. The experiments designed and completed in order to gather this empirical information rely heavily on input from the participants in the experiments; the potential users of the conversational applications. To retrieve the required information about the functionality of systems, the series of experiments documented in the thesis were designed with particular attention being paid to the impact of the presence of the user at the early developmental stages. This approach is otherwise known as usability engineering, ISO 9241. In support of the effectiveness of this approach Norman (1998) has explained that by understanding the user's motives, interests and needs, it becomes possible to improve the functionality of a system. In essence, by collaborating with the design engineer the user can positively influence the development of the product or system. Augmenting this Gould states: "you must at the very beginning and throughout development focus on prospective users" (Gould, 1988). In this thesis, the usability of retail applications inhabited with the ECA was monitored, the effectiveness of the interaction between the user and the agent was analysed, the efficiency with which a user could complete a task using the interactive retail systems was observed, and user's satisfaction with the interaction was documented in order to produce a body of empirical data with respect to the use of ECA in retail applications.

In total a progression of four empirical studies are presented. The earlier experiments documented in Chapter 4 and Chapter 5 gathered information relating to the representations of humanoid agents, through observation techniques. Participants were invited to observe retail interfaces in which a number of agents appeared. In the evolution of this empirical research, the later stages of the thesis make a transition from evaluation by observation to evaluation by interaction. To successfully conduct these

interactive experiments a functional interactive interface had to be designed and created as documented in Chapter 6, together with the results of the first interactive experiment, focusing on the representation of embodied conversational agents in contrasting virtual retail applications. In Chapter 7, the series of empirical studies concludes with a second interactive experiment, which assesses attitudes toward the inclusion of multi-modal communicative strategies in the virtual retail interfaces in order to communicate more comfortably with the agents. In order to create effective experiment interfaces, for either method of empirical investigation described in the thesis (observation or interaction), it was essential to test the functionality of the interface prior to designing experiments. These testing techniques are useful to determine that a system has the capability to allow the successful completion of experiment tasks. To guarantee that the experiment interfaces were correctly designed it is effective to employ certain test procedures.

3.2 Test Procedures

Specific techniques have become popular for testing and monitoring interfaces and many of these techniques were used at various stages during the design of the experiment platforms in this thesis. A discussion of the procedures that were used is now presented, detailing both the advantages and disadvantages of each to provide clear reasoning why these methods were completed at various stages of the experiment construction. As mentioned two types of experiment were designed, one specifically for evaluation by observation, and the other for evaluation by interaction. It was essential to test the functionality of both types of interface to ensure that they were suitable systems from which to empirically assess and evaluate aspects of the embodied conversational agents that inhabited the interfaces. The goals of the tests were to firstly assess the functionality of the experiment interface, secondly to assess the effect of the interface may have on the user, and finally to identify any problems with the interface before using it in controlled empirical investigations (Dix, 1998).

It is important that the system has the necessary functionality for the user to complete required tasks. The functionality of the system must support ease of use, thus avoiding user frustration during the interaction, also ensuring the opinions and attitudes of the participants focus on the experiment aims. Participants must have adequate time during tasks to form opinions in order to complete attitude questionnaires and take part in the

interviews in the most informative way possible. Finally, it is essential design practice to identify any problems and their causes in order to correct them, before interactions take place. The three major techniques to test the effectiveness, efficiency and satisfaction of a human computer interaction scenario are non-expert evaluation, expert heuristic evaluation and cognitive walkthrough. Table 3.1 lists the advantages and disadvantages of the three test approaches used in this thesis, highlighting the complementary nature of these techniques.

According to Molich & Nielsen (1990) non-expert testing is an approach used to equip software developers with a set of design guidelines that they can refer to throughout the design process. This approach was used effectively in the design of the two types of experiment interfaces used for the experiments in this thesis. The guidelines set out by non-expert test make sure that the designer is aware through the experiment interface design that the systems are designed to be consistent, to allow the user to complete tasks with minimal error and to provide appropriate exit procedures should the task be uncompleted.

The second test procedure used involves experts in the field of user interface design, who were invited to critique and assess the retail interfaces. The advantages are that these specialists could easily identify usability problems with the interfaces. The most common problem with this procedure is that availability of experts is often limited. Fortunately this did not arise during the evaluations of the experiment interfaces described in this thesis and expert evaluation was conducted successfully.

The cognitive walkthrough approach, which introduces psychological theory into the evaluation, was also used. This was particularly the case with respect to evaluating each of the four experiment designs reported in this thesis. Cognitive psychologists were invited to evaluate the experiment design in order to investigate how well users could perform required experiment tasks. The expert considers at every stage:

1. The impact the interaction has on the user
2. The cognitive processes required
3. Learning problems that may occur

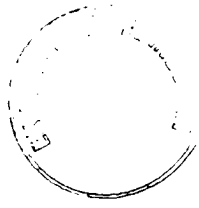
	Advantages	Disadvantages
Non-Expert Heuristic Testing	Identifies general problems Identifies recurring problems Can be completed at any stage in the design process Can be completed by software developers	Severe problems can be missed, even using the criteria as a guide Many evaluators necessary User knowledge and capabilities required
Expert Heuristic Testing	More problems identified More serious problems identified Effective solutions suggested Compliance with recognised usability guidelines checked	User interface expertise required Several evaluators necessary Only applicable at the latter stages of design process
Cognitive Walkthrough	User goals defined clearly Can be used at the early design stage or after system developed	User interface expertise required Several evaluators necessary Best results emerge at the early stages of design process

Table 3.1 Advantages and Disadvantages of Test Methods

3.3 Design of Experiments

As mentioned previously the experiment design was evaluated before participants were invited into the research environment. Prior to this evaluation it is necessary to correctly design the controlled laboratory experiments from which the relevant quantitative and qualitative information regarding participants' experiences can be gathered. The design process of such experiments can be complex, but to assist with this the experiments discussed in this thesis were designed using methods commonly employed in applied experimental psychology. The design process that was used for each of the four usability evaluations closely followed the steps involved with planning and designing an applied experiment, as outlined by Monk (1985).

1. Identify the topic to be investigated
2. Gather information about the topic
3. Specify the aim of the experiment
4. Specify the hypotheses
5. Define the variables



6. Design the experiment

7. Set realistic tasks

Each chapter describes, identifies and documents clearly the topics and issues that are investigated during the course of experiments. Following this, each chapter contains a literature review documenting related research relevant to the experiment. This information helps to clearly state the nature of the experiment, why there is a need for the experiment to take place and how the results of the experiments contribute to knowledge. This information essentially shapes the hypotheses and objectives of the experiment.

These hypotheses provide a claim as to the outcome of the experiment. Basic experiment design depends on distinguishing independent variables from the dependent variables. Independent variables are manipulated variables and dependent variables are measured. The experiments described in this thesis often had more than one independent variable and each of these variables also had a number of levels. For instance an independent variable could be the number of ECA types evaluated in a particular experiment and the level of that variable could for instance depend on the agent gender, where male and female versions of each agent type are included in the evaluation. It is possible to examine for main effects of these variables and also any interactions that may occur between the variables. The dependent variables in all the experiments were the responses to the individual statements in the questionnaires and the responses given during the interviews.

An experiment claim states that a variation in the independent variables will cause a difference in the dependent variables and it is the aim of the experiments to show that the prediction is correct. By testing the experiment hypotheses against the null hypothesis (i.e. that there exists no difference in the dependent variables between the levels of the independent variables), it is possible to determine if there are significant differences between the varying conditions and that there exists evidence that particular predictions or claims are supported. This approach is evident throughout the experiments detailed in this thesis.

After the hypotheses have been listed, the next stage is to document how these claims will be tested. At this point in each of the chapters documenting the empirical studies,

the experiment procedure is established. Specific issues are established, such as the duration of the experiment; remuneration of participants; experiment apparatus; number of experimenters; number of participants; representation of participants; preference for a between-subject or within-subject design, and how quantitative and qualitative information should be gathered.

It is vital to select a participant sample that will not bias the results, and this sample should represent the user population as closely as possible. The experiments in this thesis attempted to balance the participant sample as accurately as possible for participant age and participant gender. Three participant age groups were used (Group 1 – 18 to 35; Group 2 – 35 to 50; Group 3 – 50+). The experiment tasks did not require the participants to have computer experience (except in Chapter 7). Separate groups of participants took part in each of the experiments and the sample sizes ranged from 32 to 48, depending on the experiment criteria.

In controlled laboratory experiments it is essential to have a reliable experiment method. There are two main types of experiment method: between-groups and within-groups design. The first ensures that each participant is assigned a different condition. The second method, within-groups was used in all the experiments described in this thesis and ensures that each participant is assigned all the conditions. The advantage of the within group method is that there is less likelihood that biased results will arise due to individual differences. In a within-group design it is also essential to eliminate any learning effects. To do this the order of presentation of the conditions must be carefully randomised for each participant. Any effects will thus be counterbalanced. Finally it is essential to set realistic and clear tasks for the participants in the experiments, to ensure they have sufficient time to form opinions about the experiment. For experiment clarity, consistency and increased readability, each of the chapters detailing the experiments adheres to these experiment design guidelines.

3.4 Retrieving Experiment Data

Central to the four experiments documented in this thesis are the methods used to retrieve information in order to investigate the set experiment hypotheses. Combinations of quantitative and qualitative information retrieval methods are used in the experiments

to form conclusions specifically relating to the deployment of embodied conversational agents in retail applications. The primary methods of quantitative information retrieval used for each of the experiments are attitude questionnaires, which are proven methods to gather and measure attributes and characteristic of systems, in order to fully understand 'user-perceived quality' (Dix, 1998). In the four experiments described in this thesis, Likert-type questionnaires were designed and used to measure participants' attitudes toward aspects of embodied conversational agents in order to form conclusions about the representation of the agents in virtual retail applications (Likert, 1932).

3.4.1 Quantitative Analysis

Questionnaires need to be designed carefully so that they address the topics of interest relating to the experiment hypotheses. Although it is important to include questionnaire statements that elicit information concerning all topics of interest it is much more effective to be selective regarding the number of statements as questionnaires that are excessive in length can annoy and frustrate the experiment participant and could lead to inaccurate responses. To avoid this occurring the information desired must be defined precisely. According to Rust (1989), an effective strategy, known as the 'blueprint' method, ensures that all aspects that need to be questioned are included in the questionnaire. This method requires the definition of *content areas* and *manifestations*. The content areas are a list of everything relevant to the purpose of the questionnaire, for instance in Chapter 4, the agent and the applications are the most relevant content areas. Rust explains that an exhaustive list should be created to ensure that all areas that need to be covered by the questionnaires are addressed. The manifestations are described as the way in which the content areas manifest themselves (e.g. functionality, appearance). Cross-referencing each of the content areas with the manifestations and listing the questionnaire statements for each cell in an experiment design, ensures that an exhaustive list of statements relevant to the purpose of the experiment is created. It is not productive to have questionnaires excessive in length and so the next step in questionnaire design is to refine the list to contain a suitable number of statements. The observations of interactions between customers and agents in each of the experiments in this thesis were approximately two minutes in length and therefore the questionnaires were also designed to take approximately two minutes to complete. For this reason each questionnaire contained between fifteen and twenty statements. Throughout this thesis,

this approach of questionnaire design was used as an effective mechanism, albeit a simple guide, to list questionnaire statements for each evaluation.

All the questionnaires described in this thesis used a Likert-type (Likert, 1932) attitude scale, which is a useful method for assessing user attitudes. Coolican (1994) has reported the advantages of using Likert-type questionnaires and has shown that the Likert scaling method:

- is natural to complete,
- maintains the participant's direct involvement,
- has shown a high level of validity,
- has shown a high level of reliability.

When using Likert-type attitude scales a number of issues need to be addressed. Firstly, a balance between positive and negative attitude statements must be obtained to avoid a response acquiescence effect. This effect is caused by respondent's tendency to agree rather than disagree with the attitude statements, regardless of the questionnaire statement content. To ensure that an equal number of statements are scored in each direction it is important to reverse polarity of some statements, being cautious so as not to change the meaning of the original statement. To avoid confusion, it is also advised to avoid double negative statements.

> Acquiescence is less likely to occur with items which are clear, unambiguous and specific. (Rust, 1989).

A seven-point scale was the attitude measurement tool for all experiments described in this thesis. When a participant strongly agrees with a positive statement the score is 7 and when they strongly agree with a negative statement the score is 1. Upon completion of the questionnaires the mean for each usability attribute across all participants is calculated. This mean score provides quantitative evidence of the overall attitude to this attribute.

A number of other aspects are also fundamental to the design of a "reliable and valid" questionnaire. It is important to clearly layout the questionnaire statements, which helps the participants read and understand them. The instructions on how to complete the

questionnaire must also be clear. The usability attribute statements should be evenly spaced and numbered. It is also essential to provide "a clear visual relationship between each item and the possible response options". This can be clearly seen in Figure 3.1, where the possible options appear clearly over each response box. The actual questionnaires used for the four experiments described in this thesis are listed in the appendices.

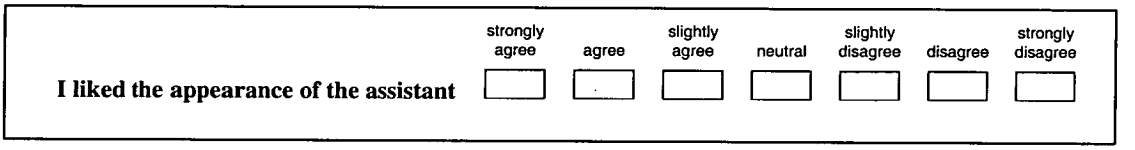


Figure 3.1 Example of a Likert Questionnaire Statement

To confirm the validity of the questionnaire it is important to pilot it with researchers who have expertise in the required field. This procedure helps to select the best usability attribute statements for the final questionnaire version and ensures the statements represent the experiment aims accurately. This procedure took place during the cognitive walkthrough procedure of the experiment design.

3.4.2 Qualitative Information

Qualitative data which is non-numeric data, can be difficult to analyse but can be useful to provide explanations for findings that emerged from the quantitative analysis. Qualitative techniques never validate theory in the same way as quantitative techniques can, but they essentially serve to provide explanations and relationships between numerical findings and help to build theories, as opposed to testing theories. In all four experiments in this thesis participants took part in interviews. These interactions served as effective strategies to give participants a voice, allowing them to comment on aspects of their experience, which may or may not be relevant to the purposes of the experiment.

> Interviewing users about their experience with an interactive system provides a direct and structured way of gathering information. Interviews have the advantages that the level of questioning can be varied to suit the context and that the evaluator can probe the user more deeply on interesting issues as they arise (Dix et al, 1993).

To be effective the interviews conducted in this thesis were planned in advance where the interview questions focused on central themes in order to constructively explore

important issues. The chosen themes help to encourage the participant to explain informally their attitudes and opinions about the interaction. Importantly interviews provide opportunities for participants to suggest improvements to the application and in the context of the progressive experiments reported in this thesis, the feedback from participants was hugely important as regards design decisions and research strategies for the continuum of the empirical research. It should be noted that it is often not possible to obtain detailed comments from all participants as some will be more enthusiastic with divulging information than others, but overall interviews were effective tools for retrieving qualitative information that served to augment the quantitative findings.

Focus groups are a second method with which to retrieve qualitative information. In addition to interviewing, focus groups can also be effective to explain in more detail participants' attitudes and opinions about the experiment topic. A focus group discussion must be a carefully planned event with a selection of participants from the experiment (normally 6-8) and is designed to obtain perceptions on a defined area of interest in a permissive, non-threatening environment. The topics of discussion are selected by the researcher and reflect important issues about the interaction or participants' experience. According to Krueger (1994) focus groups have the advantage of engaging in discussion in a group situation, which can be less intense than in a one-to-one interview situation. The salient features of the experiment tend to be discussed and often, unexpected issues are raised.

3.5 Statistical Analysis

Once the quantitative data was gathered it was important to select the methods with which to analyse it. A significance-testing paradigm, known as the *F* test, after R.A. Fisher, the statistician who developed it, is used as the primary form of statistical analysis throughout this thesis. To conduct any statistical analysis it is essential to define the null hypothesis, H_0. The null hypothesis means that there exists no difference in the dependent variables between the levels of the independent variables or when testing for a difference between two conditions, the mean difference will be 0. A result becomes significant, if enough evidence exists to reject the null hypothesis. To do this a significance level must be established. This significance level determines how strong the evidence must be for the null hypothesis to be rejected. When a statistical test is

completed a p value (significance figure) is produced. The smaller this figure is, the stronger the evidence is to reject the null hypothesis. It is common to use two significance levels in parallel. Typically $p < 0.05$ is commonly used to describe a significant result and, $p < 0.01$, describes a result that is highly significant.

A factor that must be taken into account is whether the hypothesis is one-tailed or two-tailed. The former is concerned with the effect and the direction of the effect, and the latter is concerned with the effect only. For a one-tailed hypothesis, in addition to the null hypothesis, H_0, a second hypothesis, H_1, is introduced. A significant result occurs when evidence to reject H_0 *and* evidence to support H_1 emerges. Typically H_1 is defined as a departure *in one particular direction* from what would be expected on the assumption that H_0 is true. For this reason there must already exist strong prior evidence for a predicted departure in one particular direction only. A two-tailed test, or a non-directional test is defined by a departure from H_0 in an unspecified direction and the significance values appear at either end (or tail) of a distribution curve. Unless there is theoretical evidence to expect a departure in a particular direction it is better to use a two-tailed test. With a one-tailed test, any data that shows a discrepancy from H_0 in the unspecified direction is ignored and this could lead to misinterpretation of statistical results. As there was no strong evidence to suggest effects in specified directions in any of the experiments described in this thesis, two-tailed tests were used. Figure 3.2 illustrates a histogram where the regions of significance at $p < 0.05$ and $p < 0.1$ in a two-tailed test are highlighted.

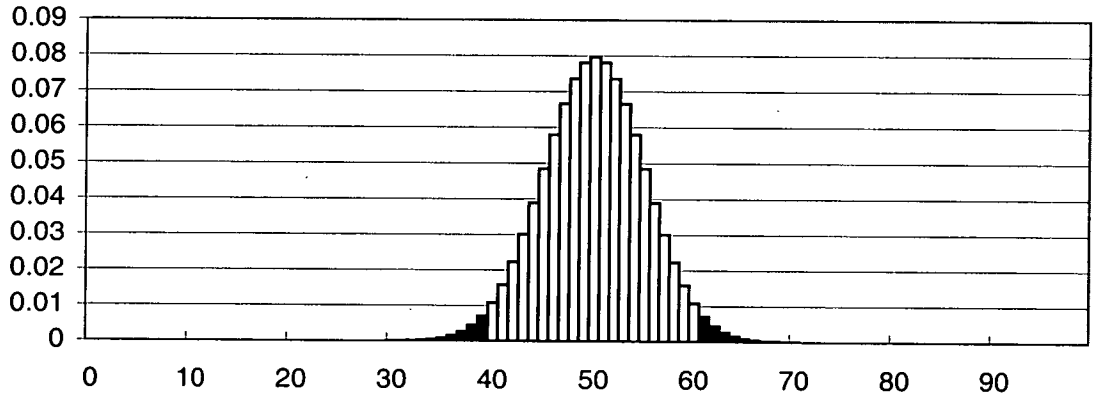


Figure 3.2 Regions of Significance in Two Tailed Tests

The type of data to be analysed also determines which type of statistical test is most suitable and for this reason it is essential to understand the format of the collected data. Different data types will have to satisfy different conditions in order for significance to be detected. Data can be categorised into three categories: nominal, ordinal and interval. *Nominal* data is described as data points that are drawn from a set of two or more discrete categories and this type of data has the weakest conditions attached it. The data points in *ordinal* data are drawn from a set with an ordered association and values in this set can be thought of as "higher" or "lower" than another value. Likert-type questionnaires can be classed as ordinal data, which does not guarantee normal distribution. *Interval* data sets have well defined distances between points as well as a defined order. The conditions to be satisfied for interval data are the strongest and the tests on this type of data assume data samples are from a normal distribution.

In addition, statistical tests can be categorised into whether they are parametric or non-parametric. A parameter is a characteristic of a population and a parametric test is one that depends on population characteristics, or parameters for its use. Typically it is a requirement of parametric tests that the populations from which the data has been taken are from normal distributions. The requirements for non-parametric tests are minimal and the shape of the population is unimportant. Non-parametric tests are sometimes referred to as distribution-free tests, (Pagano, 1990).

3.5.1 Nominal Data Tests (Non-parametric)

A ***chi-square test*** is a suitable test for nominal data to investigate if certain categories occur more frequently than others. The null hypothesis is that the frequencies of the categories are equal. This test assumes that the data points in the sample are independently selected from the population; a characteristic of non-parametric tests. An example of the use of a chi-square test can be found in Chapter 6.

3.5.2 Ordinal Data Tests (Non-parametric)

Ordinal data is often concerned with the direction of change or difference. The differences of interest are (1) a difference in repeated measures from a particular participant or (2) differences between different participants in a sample. The ***Wilcoxon signed-rank test*** is effective for testing within-subject differences. The null hypothesis

is that no differences emerge for the different stimuli (or measures) the participant experienced. The ***Mann-Whitney U test*** is similar but tests for between-subject differences.

3.5.3 Interval Tests (Parametric)

One of the more popular tests on interval data is the ***t-test.*** A related samples t-test is popular with a repeated measures design and is frequently used to analyse data in this thesis. It is an appropriate test to use when it is necessary to compare two interval dependent variables measured on the same set of participants. The null hypothesis is that the two variables have the same mean value in the population. Extending from this an unrelated samples t-test is effective for completing tests on two sets of interval measurements on different sets of participants. The null hypothesis is that the means for the two conditions are equal.

Throughout this thesis it is also a frequent procedure to compare more than two sets of data. An analysis of variance (ANOVA) is an ideal test for data where there are multiple independent variables. In certain circumstances these independent variables may have a number of values. Interactions between the multiple variables and variable levels may occur, affecting the dependent variables in different ways. The ANOVA method used frequently throughout this thesis is a test that is well suited to testing hypotheses on multiple independent variables. The ANOVA returns a significance figure for the main effect of each variable and also a significance figure for any interaction between independent variables. The null hypothesis for a main effect is that the mean of the dependent variable is the same for each independent variable under test.

An extension of this is a repeated measures ANOVA which allows testing of between-subject and within-subject variables. The within-subject cases arise when a subject sample experiences more than one condition at more than one level (e.g. in Chapter 4 the within-subject variables are agent gender and agent type). The between-subject variables can account for the differences in participant variables, for example participant gender and participant age groupings. An ANOVA centres on the F statistic. This figure is the ratio of two independent variance estimates and it measures the overall deviation from uniformity. It will normally be higher if an effect is present.

3.5.4 Note on Statistical Tests

The quantitative data retrieved from the experiments conducted for the purpose of this thesis were primarily gathered using ordinal scale questionnaires (i.e. Likert scales). Nonetheless parametric statistical tests, such as t-tests and ANOVA's were used to analyse the data. The primary assumption of a parametric test is that the data samples come from normal distributions and as explained earlier ordinal data does not guarantee a normal distribution. However, it is possible to use parametric tests on ordinal data as they are robust with regard to violations of the underlying assumptions. Unless the data substantially departs from normal, the sampling distribution of t will remain approximately the same. Parametric tests are more versatile and powerful than non-parametric tests and in some cases no equivalent non-parametric test is available to analyse specific data, such as data with multiple variables.

In general, parametric tests are used whenever possible (Pagano, 1990), but if an extreme violation of the assumption occurs it is essential to use non-parametric tests. This thesis explored some of the data using non-parametric tests first and compared these results with figures retrieved from parametric t-tests. No substantial differences emerged and this result provided enough confidence to use more powerful parametric tests throughout the thesis.

3.6 Summary

This chapter has provided a description of the experiment methods employed throughout this thesis in order to empirically evaluate embodied conversational agents. Firstly, the importance of using effective evaluation methods for both the experiment interfaces and the actual experiment design was described. The complementary nature of a number of evaluation strategies was documented together with their impact on the empirical research that was conducted in this thesis.

A description of the experiment design method, taken from methods commonly used in applied experimental psychology followed. The strategies with which to retrieve information regarding the topic under investigation, including the particular qualitative and quantitative methods specific to this thesis were discussed. An in-depth discussion

about the statistical analysis techniques then followed, explaining the different types of data and the corresponding strategies needed to analyse the information in order to document the research findings.

The four chapters that follow detail a series of progressive empirical studies in which the effectiveness of embodied conversational agents was assessed. Using the experiment methods which have been described in this chapter, conclusions are formed regarding the representations and functionality of embodied conversational agents in the context of virtual retail applications.

Chapter 4

Implementation of a Retail Interface Template to Evaluate the Effectiveness of Humanoid Photo-Realistic Agents

4.1 Introduction

The presence of conversational agents in graphical user interfaces (GUI) can result in an efficient, engaging and social collaboration between humans and machines, as demonstrated by 'Steve' (Johnson & Rickel, 2000), 'Rea' (Cassell et al., 2001), 'Gandalf' (Thorisson, 1996) and 'PPP-Persona' (André et al.,1999). However, it cannot be assumed that the use of these agents in GUI guarantees successful human-computer interaction. The research detailed in this chapter has been motivated by the growing need for objective and subjective measures of usability for embodied agents. As Cassell (1999) observes, empirical investigations of any kind of embodied interfaces are rare and the results so far have been equivocal. The enormous difficulty in drawing comparisons and general conclusions between the current empirical evaluations of embodied agent interfaces is also addressed by Dehn and van Mulken (2000) and in order to classify existing evaluations and group future studies, they define three separate categories. These include the importance of measuring:

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

While the importance of the interactive and contextual aspects of the second and third categories must be recognised, there remain many important, outstanding issues with respect to the subjective experience of interfaces using embodied agents. The evaluation reported in this chapter closely addresses participants' subjective experience of observing a range of agents. To emphasis the structure of this thesis, the controlled experiment described in this chapter concentrates solely on one section of the array of humanoid agents as per Figure 1.2, namely photo-realistic humanoid representations of

the agents. The chapter that follows describes an experiment with a similar evaluation approach, however the agents assessed were represented as humanoid animated agents. The humanoid agents were assessed in novel retail web-based environments, as it is observed that there is a lacunae of empirical evaluations of agents in retail environments, but there exist reasons to believe that such agents may have a positive impact for the benefit of the user (see Chapter 2).

The existing empirical evaluations (Koda, 1996; Sproull et al., 1996; Takeuchi & Nagao, 1993) that focus on the user's subjective experience, draw attention to important measurable dimensions of embodied agents such as perceived intelligence, believability, likeability, entertainingness and usefulness. The experiment reported in this chapter extends this list of measurable dimensions and in doing so provides quantitative and qualitative comparative data with respect to humanoid photo-realistic agents. Not only is a spectrum of photo-realistic agents carefully selected, created and evaluated, for the first time consideration is also paid to the gender of the agent. Both male and female versions of each type of photo-realistic agent were created and represented in the experiment and empirical results regarding attitudes to the agent's gender are documented.

In total, five types of humanoid photo-realistic agents were selected to represent the range of possible technologies that could be used to implement photo-realistic agents in order to anthropomorphise retail interfaces with virtual assistants. The agents included male and female versions of a video, a 3D talking head, a photo-realistic image with animated facial movement, a still image and a disembodied voice (Figure 4.1). The experiment was dependent on observation techniques, whereby participants were invited to observe (or 'eavesdrop') a recorded simulation, which depicted the agents interacting with a customer. Each observation sequence lasted approximately two minutes. Following each observation participants completed a series of questionnaires designed to gather quantitative evidence that measured their subjective experience of observing the agents. This passive methodology employed to assess the agents was extremely practical as it avoided complex technological issues involved in creating fully functional interactive applications for each of the five agent types, although it still allowed a substantial evaluation of each of them. Ideally an interactive application would have opened the opportunity to obtain more in-depth results regarding participants' cognitive experience had they themselves conversed with the agents. However to evaluate such a

substantial cast of agents this compromise of using observation methods instead of interactive methods was made.

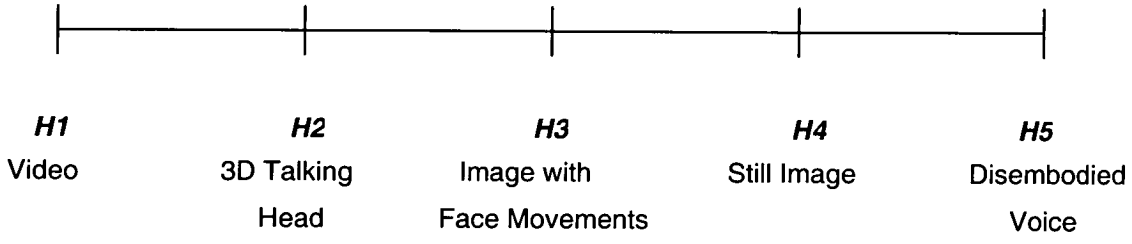


Figure 4.1 Range of Five Agent Types Evaluated

The agents were evaluated in two contrasting retail applications in order to investigate any application dependency effects. An interface template incorporating graphical and auditory output was designed and developed to support the retail applications and each of the five agent types. The agents were first evaluated in a Home Furnishings Service (Part I) and at a later date, a second participant sample evaluated the agents in a CD Service (Part II). This chapter describes the design and implementation of the experiment interface template, the five photo-realistic agent types and the selection of retail applications. The experiment design is presented, as are the results that focus on participants' perceptions of the usability of the agents, the effectiveness of the interaction, the efficiency of the system and satisfaction with the interaction.

4.2 Experiment Interface Design

A template of the experiment interface was designed and used to combine the audio and visual components of each of the five agents types and the visual content of the retail applications, which had to be graphically updated throughout each of the two minute observation sequences that were presented to the participants during the experiment. Throughout the design of this template interface the guidelines of non-expert evaluations were referred to, as mentioned in Chapter 3. These guidelines ensured that at all times the template interface was designed to be consistent. Figure 4.2 describes the components of the interface template, which were combined using the multimedia

authoring tool Macromedia Director 6.5 to create an individual 'projector' file. Using the concept of combining cast, stage and score (Chapter 2), Macromedia Director 6.5 effectively combined the three sections of the interface template. The cast included the audio and visual components of the five agent types. The stage was visible in the main window which displayed the graphical contents specific to the application and the score specified the temporal display of the media objects on the stage, which were the objects visually displayed in the selection panel. The interface was presented on a standard 32cm x 24cm PC monitor, with GeForce2 MX™ graphics card (Table 4.1, Figure 4.3).

Graphics Card Description	GeForce2 MX™ 700 million texel per/sec rate 2D/3D resolution of 2048 x 1536 at 75Hz 2.7GB/sec memory bandwidth	
Interface Dimensions (cm)	**Length**	**Breadth**
Monitor	32	24
Agent Window	4.5	4.5
Main Window	16	16
Selection Panel	18	3.5

Table 4.1 Interface Description

The template can be further described as a combination of the agent window, the main application window and the selection panel. Stand-alone Apple QuickTime movies for all of the visible agents were created and displayed in the agent window. The desktop video editor, Adobe Premier 5.1 was used to create a series of .JPG image frames, which were played in sequence using Apple QuickTime to display the changes taking place in the retail application. Again using Adobe Premier 5.1 a series of .JPG image frames were created and sequenced using Apple QuickTime to illustrate a selection panel unique to the particular application.

The technical production of the audio and visual components used to create each of the five humanoid agents is presented next, together with reasons for the inclusion of particular agents in the range of humanoid photo-realistic agents evaluated in this chapter. This is followed by a description of the components of the main window and selection panel for the two contrasting applications, the Home Furnishings Application and the CD Service Application.

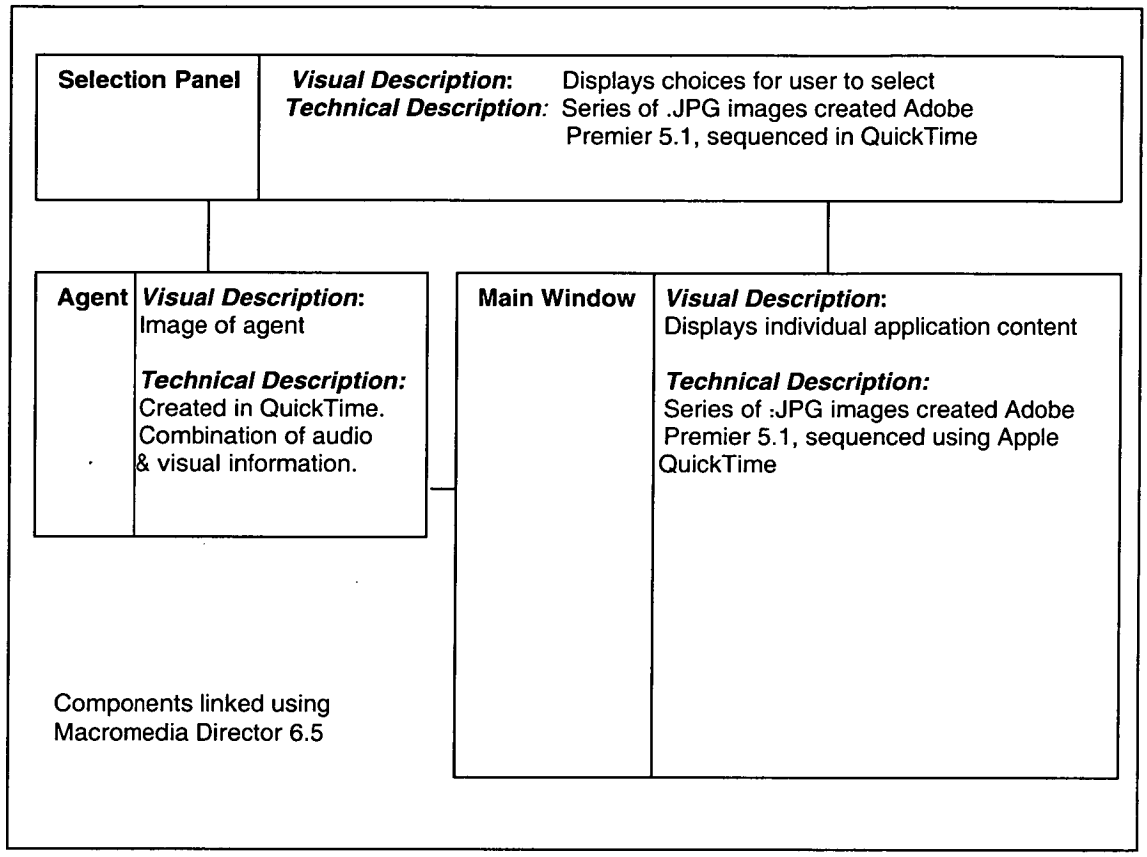


Figure 4.2 Description of Experiment Interface Template

4.3 Agent Types

All of the agent types (excluding the disembodied voice) were created from human photo-realistic images. The video agent type was included in the experiment to investigate expectations toward human facial movements in comparison to the visually less sophisticated facial movement of the other agent types. Although in an interactive situation the use of live or even pre-recorded video may be impractical in online applications, the MikeTalk project (Ezzat & Poggio, 1998) has demonstrated that the illusion of video can be created successfully without recourse to the use of pre-recorded segments. Consequently, in the event of the video humanoid photo-realistic agent type being preferred, technology such as that used in the MikeTalk project could be implemented as a suitable substitute for the successful development of a fully functional application interface with an interactive video agent present.

The second humanoid photo-realistic agent type selected for evaluation was a 3D talking head agent. Parke and Waters (1996) pioneered work in the area of 3D talking heads and since then have created DECFace (Waters & Levergood, 1995), a synthetic face that uses an automated lip-synchronisation algorithm. In support of the inclusion of this agent type in the evaluation they have stated:

> One of the more intriguing possibilities (for future research) is the construction of interactive face agents capable of assisting and conversing with the user.

As discussed in Chapter 2, more recent research and development with respect to 3D talking heads has been led by Massaro (1998). This work has been instrumental in investigating users' perceptions of speech output from talking heads, using the sophisticated computerised 3D talking head, 'Baldi'. The development of 'Baldi' demonstrated that technology is capable of creating sophisticated lifelike 3D talking head agents and although this agent is primarily being successfully utilised in educational domains, there exists potential for the deployment of such agents in other domains, such as electronic retailing.

In addition to the work by Massaro, Thalmann (2000) is focusing on facial communication in virtual environments and has stated:

> 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.

To create this sense of realism the research is largely concerned with the development of emerging MPEG-4 standards in particular those for facial cloning, real-time animation and face feature tracking. The work draws heavily on the model of human facial expression developed by Ekman (Ekman & Friesen, 1969; Ekman & Friesen, 1978) to represent non-verbal communicative behaviour. To simulate the natural representation described by Thalmann much contemporary research is steering toward the development of emotive and expressive faces for agents for in an autonomous dialogue it is necessary for the agent to blink, display head movements and exhibit emotive nuances (Thalmann, 2000; Ruttkay & Hoot, 2001; Massaro et al., 2000).

Following research findings by Takeuchi and Nagao (1993) the evaluation also includes images of agents with and without facial movement. Their efforts at introducing facial

displays, such as frowning and smiling, in multi-modal human computer interaction were successful in a games domain. Extending their research, this evaluation asks whether participants can benefit from facial displays in a retail domain. Takeuchi and Nagao firstly 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. To investigate this occurrence in a retailing domain two agent types were created, one without facial displays, i.e. a still image and secondly, an agent based on a photo-realistic image of a human with graphically animated facial movement. Takeuchi and Nagao (1993) 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 anthropomorphised interfaces and suggest that it shows that the user is appreciative of the human image that he or she tries to interpret. Their finding provides a second reason for the inclusion of the still image agent type and the image with facial movement in the evaluation in this chapter, namely to explore further user attitudes to agents that may or may not cause a difference in concentration levels. The final agent type in the chosen spectrum of agents was a disembodied voice agent. This agent type was included in the evaluation, to compare user perceptions to visible agents with agents that only have conversational capabilities and are not visible in the interface. Including this agent type encourages feedback about the conversational aspects of the retail interfaces and whether this type of communication is enhanced by the visual facial displays of certain agent types.

4.4 Agent Implementation

The following section provides details of how the five agent types were created. Both male and female versions of all agent types were developed.

4.4.1 H1 - Video

Using an auto-cue to read pre-defined scripts, one male and one female person played the role of assistants and were recorded on video. The two-minute scripts were dialogues between a 'customer' and an assistant.

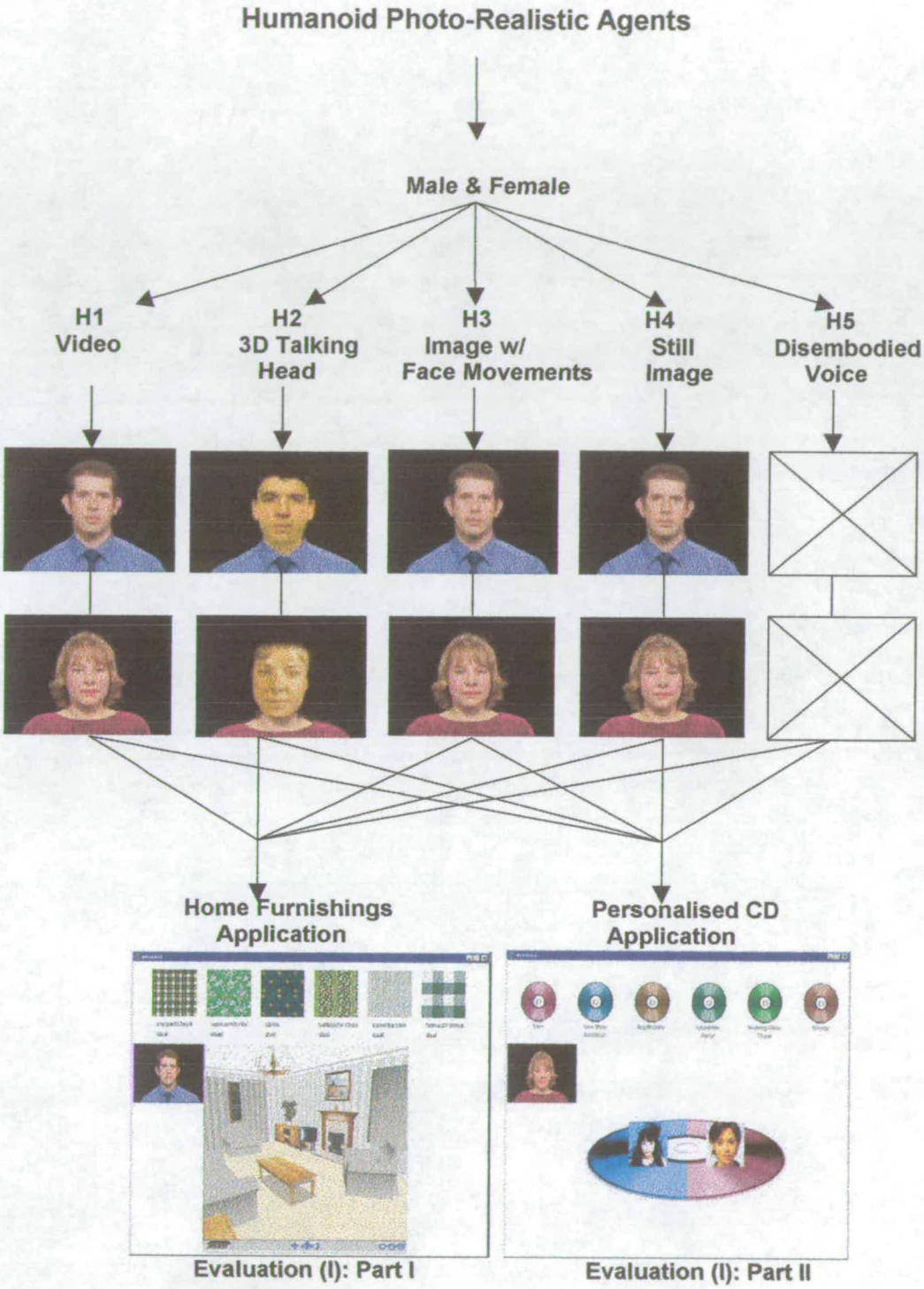


Figure 4.3 Illustration of Structure of Passive Viewing Evaluation I

The soundtracks from these movie (.AVI) script files were extracted and used as the speech output for all other human-like agents, e.g. the female assistant voice soundtrack was used for the other four female human-like agents. Natural facial movement and speech intonation was recorded when creating these videos. For example when asking a question talkers tend to display rising intonation at the end of the utterance and for a statement, talkers tend to use falling intonation. The decision to use natural speech instead of synthetic speech was made based on the evidence (McInnes et al., 1999) that user attitudes to concatenated natural speech are significantly more positive than synthetic speech. This is confirmed further by the results of Granström (1999), who concluded that embodiments with natural speech were significantly preferred by participants in a experiment than embodied agents with synthetic speech. This thesis does not dispute the use of synthetic speech for embodied agents as advances are being made in particular with respect to the problems of imitating human pitch and prosody. However, in this thesis where the agents' appearance was of primary concern in the evaluations, it was necessary to avoid audio-visual effects between appearance and the speech output.

4.4.2 H2 - 3D talking head

Using 3D Studio Max Character Studio, the character modelling software kit, a 3D wireframe head was constructed (Figure 4.4). Still images of the persons used to create H1 were mapped onto the wire-frame model of a human head. A set of facial movements was created, displaying mouth movement and blinking. These served as the key frames for the animations which were saved in .AVI format and then played in sequence to create the agent video. Facial displays were introduced in a controlled manner: when asking questions the assistant raised its eyebrows and during an affirmation the agent nodded, raised its eyebrows and blinked at the end of the sentence (Cassell, 2000). Examples of some of the facial movements displayed by this agent type are illustrated in Figure 4.5.

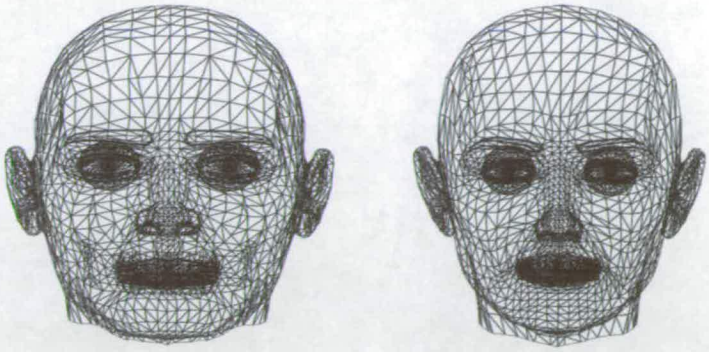


Figure 4.4 Male and Female Wireframe Heads

Figure 4.5 Facial Animations for Male 3D Agent

4.4.3 H3 - Image with animated facial movements

Photo-realistic still images of the male and female videos were taken. Using Adobe PhotoShop 4.0 varying .PSP image frames were produced illustrating blinking and mouth movement, known as visemes. These frames were imported into Adobe Premier 5.1, edited and run in a sequence to create an animation of the face with graphic facial movement that included eyebrow raising and blinking.

4.4.4 H4 - Still image

This agent type was a static image of the male and female assistants who appeared in the video recordings. The still image was converted to a .PSP image which had a pixel dimension of 107 x 85 and a pixel depth to colour ratio of 24 to 16 million.

4.4.5 H5 - Disembodied voice

Created using only the audio soundtrack files extracted from the recordings for H1 the inclusion of the disembodied voice agents raises issues about the need for visible agents in the interactive retail interfaces. All the visible agents (H1 to H4) appeared in the agent window on the interface template (4.5cm x 4.5cm). For this agent, H5, the agent window appeared blank. The audio file created, as with all previous agents was sampled at 48kHz in stereo with a 16-bit resolution.

4.5 Application Implementation

Two contrasting retail application interfaces were created using the interface template described previously. The GUI's for both applications are illustrated in Figure 4.6, where the three distinct components of the template, including the agent window, the main application window and the selection panel can be examined.

4.5.1 Application 1 – Home Furnishings Service

The GUI was created in the style of the MUeSLI Multi-modal Retail System (Wyard & Churcher, 1999), a conversational retail application. The original multi-modal spoken language system combined speech and touch input. It had an advanced dialogue manager to allow natural flexible style interaction with the agent. The system incorporated mixed initiative dialogues, with system expertise to allow the agent to make appropriate suggestions to guide the user. The MUeSLI system was designed primarily to investigate multi-modal issues of combined speech and tactile input from the user. No evaluations were conducted to evaluate or assess the representation of the agent prior to the experiment reported in this chapter. The original MUeSLI system had an advanced dialogue manager that allowed flexible spoken output from the agent. When designing the two-minute dialogues for the observation files created for this particular evaluation, these dialogue features were included as much as possible. The dialogue participants 'eavesdropped' upon a male 'customer' conversing with the agent to decorate the room. Throughout the dialogue, which can be examined in Table 4.2, mixed initiative can be observed, which displays the agent's capability to make suggestions relevant to the needs of the user.

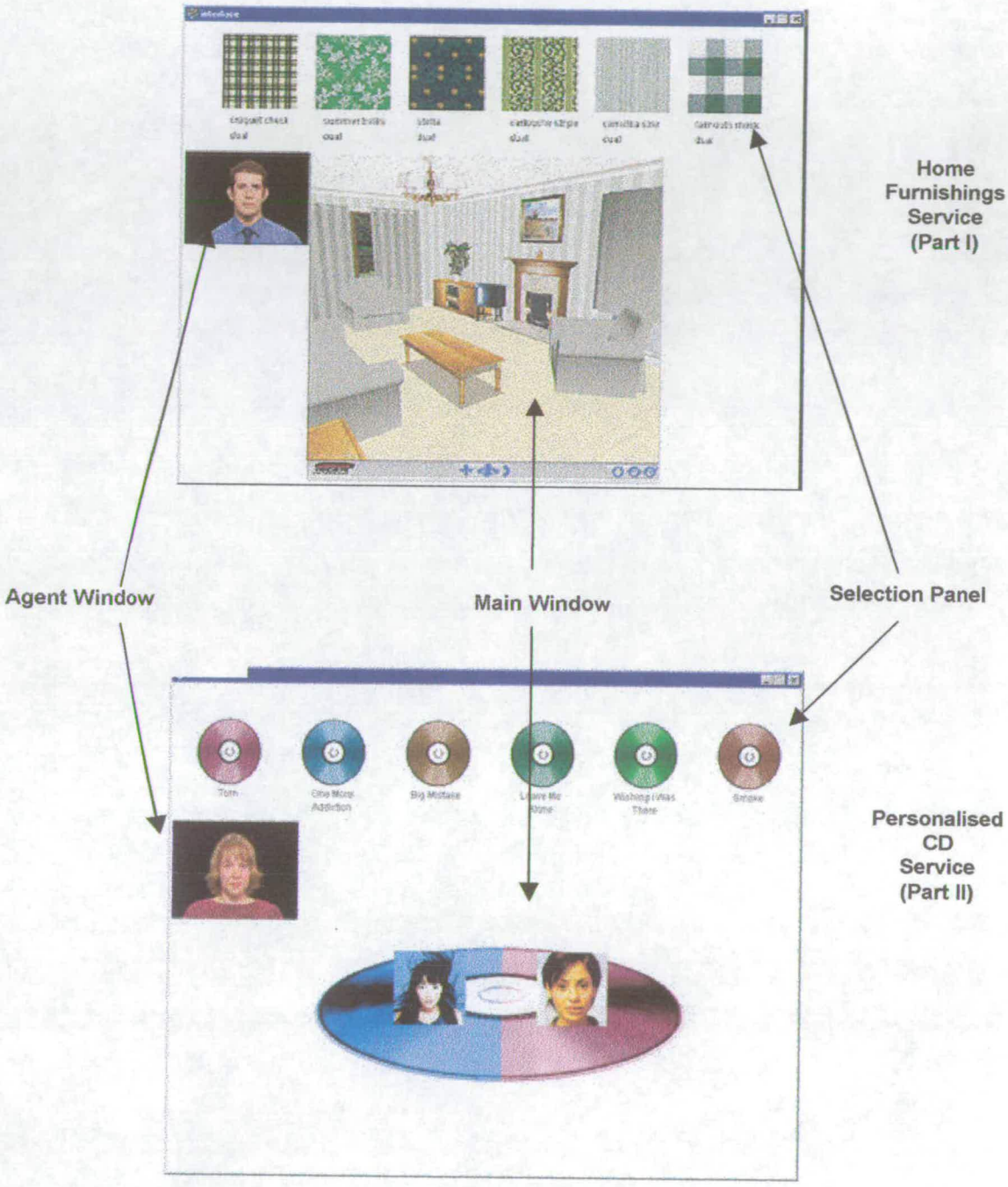


Figure 4.6 Application Graphical User Interfaces

The participants in the experiment 'eavesdropped' on identical dialogues between the customer and all the agents included in the evaluation. The theme of the dialogue was for the customer to engage in a conversation with the agent in order to decorate a virtual living room space. All interactions between the customer and the agents were identical. The agents appeared in the window to the left of the main application window and the selection panel appeared above the main window (see Figure 4.3 & 4.6). The panel was divided into six evenly spaced sections, each containing a fabric sample and the name of the fabric. The user could select these fabrics and ask to see the fabric on display in the room. The main window in the interface depicts a 3D view of a living room, dynamically updated following changes to the fabrics requested by the customer. Example .JPG files created using Adobe Premier are presented below to graphically illustrate the changes that were being made to the interface as the conversation progressed (Figure 4.7).

4.5.2 Application 2 – Personalised CD Service

Although this application was deliberately created to be a contrast to the content of the first application, the interface was designed to be similar to minimise visual effects between the evaluations. To achieve this the interface template described earlier was used and as before, three main components were created and combined in Macromedia Director to produce an observation projector file that was used to present the agents to the participants in the experiment. The main application window in the interface was of the customer's personalised CD, updated after the addition of new tracks. The selection area contained a row of six tracks from one artist, which could be selected for inclusion on the personalised CD. The name of the tracks appeared directly underneath a colour image of a CD. The 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 4.3). The customer, upon request, could listen to musical excerpts of the tracks that were available. As with the first application, the dialogue was specifically designed to exhibit system initiative in the form of expert suggestions from the agent. A series of .JPG images of the main application area illustrates the changes visible to the participants during the duration of the observation (Figure 4.8).

(A) No patterns

(B) 'Summer Trellis' on Chairs and Sofa

(C) 'Cameilla' Curtains

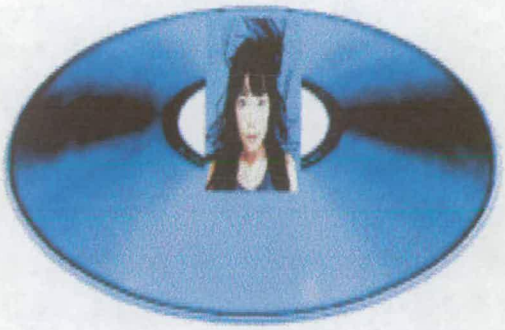
(D) 'Stella' on Chairs and Sofa

Figure 4.7 Sequence of .JPG's Illustrating Changes in the Interface

Speaker	Dialogue Content	Interface Action
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.	Six swatches appear in selection panel
Customer	Try the 'Cartouche Stripe' please.	'Cartouche Stripe' appears on the sofa
Customer	Hmm, No. Show me 'Summer Trellis' instead.	
Assistant	Certainly.	'Summer Trellis' appears on the sofa
Customer	Hmm. I'm not sure.	
Assistant	Would you like to see it on the chairs as well?	
Customer	Yes, ok.	'Summer Trellis' appears on the chairs
Customer	Yes, I like that one.	
Customer	What curtains will go with it?	
Assistant	Can I suggest a plain dark green. This one perhaps.	Green fabrics appear in selection panel
Customer	No that's too dark. Show me the 'Camellia Stripe'.	
Assistant	Certainly.	'Camellia Stripe' appear on curtains
Customer	They're nice.	
Customer	Now I want to change the sofa fabric.	
Assistant	Here are some matching sofa fabrics.	Sofa fabrics appear in selection panel
Customer	Show me 'Stella'.	'Stella' appears on the sofa
Customer	Show it on the chairs too.	'Stella' appears on the chairs
Assistant	Would you like to see some wallpaper?	
Customer	Yes, I'd like some striped wallpaper.	
Assistant	Here is a selection.	Wallpapers appear in selection panel
Assistant	'Colourwash Stripe in Ivory' would go well with your other choices.	
Customer	Ok, show me that one.	Colourwash Stripe appears on the walls
Customer	Well maybe...	

Table 4.2 Dialogue for Home Furnishings Service (Part I)

(A) Blank CD



(B) Addition of 'Bjork' Track

(C) Addition of 'Imbruglia' Track

(D) Addition of 'The Corrs' Track

Figure 4.8 Sequence of .JPG's Illustrating Changes in the Interface

Speaker	Dialogue Content	Interface Action
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?	
Assistant	Would you like to hear a track?	
Customer	Yeah, let's hear 'Human Behaviour'.	Play extract of 'Human Behaviour'
Customer	Hmm, I'm not sure.	
Customer	Can you play 'Big Time Sensuality'.	
Assistant	Sure, here you go.	Play extract of 'Big Time Sensuality'
Customer	Hmm, it's hard to choose, I only want one Bjork track	
Assistant	Do you want to hear 'Human Behaviour again'	
Customer	Yeah.	Play extract of 'Human Behaviour'
Customer	Yeah, I do want that.	
Customer	And I want something by Natalie Imbruglia on the second track.	
Assistant	Oh something of 'Left of Middle', how about this track.	Play extract
Customer	No, not that, maybe 'Wishing I was There'	Play extract
Assistant	Nice sound.	
Customer	Yeah, I want that one.	
Customer	Then something by The Corrs.	
Assistant	Sure, we've got everything by 'The Corrs'.	
Customer	Let's hear 'What Can I Do'.	Play extract of 'What Can I Do'
Customer	Oh, isn't that the single. I want the album track.	Play extract of 'What Can I Do'
Customer	Yeah, that's the right one.	
Assistant	Now, how about the next track.	
Customer	Next I want 'Ray of Light'.	
Assistant	After 'What Can I Do', you ought to have the single version.	
Customer	Yeah, ok.	Play extract of 'Ray of Light'
Customer	Well, maybe	

Table 4.3 Dialogue for Personalised CD Service (Part II)

After all the components of the retail interface were designed and constructed, an expert evaluation of the interface took place. As stated in Chapter 3, this procedure is important from a design point of view to identify any problems with the interface and for suggested solutions to be made. After some minor alterations were made to the template interface, the experiment predictions and procedure could be documented.

4.6 Experiment Predictions

1. It was predicted that the conversational applications would be received positively based on the findings of other research discussed previously. However as no other evaluations to date specifically address the inclusion of agents in retail interfaces interesting questions may be raised.

2. 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 similar, and as the same female voice was used for all female agents the attitude to the voices of all female agents would be similar.

3. It was predicted that within agent types the male and female agents would be rated similarly as they were designed to have the same verbal and non-verbal behaviour, the only difference was their gender realisation. It was predicted that attitudes might differ between the agent types.

4.7 Experiment Design

In order to test these hypotheses two groups, each of 32 participants were invited to contribute to each part of the evaluation: the Home Furnishings Service (Part I) and the CD Service (Part II). Table 4.4 details the participant age and gender distribution. As stated in Chapter 3, cognitive walkthrough evaluation was completed for experiment design. From this it was concluded that no constraints should be placed over the participant sample as regards computing experience, as the participants were only asked to listen and observe and were not asked to use the mouse or keyboard.

	Home Furnishings Part I		CD Service Part II		Total
Participant	Male	Female	Male	Female	
Age 18-35	8	5	6	6	25
Age 36-49	1	5	5	5	16
Age 50+	7	6	5	5	23
	16	16	16	16	64

Table 4.4 Analysis of Participants by Gender and Age Group

In both parts of the evaluation participants from both samples were firstly welcomed, then asked to take a seat in front of the experiment PC where they were invited to first read a brief explanation of the purpose of the experiment. The experiment task sheets can be examined in Appendix 1.1. The customer was represented by a male disembodied voice for all interactions. To assist with this explanation the participants were shown an image of a customer at a computer, and they were asked to imagine that they were positioned behind this customer and eavesdropping on the conversation between the user and the customer. This graphical image is seen in Figure 4.9.

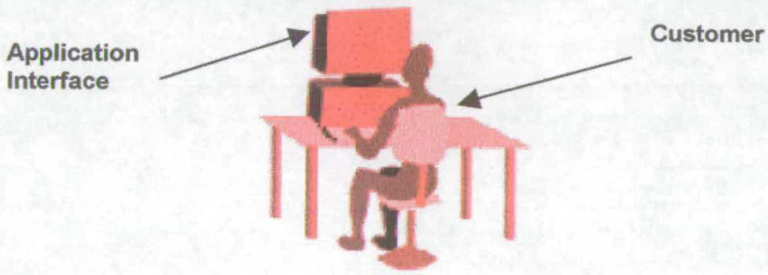


Figure 4.9 Image Used to Explain Passive Viewing Methodology

Each of the participants viewed a total of ten two-minute videos (created using Macromedia Director 6.5) and these were presented in randomised order on a Pentium II PC. The videos showed the dialogue between the customer and an agent. After each participant had witnessed a customer-agent interaction they were asked to complete a Likert format questionnaire. A discussion about Likert type questionnaire is provided in Chapter 3.

The questionnaire statements were designed using the blueprint method explained in Chapter 3. Content areas were defined as (1) application and (2) agent. The manifestations were defined as (1) behaviour, (2) functionality and (3) appearance.

Using this method an exhaustive list of questionnaire statements were created and from this the most relevant questionnaire statements were selected. Within the final questionnaires, statements were balanced for polarity, with an equal number of positively and negatively worded stimulus statements. Not all dimensions of interest were relevant to all the agents in the evaluation; therefore three different questionnaires were used. A total of eight statements were relevant to all the agent types and were therefore included in all questionnaires. Other statements relating to appearance (not relevant to H5: disembodied voice) and lip-movement were included as appropriate. The questionnaire statements are listed in Table 4.5 representing five main sections: attitude to the application, agents' voices, agents' competence and friendliness, agents' appearance and finally agents' lip movement. The three questionnaire types illustrating the actual order of presentation can be examined in Appendix 1.2.

Questionnaire Statements	H1	H2	H3	H4	H5
1. I think this service is a good idea	*	*	*	*	*
2. I think this service would be difficult for me to use.	*	*	*	*	*
3. I would like to use this service myself	*	*	*	*	*
4. The assistant's voice was not clear enough.	*	*	*	*	*
5. I liked the assistant's voice.	*	*	*	*	*
6. I felt the conversation was unnatural.	*	*	*	*	*
7. I felt the assistant was competent	*	*	*	*	*
8. I felt the assistant seemed unfriendly	*	*	*	*	*
9. I thought being able to see the assistant was helpful.	*	*	*	*	
10. The appearance of the assistant was unsuitable for the application.	*	*	*	*	
11. I liked the appearance of the assistant.	*	*	*	*	
12. I thought the assistant looked natural.	*	*	*	*	
13. I felt the speech sometimes didn't match the lips.	*	*	*		
14. I noticed the lips moving.	*	*	*		

Table 4.5 Questionnaire Statements

The dependent variables in the experiment were the responses to the individual questionnaire statements and also the responses given during a semi-structured interview, including an overall rating of each agent which was documented in the questionnaires. After participants had seen all ten agents (5 male, 5 female) in an application they then took part in a closing interview that was designed to investigate:

- Participant's views of the use of humanoid photo-realistic agents in e-retail applications.
- The effective deployment of these agents on screen.
- The characteristics required by such agents.
- The conversational possibilities with agents in future applications.

Title	**Passive Viewing Evaluation I: Humanoid Photo-Realistic Agents**	
Design		Two Independent Samples (due to 2 evaluation applications)
Predictions	4.1	The conversational applications would be received positively.
	4.2	Attitudes to the voices of all the male agents would be similar; attitudes to the voices of all female agents would be similar.
	4.3	Within agent types male and female agents would be rated similarly; between agent type differences might occur.
Dependent Variables		Attitude Questionnaire Responses (1-7 Likert scale) Agent Ratings (1-10 Scale)
Other Data	1	Interview Answers
	2	2 Focus Groups (Part I and Part II)
(Experiment) Independent Variables:	1	Agent Type (5 levels)
	2	Agent Gender (2 levels)
	3	Application (2 levels)
(Participant) Independent Variables	1	Gender (2 levels)
	2	Age Group (3 levels)
Extraneous Variables:	Presentation Order	Agent presentation order randomised.
	Equipment Differences	Controlled by using exactly the same apparatus in both evaluation parts.
	Physical Layout	Controlled by using the exact same room in the building for both evaluation parts.
Location		Edinburgh - CCIR Premises, Central Edinburgh
Cohort		N = 64: Part I (32); Part II: (32) 50% male, 50% female
Remuneration		£10
Duration:		50 minutes

Table 4.6 Summary Table of Passive Viewing Evaluation I

In summary, the experiment was divided into two sections, firstly an evaluation of five humanoid photo-realistic agent types in a Home Furnishings Service. At a later date, with a second participant sample the same humanoid photo-realistic agent types were

evaluated in a personalised CD Service. Participants completed questionnaires relating to each individual agent and they also took part in closing interviews and focus groups.

4.8 Results

To allow a comprehensive assessment and comparison of the humanoid photo-realistic agents in the two contrasting retail applications the quantitative data gathered in both parts of the evaluation was combined. For each usability attribute, and also for the agent ratings a series of repeated measures ANOVAs, taking agent gender, agent type, application, participant gender and age as the independent variables, were completed. Each ANOVA table presents the results for within and between subject effects, together with interactions and results for the participant between-subject variables of age and gender. An explanation and discussion of significant and non-significant results is presented below. Table 4.7 describes the abbreviations needed to understand the ANOVA tables.

Abbreviation	Explanation
A(Type)	Agent Type
A(Gender)	Agent Gender
P(Type)	Participant Type
P(Gender)	Participant Gender

Table 4.7 Explanation of Abbreviations Used in ANOVA Tables

4.8.1 Agent Ratings

Participants were asked to rate each agent on a scale of 1 to 10, where 1 was the lowest rating and 10 was the highest. There was a highly significant main effect for agent type (Table 4.8). The mean rating scores are given in Table 4.9 together with the results of pair-wise comparisons. The results show that the video agent type (H1) was rated higher than all other agent types. Following this the disembodied agent type (H5) was rated higher than H2, H3 and H4. The 3D talking head agent (H2) was rated lowest.

	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	1014.387	4	253.597	61.48	**.000**
A(Type) * Application	8.243	4	2.061	.324	.862
A(Type) * P(Age)	17.167	8	2.146	.337	.951
A(Type) * P(Gender)	15.114	4	3.779	.594	.668
Error(A(Type))	1324.080	208	6.366		
A(Gender)	3.678	1	3.678	1.865	.178
A(Gender) * Application	.534	1	.534	.271	.605
A(Gender) * P(Age)	.191	2	9.536E-02	.048	.953
A(Gender) * P(Gender)	2.678	1	2.678	1.358	.249
Error(A(Gender))	102.520	52	1.972		
A(Gender) * A(Type)	6.772	4	1.693	1.417	.229
A(Gender) * A(Type) * Application	8.859	4	2.215	1.854	.120
Error(A(Gender)*A(Type))	248.488	208	1.195		
Between Subject Effects					
Application	6.470	1	6.470	.461	.500
P(Age)	24.092	2	12.046	.859	.430
P(Gender)	56.063	1	56.063	3.996	.051
Application * P(Age)	3.079	2	1.539	.110	.896
Application * P(Gender)	8.999	1	8.999	.641	.427
Error	729.622	52	14.031		

Table 4.8 Ratings ANOVA

Humanoid Photo-Realistic Agent Types	Mean Rating Score (max 10)	Type rated better than (all $p < 0.01$)
H1 (Video)	7.25	H2, H3, H4, H5
H2 (3D talking head)	3.05	-
H3 (Image w/ facial moves)	4.27	H2
H4 (Still image)	4.70	H2
H5 (Disembodied voice)	5.85	H2, H3, H4

Table 4.9 Agent Ratings Mean Scores

4.8.2 Attitude to Applications

To provide a general impression of the results for this section, the mean scores for the independent variables of agent type, agent gender and application are presented in Figures 4.10 (i), (ii) and (iii).

4.8.2.1 Usability Attribute – "Good idea"

I think this service is a good idea	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	13.903	4	3.476	3.726	**.034**
A(Type) * Application	7.744	4	1.936	4.221	**.014**
A(Type) * P(Age)	4.494	8	.562	.431	.902
A(Type) * P(Gender)	6.571	4	1.643	1.260	.287
Error(A(Type))	260.730	200	1.304		
A(Gender)	2.896	1	2.896	2.748	.104
A(Gender) * Application	2.490	1	2.490	2.362	.131
A(Gender) * P(Age)	2.490	1	2.490	2.362	.131
A(Gender) * P(Gender)	.774	1	.774	.735	.395
Error(A(Gender))	52.704	50	1.054		
A(Gender) * A(Type)	5.160	4	1.290	1.773	.136
A(Gender) * A(Type) * Application	1.690	4	.422	.581	.677
A(Gender) * A(Type) * P(Age)	4.308	8	.539	.740	.656
A(Gender) * A(Type) * P(Gender)	2.060	4	.515	.708	.587
Error(A(Gender)*A(Type))	145.488	200	.727		
Between Subject Effects					
Application	7.397	1	7.397	.791	.378
P(Age)	6.104	2	3.052	.326	.723
P(Gender)	3.047	1	3.047	.326	.571
Application * P(Age)	1.055	2	.527	.056	.945
Application * P(Gender)	22.780	1	22.780	2.436	.125
Error	467.595	50	9.352		

Table 4.10 ANOVA for Usability Attribute "Good idea"

The ANOVA (Table 4.10) for the usability attribute "I think this service is a good idea" showed that there were significant effects with respect to agent type, which follows the trend set by the agent ratings where participants felt the applications were better with the video agents (H1). In fact pair-wise comparisons in the form of t-tests showed that the applications with the video agents (H1) produced significantly higher results than the applications with H2 ($p < 0.05$), H3 ($p < 0.01$) and H4 ($p < 0.05$). The mean results are shown in Table 4.11.

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	5.69
H2 (3D talking head)	5.19
H3 (Image w/ facial moves.)	5.39
H4 (Still image)	5.27
H5 (Disembodied voice)	5.44

Table 4.11 Usability Attribute "Good Idea"
Mean Scores by Agent Type

An interaction between the two main variables of agent type and application also emerged. This interaction suggested differences in attitude to the agents between the applications. In fact there was a much lower mean score for the still image agent (H4) in the CD Service than the Home Furnishings service. The mean results for agent type with respect to both applications are given in Table 4.12.

Humanoid Photo-Realistic Agent Types	Home Furnishings	CD Service
H1 (Video)	5.53	5.73
H2 (3D talking head)	5.32	5.13
H3 (Image w/ facial moves.)	5.53	5.15
H4 (Still image)	5.57	5.03
H5 (Disembodied voice)	5.57	5.30

Table 4.12 Usability Attribute "Good Idea"
Mean Scores by Agent Type and Application

4.8.2.2 Usability Attribute – "Ease of use"

Analysis showed that participants did not consider that they would find either application difficult to use if they were given the opportunity to use them interactively (grand mean = 5.58). The ANOVA table (Table 4.13) shows no significant effects for the repeated measures variables of agent type, agent gender or for the between subject variable of application. No effects or interactions for participant age or gender were evident.

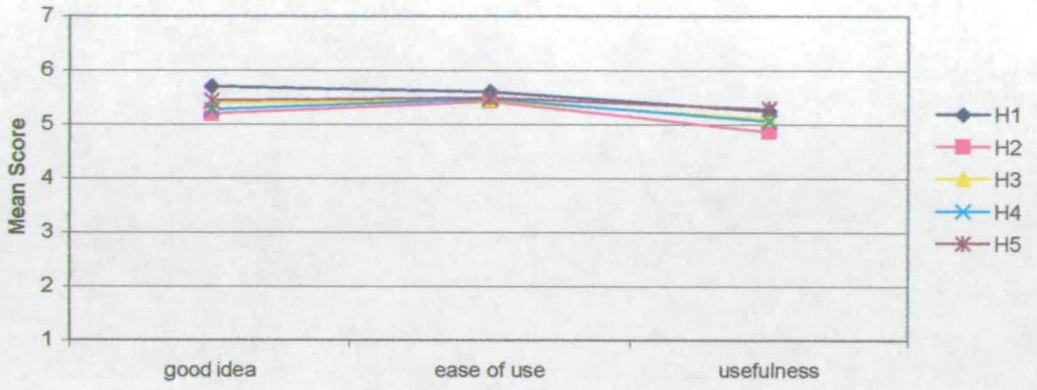


Figure 4.10(i) Usability Attributes for Application by Agent Type

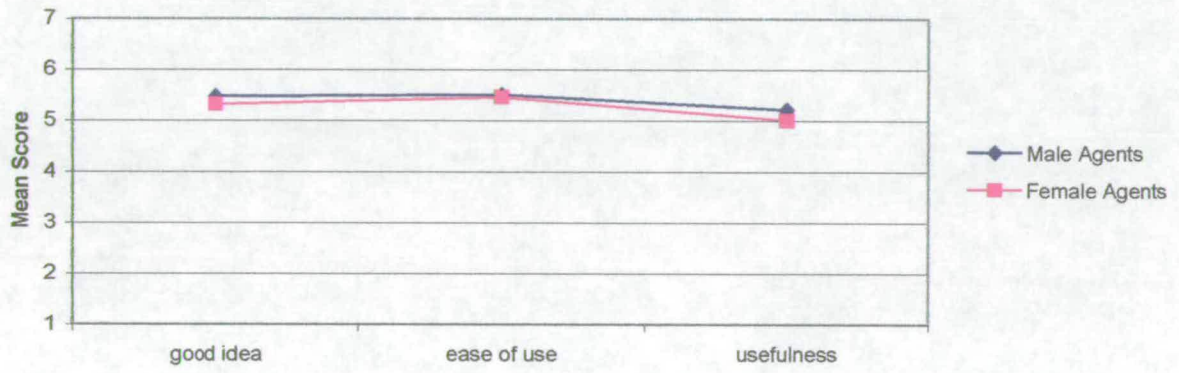


Figure 4.10(ii) Usability Attributes for Application by Agent Gender

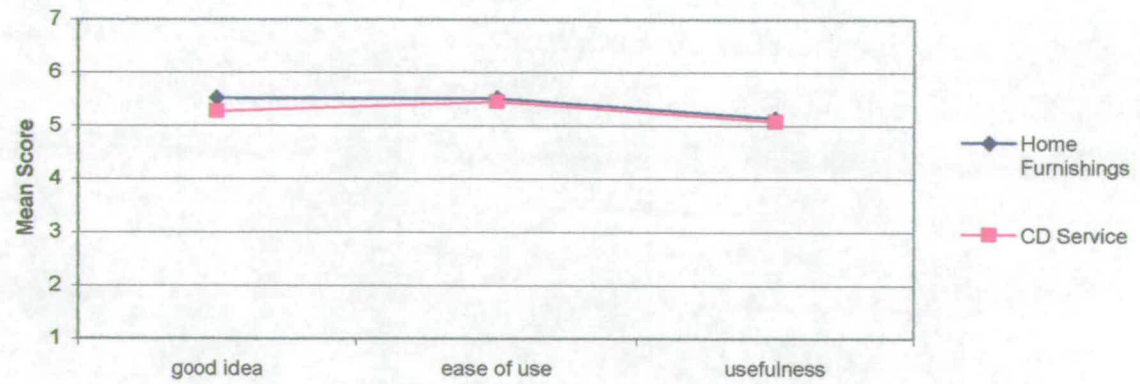


Figure 4.10(iii) Usability Attributes for Application by Application

I think this service would be difficult to use	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	1.852	4	.463	.788	.534
A(Type) * Application	.376	4	9.403E-02	.160	.958
A(Type) * P(Age)	1.388	8	.174	.295	.967
A(Type) * P(Gender)	1.105	4	.276	.470	.758
Error(A(Type))	117.511	200	.588		
A(Gender)	.285	1	.285	.658	.421
A(Gender) * Application	.169	1	.169	.390	.535
A(Gender) * P(Age)	.743	2	.372	.858	.430
A(Gender) * P(Gender)	.224	1	.224	.518	.475
Error(A(Gender))	21.646	50	.433		
A(Gender) * A(Type)	4.021	4	1.005	1.440	.222
A(Gender) * A(Type) * Application	1.182	4	.295	.423	.792
Error(A(Gender)*A(Type))	139.645	200	.698		
Between Subject Effects					
Application	.492	1	.492	.040	.842
P(Age)	32.398	2	16.199	1.320	.276
P(Gender)	16.723	1	16.723	1.363	.249
Application * P(Age)	15.798	2	7.899	.644	.530
Application * P(Gender)	1.314	1	1.314	.107	.745
Error	613.537	50	12.271		

Table 4.13 ANOVA for Usability Attribute "Ease of use"

4.8.2.3 Usability Attribute – "Usefulness"

The results for this attribute followed the overall trend and significant differences (Table 4.14) between agent types showed that participants would prefer to use the applications if the agent was a video (H1) or a disembodied voice (H5). Pair-wise comparisons indicated they would significantly use the applications with a video instead of a 3D talking head agent ($p < 0.05$). The mean scores are provided in Table 4.15.

The ANOVA table (Table 4.14) also shows highly significant effects for agent gender and Figure 4.10(ii) shows that participants would prefer to use the applications if the agents were male. Later in this chapter it will be shown that there was a significantly negative attitude to the voice used for the female agents. This occurrence is used to explain the reason why participants would rather use the applications with the male agents in the evaluation.

I would like to use this service myself	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	10.718	4	2.679	3.203	.014
A(Type) * Application	3.148	4	.787	.941	.441
A(Type) * P(Age)	7.858	8	.982	1.174	.316
A(Type) * P(Gender)	1.187	4	.297	.355	.841
Error(A(Type))	8.203	200	.836		
A(Gender)	5.289	1	5.289	4.108	.002
A(Gender) * Application	3.567	1	3.567	3.773	.058
A(Gender) * P(Age)	3.405	2	1.703	1.801	.176
A(Gender) * P(Gender)	.701	1	.701	.741	.393
Error(A(Gender))	47.278	50	.946		
A(Gender) * A(Type)	.420	4	.105	.147	.964
A(Gender) * A(Type) * Application	3.435	4	.859	1.202	.311
Error(A(Gender)*A(Type))	142.830	200	.714		
Between Subject Effects					
Application	1.177	1	1.177	.082	.776
P(Age)	14.723	2	7.361	.514	.601
P(Gender)	.504	1	.504	.035	.852
Application * P(Age)	.998	2	.499	.035	.966
Application * P(Gender)	21.127	1	21.127	1.474	.230
Error	716.492	50	14.330		

Table 4.14 ANOVA for Usability Attribute "Usefulness"

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	5.24
H2 (3D talking head)	4.86
H3 (Image w/ facial moves.)	5.10
H4 (Still image)	5.06
H5 (Disembodied voice)	5.28

Table 4.15 Usability Attribute "Usefulness"
Mean Scores by Agent Type

The results so far indicate that participants believe both services would be easy to use could they use them interactively. In addition participants thought the applications were good ideas and useful, however following the trend set out by the agent ratings, there is a preference for applications that are inhabited with the conversational agents visually represented by a video agent (H1).

4.8.3 Attitude to Voices & Conversation

The mean scores for this section particularly concerned with attitudes to the agents' voices are provided graphically in the Figures 4.11(i), (ii) and (iii).

4.8.3.1 Usability Attribute – "Voice clarity"

The assistant's voice was not clear enough	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	1.852	4	.463	3.79	**.021**
A(Type) * Application	.376	4	9.403E-02	.160	.958
A(Type) * P(Age)	1.388	8	.174	.295	.967
A(Type) * P(Gender)	1.105	4	.276	.470	.758
Error(A(Type))	117.511	200	.588		
A(Gender)	.285	1	.285	.658	.421
A(Gender) * Application	.169	1	.169	.390	.535
A(Gender) * P(Age)	.743	2	.372	.858	.430
A(Gender) * P(Gender)	.224	1	.224	.518	.475
Error(A(Gender))	21.646	50	.433		
A(Gender) * A(Type)	4.021	4	1.005	1.440	.222
A(Gender) * A(Type) * Application	1.182	4	.295	.423	.792
Error(A(Gender)*A(Type))	139.645	200	.698		
Between Subject Effects					
Application	.492	1	.492	.040	.842
P(Age)	32.398	2	16.199	1.320	.276
P(Gender)	16.723	1	16.723	1.363	.249
Application * P(Age)	15.798	2	7.899	.644	.530
Application * P(Gender)	1.314	1	1.314	.107	.745
Error	613.537	50	12.271		

Table 4.16 ANOVA for Usability Attribute "Voice clarity"

Significant effects (Table 4.16) for agent type emerged when participants were asked about the clarity of the agent's voice, suggesting cross-modal audio-visual effects. Even though the voices for all the female and male agents were identical for respective agent gender, participants had varying attitudes to the clarity of the voices for the different agent types. The video agents (H1) and the disembodied voice agents (H5) were rated similarly. Pair-wise comparisons show that the voices of H5 were judged to be significantly clearer than H2, H3 and H4 (all at $p < 0.01$) and the voices of H1 were

significantly clearer than H2 and H4, (both at $p < 0.01$) and H3, ($p < 0.05$). The table of means is shown in Table 4.17.

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	4.80
H2 (3D talking head)	4.29
H3 (Image w/ facial moves.)	4.52
H4 (Still image)	4.39
H5 (Disembodied voice)	4.86

Table 4.17 Usability Attribute "Voice clarity" Mean Scores by Agent Type

4.8.3.2 Usability Attribute – "Voice liked"

I liked the assistant's voice	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	23.026	4	5.757	3.777	**.000**
A(Type) * Application	13.816	4	3.454	3.574	**.008**
A(Type) * P(Age)	10.061	8	1.258	1.301	.244
A(Type) * P(Gender)	5.897	4	1.474	1.526	.196
Error(A(Type))	197.129	204	.966		
A(Gender)	14.336	1	14.336	7.26	.003
A(Gender) * Application	5.658	1	5.658	1.885	.176
A(Gender) * P(Age)	.979	2	.489	.163	.850
A(Gender) * P(Gender)	.000	1	.000	.000	1.000
Error(A(Gender))	153.067	51	3.001		
A(Gender) * A(Type)	4.105	4	1.026	1.184	.319
A(Gender) * A(Type) * Application	4.587	4	1.147	1.323	.263
Error(A(Gender)*A(Type))	176.868	204	.867		
Between Subject Effects					
Application	29.478	1	29.478	1.828	.182
P(Age)	2.744	2	1.372	.085	.919
P(Gender)	14.950	1	14.950	.927	.340
Application * P(Age)	147.892	2	73.946	4.586	.055
Application * P(Gender)	57.232	1	57.232	3.549	.065
Error	822.329	51	16.124		

Table 4.18 ANOVA for Usability Attribute "Voice liked"

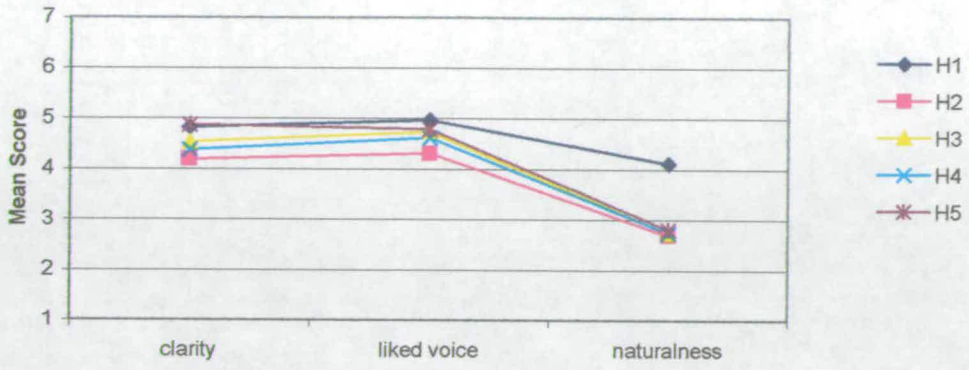


Figure 4.11(i) Usability Attributes for Agents' Voice by Agent Type

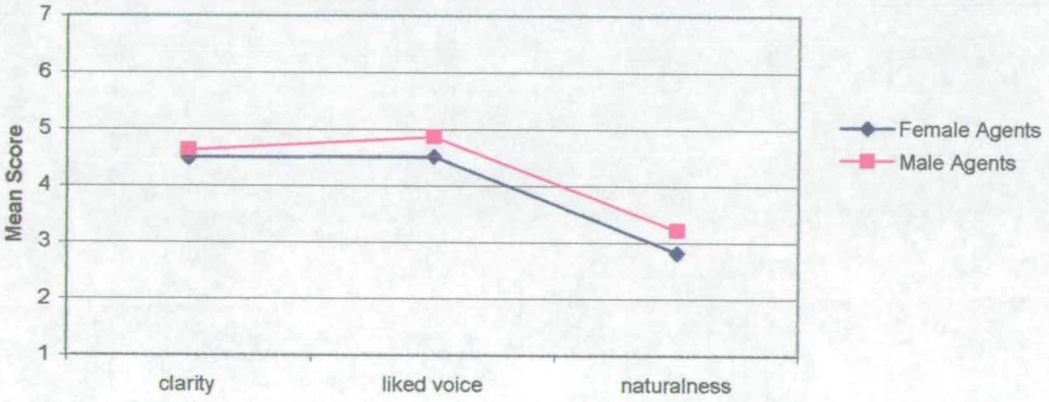


Figure 4.11(ii) Usability Attributes for Agents' Voice by Agent Type Gender

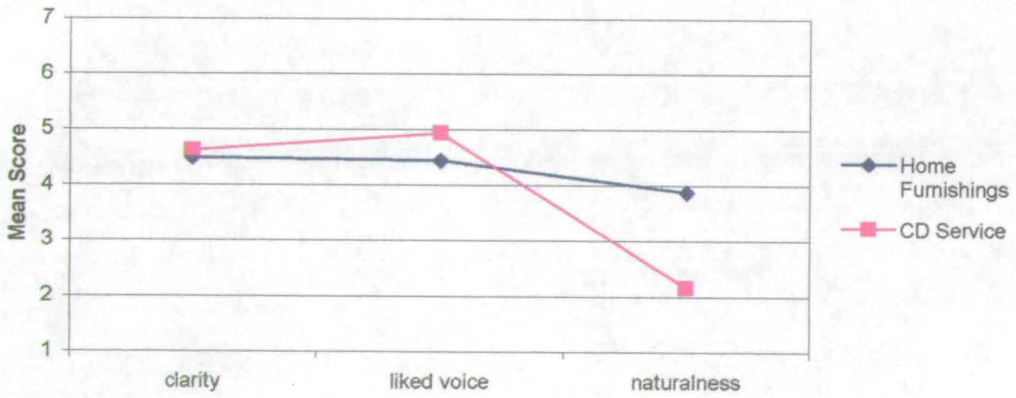


Figure 4.11(iii) Usability Attributes for Agents' Voice by Application

This attribute addressed whether participants liked the voice and a highly significant effect emerged for agent type (Table 4.18). The mean results are presented in Table 4.19. Pair-comparisons showed that the voice of the 3D talking head agent type (H2) was significantly less liked than the video agent type (H1) and the disembodied voice agent type (H5), both at $p < 0.05$. This indicates a further emergence of cross-modal effects, because as explained earlier the agents' voices were identical for the respective gender and the agents actually only differed in their physical realisation in the interface.

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	4.91
H2 (3D talking head)	4.59
H3 (Image w/ facial moves.)	4.69
H4 (Still image)	4.68
H5 (Disembodied voice)	4.82

Table 4.19 Usability Attribute "Voice liked"
Mean Scores by Agent Type

An interaction between the main variables of agent type and application also emerged. The mean results (Table 4.20) indicate, and pair-wise comparisons confirm that the significant differences occurred in the first evaluation (Home Furnishings), where the voice of the video agent type (H1) was significantly preferred over H2, H3 and H4, all at $p < 0.01$. These effects were weaker in the second part of the evaluation.

Humanoid Photo-Realistic Agent Types	Home Furnishings	CD Service
H1 (Video)	4.98	4.96
H2 (3D talking head)	3.79	4.83
H3 (Image w/ facial moves.)	4.53	4.96
H4 (Still image)	4.33	4.92
H5 (Disembodied voice)	4.56	5.02

Table 4.20 Usability Attribute "Voice liked"
Mean Scores by Agent Type and Application

There was also a significant effect for agent gender and Figure 4.11(ii) shows that the female voice was not liked as much as the male voice. Participants explained that the female voice had a more distinctive accent than the male voice, which did not appeal to them. As mentioned in the presentation of the results for the attribute of 'Usefulness',

the significant gender difference with respect to voice is offered as an explanation as to why results indicated that participants also preferred the applications with a male agent.

4.8.3.3 Usability Attribute – "Naturalness of conversation"

I felt the conversation was unnatural	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	144.120	4	36.030	22.55	**.000**
A(Type) * Application	82.286	4	20.572	18.485	**.000**
A(Type) * P(Age)	8.544	8	1.068	.960	.469
A(Type) * P(Gender)	2.733	4	.683	.614	.653
Error(A(Type))	222.576	200	1.113		
A(Gender)	18.404	1	18.404	8.05	**.004**
A(Gender) * Application	3.504E-02	1	3.504E-02	.017	.896
A(Gender) * P(Age)	7.799	2	3.900	1.914	.158
A(Gender) * P(Gender)	2.666	1	2.666	1.308	.258
Error(A(Gender))	101.871	50	2.037		
A(Gender) * A(Type)	13.834	4	3.458	4.575	**.001**
A(Gender) * A(Type) * Application	23.103	4	5.776	7.641	**.000**
Error(A(Gender)*A(Type))	151.176	200	.756		
Between Subject Effects					
Application	351.272	1	351.272	37.65	**.000**
P(Age)	32.947	2	16.474	1.433	.248
P(Gender)	1.875E-02	1	1.875E-02	.002	.968
Application * P(Age)	20.924	2	10.462	.910	.409
Application * P(Gender)	1.523	1	1.523	.132	.717
Error	574.748	50	11.495		

Table 4.21 ANOVA for Usability Attribute "Naturalness of conversation"

For this usability attribute there were highly significant effects for the three main variables (Table 4.21). The results showed that overall the conversation with the video agents (H1) was felt to be most natural and was significantly more natural that H2, H3, H4 and H5, all at $p < 0.01$. In addition to this a highly significant interaction between agent type and application showed that this was particularly the case in the CD Service, the mean results are presented in Table 4.22.

The conversation with the male and female agents was not thought to be natural, (mean female = 2.82, mean male = 3.14). In fact it was significantly poorer for the female agents. Participants did not think that the conversations in either the Home Furnishings

Service or the CD Service were natural (mean Home Furnishings = 3.87, mean CD Service = 2.15), this is confirmed later in the analysis of the qualitative findings, where participants commented on the repetitive dialogue being monotonous and this is offered as an explanation for the low scores of this usability attribute with respect to the applications. The conversation in the CD Service was significantly less natural, and again in the qualitative findings it was disclosed that participants were not familiar with the musical tracks chosen in the dialogue and this could have caused the poorer score in the CD Service.

Humanoid Photo-Realistic Agent Types	Home Furnishings	CD Service
H1 (Video)	4.19	4.04
H2 (3D talking head)	3.53	1.86
H3 (Image w/ facial moves.)	3.55	1.72
H4 (Still image)	3.83	1.64
H5 (Disembodied voice)	4.07	1.53

Table 4.22 Usability Attribute "Naturalness of conversation" Mean Scores by Agent Type and Application

The ANOVA table also displayed highly significant interactions for agent gender and agent type and a second significant interaction between all three variables of agent gender, agent type and application is reported. The mean results, presented in Table 4.23, show similar trends for agent type in both applications, with the conversation with video agent (H1) being preferred in both applications. However the differences between H1 and all other agents in the CD Service were more obvious. Within agent types, agent gender differences also occurred and there is much stronger evidence in the CD Service that suggests that the conversation with the male video agent type was more natural than all other agents.

	H1 (Video)		H2 (3D talking head)		H3 (Image w/ facial moves)		H4 (Still image)		H5 (Disembodied voice)	
	F	M	F	M	F	M	F	M	F	M
Mean Score Home Furn.	3.94	4.19	3.36	3.74	3.87	3.55	3.45	3.87	3.93	4.26
Mean Score CD Service	3.09	5.00	1.87	1.87	1.54	1.87	1.64	1.65	1.58	1.48

Table 4.23 Usability Attribute "Naturalness of conversation" Mean Scores by Agent Type, Agent Gender and Application

In summary, the main findings with respect to participants' attitudes to the agents' voices, show that both the video agents (H1) and the disembodied voices (H5) were thought to be the clearer, and were also both liked more. The video agent (H1) was also more natural than all other agents, including H5. Overall the voices of the male agents were preferred to the female agents, and the conversation with the male agents was also thought to more natural.

4.8.4 Attitude to Competence & Friendliness

To investigate differences that may occur between agent types (Prediction 4.3) usability attributes relating to the agents' competence and friendliness were included. As it is expected that assistants in retail spaces should be competent and friendly and it is investigated if there are differences based on agent type, gender or application.

4.8.4.1 Usability Attribute – "Competence"

I felt the assistant was competent	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	15.831	4	3.958	2.860	**.025**
A(Type) * Application	10.392	4	2.598	1.877	.116
A(Type) * P(Age)	19.372	8	2.421	1.750	.089
A(Type) * P(Gender)	1.377	4	.344	.249	.910
Error(A(Type))	282.319	204	1.384		
A(Gender)	1.493	1	1.493	.713	.402
A(Gender) * Application	1.214	1	1.214	.580	.450
A(Gender) * P(Age)	3.665	2	1.833	.875	.423
A(Gender) * P(Gender)	3.922	1	3.922	1.872	.177
Error(A(Gender))	106.851	51	2.095		
A(Gender) * A(Type)	1.598	4	.400	.376	.826
A(Gender) * A(Type) * Application	1.453	4	.363	.342	.849
Error(A(Gender)*A(Type))	216.694	204	1.062		
Between Subject Effects					
Application	49.845	1	49.845	5.816	**.020**
P(Age)	11.257	2	5.628	.657	.523
P(Gender)	8.608	1	8.608	1.004	.321
Application * P(Age)	12.391	2	6.196	.723	.490
Application * P(Gender)	2.673	1	2.673	.312	.579
Error	437.085	51	8.570		

Table 4.24 ANOVA for Usability Attribute "Competence"

A significant effect (Table 4.24) for agent type emerged with respect to attitude to agent competence. Mean scores (Table 4.25) and pair-wise comparisons indicate that the video agent (H1) and disembodied voice agent (H5) were the most competent and significantly more competent than the 3D talking head agent (H2), both at $p < 0.01$.

A significant effect also emerged for application and the agents in the Home Furnishings Service were thought to be more competent than those in the CD Service (mean for Home Furnishings = 5.23; mean for CD Service = 4.72). As mentioned in the discussion of the attribute of 'Naturalness of conversation', this could have been due to unfamiliarity of the selection of musical tracks made in the dialogue.

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	5.16
H2 (3D talking head)	4.84
H3 (Image w/ facial moves.)	4.89
H4 (Still image)	4.92
H5 (Disembodied voice)	5.32

Table 4.25 Usability Attribute "Competence" Mean Scores by Agent Type and Application

4.8.4.2 Usability Attribute – "Friendliness"

Highly significant main effects for agent type, agent gender and application emerged with respect to the perceived friendliness of the assistants (Table 4.26). The mean results for agent type are given in Table 4.27 and pair-wise comparisons showed that the video agent (H1) was significantly friendlier than H2, H3 and H4 (at all $p < 0.01$). The disembodied voice agent (H5) was perceived to be as friendly as H1. There was also a significant result for agent gender. The male agents were significantly friendlier than the female (mean female = 4.20, mean male = 4.74). It will become more evident in the discussion of the qualitative data that participants felt the female voice annoying and as a result less friendly. A significant effect for application showed that the agents in the Home Furnishings application were friendlier. A further interaction between these two main variables showed an effect between agent gender and application indicating the female agents in the CD Service were significantly less friendlier than those in the Home Furnishings Service. The mean results for agent gender and application are displayed in Table 4.28.

I felt the assistant seemed unfriendly	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	18.505	4	4.626	3.879	**.006**
A(Type) * Application	9.341	4	2.335	1.893	.113
A(Type) * P(Age)	10.171	8	1.271	1.030	.414
A(Type) * P(Gender)	5.001	4	1.250	1.013	.402
Error(A(Type))	241.857	196	1.234		
A(Gender)	30.755	1	30.755	11.21	**.008**
A(Gender) * Application	34.656	1	34.656	12.36	**.005**
A(Gender) * P(Age)	4.173	2	2.087	.512	.602
A(Gender) * P(Gender)	.194	1	.194	.047	.828
Error(A(Gender))	199.697	49	4.075		
A(Gender) * A(Type)	2.410	4	.602	.458	.767
A(Gender) * A(Type) * Application	3.322	4	.831	.631	.641
Error(A(Gender)*A(Type))	258.039	196	1.317		
Between Subject Effects					
Application	120.061	1	120.061	11.72	**.004**
P(Age)	14.259	2	7.130	.545	.584
P(Gender)	11.099	1	11.099	.848	.362
Application * P(Age)	15.951	2	7.976	.609	.548
Application * P(Gender)	.183	1	.183	.014	.906
Error	641.536	49	13.093		

Table 4.26 ANOVA for Usability Attribute "Friendliness"

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	4.72
H2 (3D talking head)	4.23
H3 (Image w/ facial moves.)	4.35
H4 (Still image)	4.40
H5 (Disembodied voice)	4.63

Table 4.27 Usability Attribute "Friendliness"

Mean Scores by Agent Type

Humanoid Photo-Realistic Agent Gender	Home Furnishings	CD Service
Female Agents	5.00	3.44
Male Agents	4.97	4.50

Table 4.28 Usability Attribute "Friendliness"

Mean Scores by Agent Gender and Application

4.8.5 Attitude to Appearance

The mean results for usability attributes of agents' appearance are graphically illustrated for each of the three independent variables in Figure 4.12 (i), (ii) and (iii).

4.8.5.1 Usability Attribute – "Helpfulness"

Significant effects for agent type and an interaction between agent type and application emerged for 'Helpfulness' (Table 4.29). In the Home Furnishings Service seeing the video agent (H1) was significantly more helpful than seeing the 3D talking head (H2), $p < 0.01$. In the second application seeing the video agent (H1) was significantly more helpful than H2, H3 and H4, all at $p < 0.01$ (Table 4.30).

Being able to see the assistant was helpful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	107.803	3	35.934	15.86	**.000**
A(Type) * Application	31.783	3	10.594	5.84	**.005**
A(Type) * P(Age)	11.203	6	1.867	.773	.592
A(Type) * P(Gender)	8.474	3	2.825	1.170	.323
Error(A(Type))	376.661	156	2.414		
A(Gender)	2.793	1	2.793	1.811	.184
A(Gender) * Application	.128	1	.128	.083	.774
A(Gender) * P(Age)	6.976	2	3.488	2.262	.114
A(Gender) * P(Gender)	.465	1	.465	.302	.585
Error(A(Gender))	80.192	52	1.542		
A(Gender) * A(Type)	17.481	3	5.827	3.838	.051
A(Gender) * A(Type) * Application	2.854	3	.951	.627	.599
Error(A(Gender)*A(Type))	236.874	156	1.518		
Between Subject Effects					
Application	47.163	1	47.163	3.257	.077
P(Age)	14.072	2	7.036	.486	.618
P(Gender)	.397	1	.397	.027	.869
Application * P(Age)	6.523	2	3.261	.225	.799
Application * P(Gender)	9.898	1	9.898	.684	.412
Error	752.953	52	14.480		

Table 4.29 ANOVA for Usability Attribute "Helpfulness"

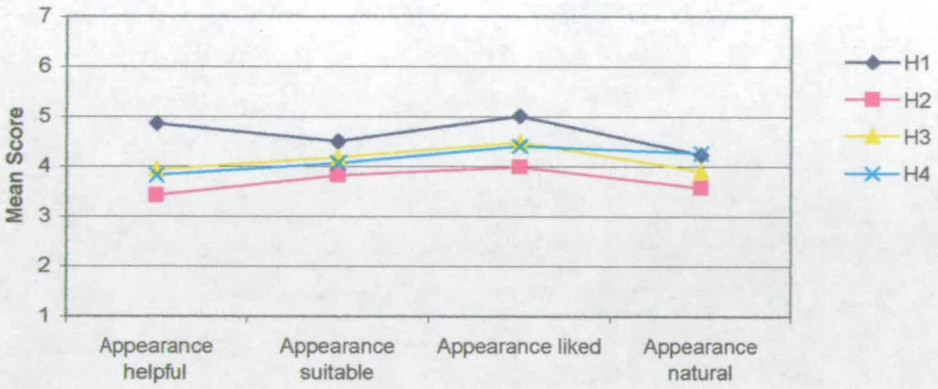


Figure 4.12(i) Usability Attributes for Agents' Appearance by Agent Type

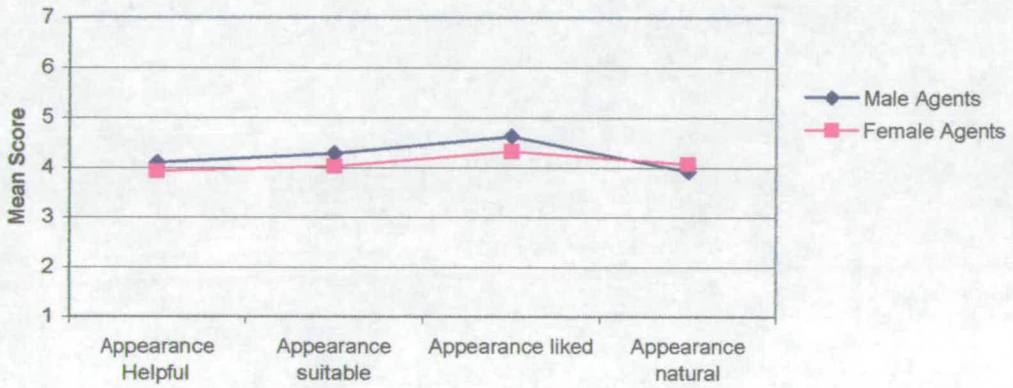


Figure 4.12(ii) Usability Attributes for Agents' Appearance by Agent Gender

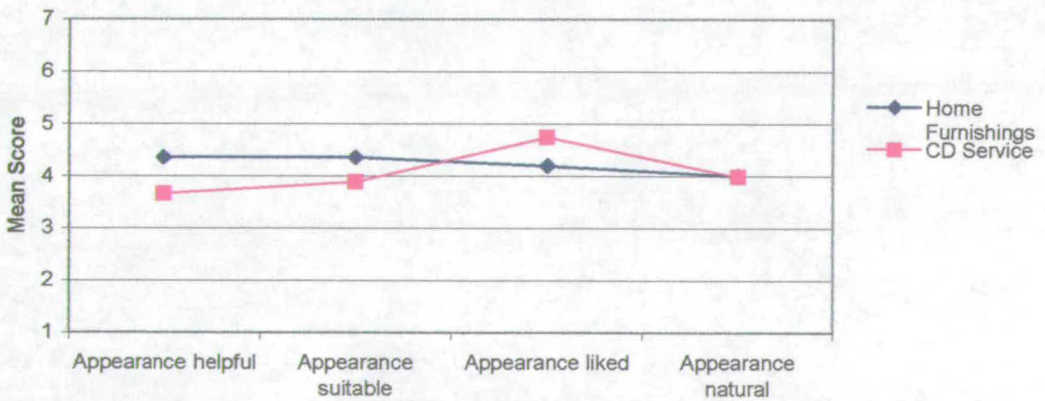


Figure 4.12(iii) Usability Attributes for Agents' Appearance by Application

Humanoid Photo-Realistic Agent Types	Home Furnishings Application 1	CD Service Application 2
H1 (Video)	4.81	4.90
H2 (3D talking head)	3.99	2.84
H3 (Image w/ facial moves.)	4.60	3.28
H4 (Still image)	4.03	3.61
H5 (Disembodied voice)	NA	NA

Table 4.30 Usability Attribute "Helpfulness"
Mean Scores by Agent Type and Application

4.8.5.2 Usability Attribute – "Appearance suitable"

The appearance of the assistant was unsuitable for the scenario	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	33.184	3	11.061	6.401	**.000**
A(Type) * Application	9.678	3	3.226	2.288	.081
A(Type) * P(Age)	5.631	6	.938	.666	.677
A(Type) * P(Gender)	.909	3	.303	.215	.886
Error(A(Type))	219.923	156	1.410		
A(Gender)	4.414	1	4.414	2.869	.096
A(Gender) * Application	1.120E-03	1	1.120E-03	.001	.979
A(Gender) * P(Age)	2.023	2	1.012	.658	.522
A(Gender) * P(Gender)	.113	1	.113	.073	.788
Error(A(Gender))	79.986	52	1.538		
A(Gender) * A(Type)	7.286	3	2.429	2.234	.086
A(Gender) * A(Type) * Application	6.982	3	2.327	2.141	.097
Error(A(Gender)*A(Type))	169.568	156	1.087		
Between Subject Effects					
Application	21.279	1	21.279	1.394	.243
P(Age)	45.005	2	22.503	1.474	.238
P(Gender)	29.035	1	29.035	1.902	.174
Application * P(Age)	13.410	2	6.705	.439	.647
Application * P(Gender)	11.872	1	11.872	.778	.382
Error	793.684	52	15.263		

Table 4.31 ANOVA for Usability Attribute "Appearance suitable"

The appearance of the video agent was perceived as being the most suitable agent type in both applications (Table 4.32). Pair-wise comparisons show this agent was significantly more suitable than H2 ($p < 0.01$), H3 and H4 ($p < 0.05$).

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	4.50
H2 (3D talking head)	3.82
H3 (Image w/ facial moves.)	4.18
H4 (Still image)	4.07
H5 (Disembodied voice)	NA

Table 4.32 Usability Attribute "Appearance suitable" Mean Scores by Agent Type

4.8.5.3 Usability Attribute – "Appearance Liked"

I liked the appearance of the assistant	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	50.707	3	16.902	12.80	**.000**
A(Type) * Application	21.999	3	7.333	4.59	**.000**
A(Type) * P(Age)	9.037	6	1.506	1.714	.121
A(Type) * P(Gender)	2.298	3	.766	.872	.457
Error(A(Type))	137.071	156	.879		
A(Gender)	9.112	1	9.112	11.98	**.003**
A(Gender) * Application	3.597	1	3.597	3.701	.060
A(Gender) * P(Age)	.392	2	.196	.202	.818
A(Gender) * P(Gender)	.665	1	.665	.685	.412
Error(A(Gender))	50.543	52	.972		
A(Gender) * A(Type)	3.715	3	1.238	1.386	.249
A(Gender) * A(Type) * Application	2.019	3	.673	.753	.522
Error(A(Gender)*A(Type))	139.372	156	.893		
Between Subject Effects					
Application	29.628	1	29.628	2.252	.140
P(Age)	5.372	2	2.686	.204	.816
P(Gender)	4.303E-02	1	4.303E-02	.003	.955
Application * P(Age)	65.334	2	32.667	2.483	.093
Application * P(Gender)	4.488	1	4.488	.341	.562
Error	684.163	52	13.157		

Table 4.33 ANOVA for Usability Attribute "Appearance liked"

The results for this usability attribute showed highly significant effects for agent type with respect to the appearance of the agents (Table 4.33). The mean scores are listed in Table 4.34. Following the overall trend, the video agents (H1) had the most popular

appearance in both applications. An additional significant interaction between agent type and application also emerged and pair-wise comparisons showed that the appearance of H1 was significantly better than the still image agent H4 ($p < 0.01$) in the Home Furnishings Service, and significantly better than H2, H3 and H4 (all at $p < 0.01$) in the CD Service. There was also a highly significant effect for agent gender where the male agents' appearances were preferred (mean female = 4.35, mean male = 4.65).

Humanoid Photo-Realistic Agent Types	Home Furnishings Application 1	CD Service Application 2
H1 (Video)	4.87	4.98
H2 (3D talking head)	3.58	4.64
H3 (Image w/ facial moves.)	4.23	4.75
H4 (Still image)	4.26	4.75
H5 (Disembodied voice)	NA	NA

Table 4.34 Usability Attribute "Appearance liked" Mean Scores by Agent Type and Application

4.8.5.4 Usability Attribute – "Appearance natural"

I thought the assistant looked natural	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	30.034	3	10.011	7.36	**.003**
A(Type) * Application	97.344	3	32.448	16.738	**.000**
A(Type) * P(Age)	18.441	6	3.074	1.500	.182
A(Type) * P(Gender)	20.381	3	6.794	3.316	**.022**
Error(A(Type))	307.307	150	2.049		
A(Gender)	2.108	1	2.108	.800	.375
A(Gender) * Application	.949	1	.949	.360	.551
A(Gender) * P(Age)	12.152	2	6.076	2.307	.110
A(Gender) * P(Gender)	4.709	1	4.709	1.788	.187
Error(A(Gender))	131.705	50	2.634		
A(Gender) * A(Type)	27.518	3	9.173	5.028	**.002**
A(Gender) * A(Type) * Application	1.039	3	.346	.190	.903
Error(A(Gender)*A(Type))	273.666	150	1.824		
Between Subject Effects					
Application	2.927E-02	1	2.927E-02	.004	.951
P(Age)	14.072	2	7.036	.904	.411
P(Gender)	7.392	1	7.392	.950	.334
Error	388.980	50	7.780		

Table 4.35 ANOVA for Usability Attribute "Appearance natural"

Once again highly significant results emerged for agent type when participants were asked if the assistant looked natural (Table 4.35). Pair-wise comparisons showed that the video agents (H1) were significantly more natural than H2, H3 and H4 (all at $p < 0.01$).

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	4.93
H2 (3D talking head)	4.08
H3 (Image w/ facial moves.)	4.49
H4 (Still image)	4.51
H5 (Disembodied voice)	NA

Table 4.36 Usability Attribute "Appearance natural" Mean Scores by Agent Type

There was also an interaction between agent type and application. Agents in the Home Furnishings Service followed the overall trend with the appearance of the video agent (H1) thought to be most natural. In contrast to this the CD Service results did not follow this trend, where the results for the still image were significantly higher than the other three agents, H1, H2 and H3, (all at $p < 0.01$). A tentative explanation for this result may impinge on the interface design of the CD Service (Table 4.37). It is suggested that confusion may have arisen because still images of both the selected artist and the agent appeared in the interface simultaneously. It is possible that the participants may have thought the agent looked natural in comparison to the other visual stimuli in the interface, whereas the usability attribute was aiming to ask about the naturalness of the agent alone.

Humanoid Photo-Realistic Agent Types	Home Furnishings	CD Service
H1 (Video)	4.82	3.50
H2 (3D talking head)	2.92	3.95
H3 (Image w/ facial moves.)	3.82	3.64
H4 (Still image)	3.83	4.78
H5 (Disembodied voice)	NA	NA

Table 4.37 Usability Attribute "Appearance natural" Mean Scores by Agent Type and Application

It is certain that the results for the still image agent (H4) do not agree with previous findings from the first usability attribute ("I think this service is a good idea"). For this

attribute, participants significantly felt that the presence of H4 in the interface was not such a good idea, in comparison to the presence of other agents. In addition the trend from other usability attributes showed more negative responses to the still image. Although for this particular usability attribute this trend was followed in the case of the Home Furnishings application, however it is again felt that the image of the artist in the interface could have caused equivocal results for this attribute in the CD application.

This section showed that the video agent (H1) was also thought to be more competent and friendlier than the presence of the other agents. Seeing H1 was also thought to be more helpful, their appearance was more suitable for the applications, and the appearance was also more natural.

4.8.6 Attitude to Facial Movement

4.8.6.1 Usability Attribute – "Lip-synchronisation"

I felt the speech didn't match the lips	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	44.554	2	22.277	11.997	**.000**
A(Type) * Application	35.180	2	17.590	9.473	**.000**
A(Type) * P(Age)	3.813	4	.953	.513	.726
A(Type) * P(Gender)	.312	2	.156	.084	.920
Error(A(Type))	185.689	100	1.857		
A(Gender)	.154	1	.154	.224	.638
A(Gender) * Application	.946	1	.946	1.380	.246
A(Gender) * P(Age)	1.973	2	.986	1.439	.247
A(Gender) * P(Gender)	1.994E-02	1	1.994E-02	.029	.865
Error(A(Gender))	34.278	50	.686		
A(Gender) * A(Type)	1.194	2	.597	.690	.504
A(Gender) * A(Type) * Application	2.313	2	1.156	1.337	.267
Error(A(Gender)*A(Type))	86.508	100	.865		
Between Subject Effects					
Application	43.211	1	43.211	15.171	**.000**
P(Age)	10.113	2	5.056	1.775	.180
P(Gender)	5.990	1	5.990	2.103	.153
Application * P(Age)	1.622	2	.811	.285	.753
Application * P(Gender)	5.538E-02	1	5.538E-02	.019	.890
Error	142.411	50	2.848		

Table 4.38 ANOVA for Usability Attribute "Lip-synchronisation"

Highly significant effects for agent type emerged when participants were asked if they felt the speech matched the lip movement (Table 4.38). The means (Table 4.39) for all three agent types with lip movement were below neutral, suggesting that participants did not think this lip synchronisation was good, even for the video agents, (H1).

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	2.79
H2 (3D talking head)	2.00
H3 (Image w/ facial moves.)	1.84
H4 (Still image)	NA
H5 (Disembodied voice)	NA

Table 4.39 Usability Attribute "Lip-synchronisation"
Mean Scores by Agent Type

A significant effect for application showed that the lip-synchronisation of the agents in the CD Service was much poorer than those in the Home Furnishings Service (mean Home Furnishings = 2.66; mean CD Service = 1.88). Extending from this a significant interaction between agent type and application showed that the difference between the agent types was much stronger in the Home Furnishings Service, where the lip movement of H1 was significantly better than the same agent type in the CD Service, $p <$ 0.01 (Table 4.40).

Humanoid Photo-Realistic Agent Types	Home Furnishings Application 1	CD Service Application 2
H1 (Video)	3.70	1.94
H2 (3D talking head)	2.19	1.96
H3 (Image w/ facial moves.)	2.08	1.74
H4 (Still image)	NA	NA
H5 (Disembodied voice)	NA	NA

Table 4.40 Usability Attribute "Lip- synchronisation"
Mean Scores by Agent Type and Application

4.8.6.2 Usability Attribute – "Noticed lip-movement"

As the mean results for all agents were below neutral this showed that participants didn't notice the lip movement of the agents, see Table 4.41.

Humanoid Photo-Realistic Agent Types	Mean Score (max 7)
H1 (Video)	3.28
H2 (3D talking head)	2.82
H3 (Image w/ facial moves.)	3.00
H4 (Still image)	NA
H5 (Disembodied voice)	NA

Table 4.41 Usability Attribute "Noticed lip-movement" Mean Scores by Agent Type

In fact a significant effect (Table 4.42) for agent type showed that participants noticed the lip movement of H2 and H3 less than the video agent (H1), both at $p < 0.05$.

I noticed the lips moving	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	10.983	2	5.492	4.554	**.013**
A(Type) * Application	5.366	2	2.683	2.225	.113
A(Type) * P(Age)	18.783	4	4.696	3.894	.065
A(Type) * P(Gender)	1.351	2	.676	.560	.573
A(Gender)	.303	1	.303	.210	.649
A(Gender) * Application	3.141	1	3.141	2.175	.146
A(Gender) * P(Age)	5.272	2	2.636	1.826	.171
A(Gender) * P(Gender)	1.942E-02	1	1.942E-02	.013	.908
Error(A(Gender))	75.080	52	1.444		
A(Gender) * A(Type)	7.689	2	3.845	3.399	.057
A(Gender) * A(Type) * Application	.269	2	.135	.119	.888
Error(A(Gender)*A(Type))	117.641	104	1.131		
Between Subject Effects					
Application	209.990	1	209.990	41.516	**.000**
P(Age)	7.006	2	3.503	.693	.505
P(Gender)	.967	1	.967	.191	.664
Application * P(Age)	1.685	2	.843	.167	.847
Application * P(Gender)	3.286	1	3.286	.650	.424
Error	263.017	52	5.058		

Table 4.42 ANOVA for Usability Attribute "Noticed lip-movement"

This result helps to explain further the negative effects displayed for H2 and H3 with respect to 'Lip-synchronisation', that is if the participants failed to notice the lip-movement it may have been more difficult to assess whether or not the lip-movement actually matched the speech.

4.8.7 Agent Ranking

During the interviews participants were shown ten second excerpts of the interactions between each of the agents and the customer, after which they were asked to rank the best three assistants. The results for both parts of the evaluation are separately presented below.

4.8.7.1 Application 1: Home Furnishings

Rank	H1 (Video)		H2 (3D talking head)		H3 (Image with facial moves)		H4 (Still image)		H5 (Disembodied voice)	
	F	M	F	M	F	M	F	M	F	M
1	6	13	0	1	0	1	0	0	6	5
2	11	5	0	0	1	2	1	3	3	6
3	5	5	1	0	4	5	2	6	0	3

Table 4.43 Participants Rankings for Agents in Home Furnishings

The video agents (H1) received the majority of the votes (19 votes) from the 32 participants, followed by the disembodied voice agents (H5) who received 11 votes. The 3D talking head agents (H2) received one vote, as did the image with facial movements (H3). The still image (H4) was not the first preference for any of the participants (Table 4.43).

A relationship between the first and second preferences for the male and female video agents (H1) and the disembodied voice agents (H5) was evident. Thirteen participants voted for the male video agent as a first preference. From this group of participants, eight voted for the female video agent as a second preference. By the same token, six participants voted for the female video agent first and four of these participants voted for the male video agent second.

With respect to the disembodied voice agents (H5) six participants voted for the female disembodied voice agent and of these participants four voted for the male disembodied voice agent as their second preference. Finally, three of the five participants who voted the male disembodied voice agent as their first preference, voted for the female disembodied voice a their second preference. In all cases over half of the second preference votes went to the opposite gender of the same technology. The correlations between the first and second preference votes are shown in more detail for agents H1 and

H5, in Table 4.44. The figures in parentheses show the number of votes that came from participants who voted for the opposite gender of the same agent type for their first preference.

Rank	H1 (Video)		H5 (Disembodied voice)	
	F	M	F	M
1	6	13	6	5
2	11(8)	5(4)	3(3)	6(4)

Table 4.44 First and Second Preferences Correlations

4.8.7.2 Application 2: CD Service

Rank	H1 (Video)		H2 (3D talking head)		H3 (Image with facial moves)		H4 (Still image)		H5 (Disembodied voice)	
	F	M	F	M	F	M	F	M	F	M
1	6	14	0	1	1	1	0	1	2	6
2	9	10	0	1	0	1	1	3	3	4
3	5	4	0	0	3	5	3	6	3	3

Table 4.45 Participants Rankings for Agents in CD Service

A similar although not identical result emerged for the agent rankings in the CD Service. The majority of first preference votes (20 in total) were given to the video agent (H1), followed by the 8 votes to the disembodied voices. H2 and H4 received one vote each and H3 received two votes. Similarly the relationship between first and second preference votes was for participants to give the second preference vote to the opposite gender of the same agent type (Table 4.45). This was specifically the case for the six participants who voted for the female video agent (H1) first, with all of these participants delegating their second preference vote to the male video agent (H1). Of the fourteen participants who voted for the male video agent (H1) first, nine voted for the female video agent (H1) second. With respect to the disembodied voice agents (H5), three of the six participants who voted for the male disembodied voice (H5) first voted for the female disembodied voice agent second and both participants who voted for the female disembodied voice agent first, voted the male disembodied voice agent second. Again the third preference votes were distributed amongst all agents, except the 3D talking head agents (H2). The correlations for H1 and H5 are shown in Table 4.46.

Rank	H1 (Video)		H5 (Disembodied voice)	
	F	M	F	M
1	6	14	2	6
2	9(9)	10(6)	3(3)	4(2)

Table 4.46 First and Second Preferences Correlations

If the majority of participants voted for the same gender but a different agent type in the second preference voting, it could have indicated a greater agent gender division. However, in both parts of the evaluation the swing for voting between first and second preferences went to the opposite gender of the same agent type. This indicates that participants were first concerned with agent type, and then agent gender. To summarise the results of the agent ratings, the main findings show that the video agents (H1) and the disembodied voices (H5) were most preferred. In addition to the quantitative data presented previously, qualitative results in the form of focus groups and interviews clarify the reasons why participants would prefer to interact with the video agents (H1) and the disembodied voice agents (H5).

4.8.8 Interview Feedback

Overall, positive comments were retrieved with respect to participants' views of the use of humanoid photo-realistic agents in e-retail applications. In both applications the majority of participants expressed confidence for the successful deployment of conversational agents in the future. Some comments made by the participants are given in Table 4.47.

Home Furnishings	CD Service
"Excellent idea" "I would definitely use it" "Ideal for shopping" "Nice for home shopping"	"Useful since it felt you were dealing with someone" "Good concept" "With some improvements I would use it" "More personal than just typing".

Table 4.47 Participants' Comments about the Applications

When asked if they felt the applications were enhanced by the visual presence of an agent, 36 of the 64 participants (56%) replied positively saying that they preferred to see an agent in the interface and these participants did express a desire to actually converse

with the agents interactively themselves. This mirrors the previous quantitative results, which produced a dichotomy between the video agents (H1) and the disembodied voices (H5). To improve the conversational agents most suggestions were for the agents to behave *"naturally"*. It was suggested that the interface may be improved by *"having a character instead of an actual person to avoid looking unnatural"*. Because the agents had photo-realistic visual appearances participants felt it was extremely important for the agents to have corresponding natural synchronised voice and lip movement.

Results from both evaluations indicated that participants favoured the use of video agents (H1) in the interfaces. Feedback confirmed a negative attitude towards the lack of facial expression of the other visible agents. One said that "*the assistant's appearance looked very unnatural because of minimal facial expressions*", and another participant stated that *"the faces lacked emotion"*. Participants also commented that the lip synchronisation of 3D talking head agents (H2) and the image with facial movements (H3) did not match the speech and this was distracting and annoying. If the visible agent did not have human-like facial expressions to match its photo-realistic human-like appearance it was thought to be distracting, and that disembodied voice agents (H5) would then be better to interact with in order to complete tasks in the applications.

All the participants who expressed preference for the disembodied voice agents (H5) commented that the picture of the sales assistant was distracting and that having to look at the picture distracted them from the visual changes that were happening in the virtual environment. It is important to note that no matter how sophisticated the appearance and behaviour of the agent type, if the interface demands too much attention from the user, conversational agents may not actually enhance the service.

In both evaluations the female voice received negative comments, further explaining the quantitative results. The female voice was described as: *"disinterested"*, *"the tone was over confident"*. These results indicated that it is important that conversational agents have voices that are liked by the user group and it is suggested that these voices should be fluent and conversational.

4.8.9 Focus Group Feedback

Approximately five days after the completion of each evaluation some participants returned to attend focus groups. Again, the participants in the focus group were selected to represent, as much as possible, gender and age (Table 4.48).

	Home Furnishings		CD Service	
Participant	**Male**	**Female**	**Male**	**Female**
Age 18-35	2	1	1	1
Age 36-49	1	1	1	1
Age 50+	1	1	1	1
	4	3	3	3

Table 4.48 Analysis of Participants by Gender and Age Group

Participants in the focus groups were introduced to each other and were reminded about the experiment in which they took part. The focus group discussion focused on (1) the application, (2) the agents, and (3) suggested improvements. With respect to the applications little new material, other than that already gathered in the interviews emerged, except for the fact that the repetitive dialogue between the customer and each of the ten agents was monotonous. To stimulate discussion about the agents, an excerpt of each agent type was shown to the participants and they were asked to comment on each. Overall, participants thought it was most important for the agents to behave as they appeared, but also to be polite and show competence.

4.8.9.1 H1 – Video

Participants commented that this agent type seemed to have *"more emotion"* in the voice, but the female agent did sound less natural as *"it was obvious she was reading from a script"*. With respect to the male agents, one participant in the first focus group (Home Furnishings) felt the voice of this agent was better than the other male agents, despite the fact that the agents were created with identical verbal output using the same male voice for all male agents. This cross-modal effect was evident from the quantitative results, and indicates that the visual appearance of the agent can affect the perception of the agent's voice.

4.8.9.2 H2 – 3D talking heads

Participants felt the 3D talking heads (H2) were *"terrible"*. In fact the female talking head looked distorted and *"dummy like"*. The female talking head agent *"distracted attention"* from the (home furnishings) application. Participants agreed that the male talking head agent had annoying head movements. The low ratings and poor acceptance of this agent type suggested that the technology used to create the 3D talking heads was underdeveloped.

4.8.9.3 H3 – Image with facial movements

In agreement with earlier results this agent type received few positive comments. In comparison to other agent types and gender participants suggested that the male assistant was not interested and was even seen as hostile. The female version was *"too unnatural"*.

4.8.9.4 H4 – Still image

Overall, this agent type (male and female) was distracting. Participants said that there did not seem to be any point in having a still image and they expected the face to move.

4.8.9.5 H5 – Disembodied voices

The disembodied voice agents prompted comments which suggested that the female assistant appeared to be *"more confident"* than the other female assistants. This is interesting considering the dialogue was identical in each video recording. Some members of the group would have preferred to see a face to match the voice.

Participants were also asked to make suggested improvements for agents and the applications. Those who favoured the disembodied voice agents over other agent types suggested that the user should have the facility to remove the picture of the assistant after a certain time if desired. There were also comments that the assistants should have smiled more. Poor lip-synchronisation was also highlighted as a problem, and overall it was found to be distracting. To avoid problems of lip-synchronisation interacting with cartoon characters was suggested by the participants in the first focus group (Home Furnishings). Finally it was felt that the assistant's appearance and voice lacked emotion

and it was pointed out that hand gestures may enhance emotive and expressive nature of the agents as assistants.

4.9 Discussion

Studies by Walker et al. (1994), Koda (1996) and Lester & Stone (1997) showed that agents with strong visual presence and facial expression can be more engaging and motivating for the user. Expanding from these results, the empirical evaluation reported in this chapter shows that certain photo-realistic humanoid agent types are liked more than others and the point at which the photo-realistic faces are disliked is sensitive, but centres on the verbal and non-verbal behaviour of the agents. Although Walker et al. and Koda also demonstrated that task performance was not negatively affected by the use of a face in the interface, the results in this chapter show that certain agent types can distract the user. This may not necessarily lead to poorer task performance in agreement with the results of Walker et al and Koda, but in support of the work by Takeuchi and Nagao (1993) there may be an interference in user concentration, and the findings in this chapter suggest that this may reduce the user's overall attitude to the agent as an assistant and may discourage further interactions with the agent by participants.

The claim (Prediction 4.1) that conversational agents would be liked in retail interfaces was supported when two participant groups received the Home Furnishings application and CD Service application positively. Qualitative and quantitative findings confirmed that participants thought both applications were good ideas, they would be easy to use and participants stated they would use the applications themselves. No outstanding application dependency issues emerged during the two evaluations of the humanoid photo-realistic agents in the applications and in fact the agents were rated similarly for both applications.

It should be noted that the content of the dialogue in the CD Service application did not appeal to all the participants, namely because they were unfamiliar with or disliked the musical tracks chosen by the 'customer'. This unavoidable design feature could provide plausible reasoning for the emergence of some additional application dependency issues. Despite this the overall trend of participants' attitudes toward both contrasting applications, as indicated by the qualitative and quantitative results from both parts of

the evaluation confirms the fact that the conversational capabilities of these retail applications were received positively, in agreement with the first experiment prediction.

The second prediction (Prediction 4.2) stated that no differences would emerge with respect to the voices of all the male agents and the voices of all the female agents as one male and one female voice was used for all agent types of the same gender respectively. This claim was not supported and a number of results were reported signalling significant effects for participants' attitudes toward the voices of the various agent types. Such results indicate the occurrence of cross-modal effects between the physical realisation of the agent in the interface and the perceived quality of the agent's speech output. Although research by Massaro (1998), Allwood et al. (1990) and McNeill (1997, 1992) has thoroughly indicated that cross-modal effects are common between the quality of facial visemes and the perception of the actual phoneme output, the results of this evaluation offer a different perspective to the occurrence of cross-modal effects and show that the visual realisation of the agent types can produce other highly significant cross-modal effects with respect to the quality of identical human-like voice outputs. Usability attributes relating to agents' voices showed that the participants preferred the voices of the video agents and the disembodied voice agents to the other agent types, despite the fact that the voices from the same agent gender were identical between the agent types. They also perceived the voices of these two agent types to be clearer and the conversation between the customer and these agents to be more natural. These are interesting cross-modal effects and show that regardless of the quality of the voices used for the agents' output, if the agent's visual appearance does not complement the verbal behaviour, attitude to the agents' voices will be significantly lower.

In addition to the issues addressed with respect to the voices between the different agent types, significant differences also emerged for agent gender. Although participants felt the voices of both the male and female agents were clear, it did emerge that participants felt the male voices were liked better and that the dialogue between the customer and the agent was more natural with the male agents than the female agents. The male agents were also thought to be friendlier. The qualitative data reiterated the fact the female agent's voice was not liked when a number of negative comments were recorded, for example: *"the female voice was annoying and seemed disinterested"*. This occurrence stresses the importance of selecting fluent, conversational voices that will appeal to the majority of the potential user group.

The third prediction (Prediction 4.3) stated that male and female agents of the same type would be rated similarly and overall this claim was supported with no other dominating significant differences occurring within agent types, except for differences with respect to voice which has been discussed previously. The male and female agents of all five agent types were rated similarly. However it was shown that the male agents were friendlier than their female counterparts (the qualitative data actually confirmed that this was most likely effect of the poor perception of the female voice).

Between the agent types it was predicted that differences might emerge for agent appearance (H1-H4 only) and in support of this claim there were significant preferences with respect to the appearance for the video agents (H1) and disembodied voice agents (H5) over the other agents in the evaluation. 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 experiment findings by Reeves and Nass (1996), who discovered participants prefer to interact with agents who have consistent personalities, which in itself reflects everyday human-like behaviour. The video agents were thought to be more competent and friendly, being able to see them was more helpful than the other visible agent types and overall their appearance was more natural. From an alternative perspective the results are also in agreement with Takeuchi and Nagao (1993) and show that in retail domains users also interpret the facial displays of the agents, but extending their results show that this is actually beneficial if the facial displays complement the appearance of the agent. Otherwise as concluded by both Takeuchi and Nagao (1993) and the results in this chapter, the agent is thought to be distracting and this can negatively alter users' concentration levels.

When the participants did not see such human-like behaviour they preferred not to see a visual display of the agent, and therefore preferred only a disembodied voice agent (H5). The popularity of the disembodied voice agents raises interesting issues about the need for visual representations of embodied conversational agents in e-retail interfaces. In addition to ensuring the agents' behaviour corresponds to the visual display, the application task should afford a visual display of an agent. Some participants commented that the image of the assistant distracted them from changes that were being made in the interface. Finally, it is therefore also necessary for interface designers to

assess application interfaces carefully in order to establish if a visual realisation of a humanoid photo-realistic agent is an actual enhancement and not a distraction.

4.10 Summary

This chapter presented an evaluation that aimed to assess the effectiveness of a range of photo-realistic humanoid agents in two contrasting retail applications. A technical description of the graphical user interface template design was given. This template was used to create two contrasting GUI retail applications. A technical description of the technology used to create five agent types was also provided. The retail applications in which the agents appeared were described and a series of images presented in the text assist in the description of the functionality of the applications and the agents.

Following this the experiment predictions, procedure and both quantitative and qualitative findings were documented. The overall results showed that when an agent is represented as a photo-realistic image, participants expect this agent to display sophisticated human-like verbal and non-verbal behaviour. In the absence of such a graphical realisation, participants still welcome the conversational interface as a way in which to complete tasks and selected the disembodied voice agent as the next most suitable agent type with which to interact when completing tasks.

The passive viewing methodology or evaluation by observation technique used to assess the agents was deemed to be a suitable for the experiment and successfully provided important empirical evidence about the deployment of photo-realistic agents in retail applications. Although the repetitive identical dialogues that the participants listened to in order to evaluate the agents were thought to be monotonous, the results overall are informative regarding agents represented as photo-realistic images.

In the chapter that follows a second passive viewing evaluation is documented. Some necessary alterations to the experiment design are made, for instance producing similar although not identical scripts for the interactions between the agents and the user. The second part of the array (Figure 1.1) of possible representations of humanoid agent types are designed, created and evaluated, namely the humanoid animated agents.

Chapter 5

Utilising the Retail Interface Template to Evaluate the Effectiveness of Humanoid Animated Agents

5.1 Introduction

The array of humanoid embodied conversational agents illustrated in Figure 1.2, spans from photo-realistic creations to humanoid animated agents. The deployment of such humanoid animated agents in interfaces is becoming more popular. Initially they were used in educational and tutoring domains (Johnson & Rickel, 2000), but more recently there is activity to suggest that such agents are popular methods as marketing tools to enhance commercial websites (Rist, 2001). However as with humanoid photo-realistic agents little empirical evidence is available yet as to the effectiveness of humanoid animated agents, and there is no evidence to support the successful use of these agents in the context of electronic retail domains.

In Chapter 4, it was documented that participants have high expectations toward the appearance and behaviour of *photo-realistic* agents and the verbal and non-verbal communication of the agents must complement their human-like photo-realistic appearance. To provide a comprehensive evaluation of ECA, as dictated by Figure 1.2, animated versions of the agents must also be considered. Extending the evaluation of possible humanoid embodied conversational agents, this chapter describes a second passive viewing evaluation in which a range of humanoid *animated* agents were selected, created and evaluated. The chosen range was selected to represent as much as possible the spread of humanoid animated agents that could be created to enhance interfaces in the future. The range of animated agents consisted of 2D and 3D talking heads and 2D and 3D fully embodied humanoids. Again, the gender of the agent is considered in the experiment and male and female versions of all agent types are created.

The evaluation described in this chapter used a similar experiment procedure to the passive viewing evaluation technique described in Chapter 4. The focus of the evaluation is the assessment of the participant's subjective experience while observing a range of interactions between a user (i.e. 'customer'), and the assistants (i.e. animated agents) in a retail application. The Home Furnishings application, already assessed in

Chapter 4 was deemed suitable as an environment in which the animated agents could appear as conversational assistants. Participants' perceptions toward a variety of aspects of the agents were assessed in order to document findings which interface designers can refer to when considering the use of humanoid animated agents in interfaces. In accordance with the definitions presented in Chapter 3, the evaluation monitored in particular, users' perceptions of the usability of the agents, the effectiveness of the interaction, and also the efficiency of the interaction.

5.2 Agent Types

Male and female versions of humanoid animated agent types were evaluated in the role of an interactive conversational sales assistant in a retail environment. Included in the cast of animated humanoids were 2D and 3D heads and 2D and 3D fully embodied agents. As a control between the evaluation described in Chapter 4 and this evaluation, the disembodied voice agent was included. As it was also shown that this agent type was significantly more popular than other photo-realistic agent types, this research issue also transfers to this evaluation in terms of the abilities of a disembodied voice agent to be more popular than other visible animated agents. An illustration of the structure of this passive viewing evaluation is given in Figure 5.1.

Animated agents are popularly realised using both two and three dimensions, each of which can offer a certain degree of lifelikeness to the screen (Thomas & Johnston, 1981). It is thought that 2D animations can offer sufficiently lifelike qualities for an engaging interaction to occur and some applications are using such agents in their applications (André & Rist. 1999; Person, 2000). However it has also been suggested (Badler, 1999; Cassell, 2000, Thalmann, 2000) that 3D realisations can extend the lifelikeness of the agents and this three-dimensionality allows the agents to express more human-like verbal and non-verbal communicative behaviours, increasing the depth of the interaction. However, the conceptualisation of width, height and depth can be difficult on a flat (2D) screen (Ratner, 1998), and therefore the x, y and z axes are used to define width, height and depth (forward and backwards movement), respectively. Using this coordinate system it is possible to create angular visualisations of characters and objects, providing an illusion from varying viewpoints, in order to illustrate the depth of the 3D object. It is argued that 3D agent animations may raise user expectations above and

beyond the actual capabilities of the humanoid agent leading to disappointing and frustrating communication (Lanier, 1995; Shneiderman, 1999). To discover which of the agent types is preferred by users and why they prefer them, both 2D and 3D humanoid animated agents are included in the evaluation. By establishing which agents potential user groups prefer it will be possible for systems to develop in order to satisfy user expectations.

In addition to issues of dimensionality, the evaluation also addresses the issue of graphically realising the animated agents as heads or entire bodies. Evidence exists that human-human conversation heavily depends on non-verbal gesturing for fluid, natural conversation flow (Whittaker & Walker, 1991, McNeill, 1992) and such gesturing plays an essential part in turn taking and turn giving. There is growing theoretical (Thorisson, 1996; Lester & Stone, 1997) evidence that by transferring this non-verbal communicative behaviour to the screen in the form of fully embodied characters, interactions between animated agents and human users may become more comfortable, efficient and effective. To construct such interactions the humanoid animated agents must be graphically realised as fully embodied agents and it is essential that these animations are as natural as possible, for it has been stated within the humanoid agents research community that "virtual humans should be alive, not just movable, they should have nice, virtual bodies, they should walk and not slide" (Thalmann, 2000).

The array of 2D and 3D humanoid animated agents, is illustrated in Figure 5.2. In addition to these, a sixth agent type was included. Advances in computer graphics will make it possible for Internet applications to be described in three dimensions as opposed to the more traditional two-dimensional scenarios of today and similar to the games industry these 3D retail environments are suitable environments to be inhabited with 3D fully embodied agents. With this in mind the evaluation assesses participants' attitudes toward 3D animated agents appearing as assistants immersed in the 3D application environment (see Figure 5.7).

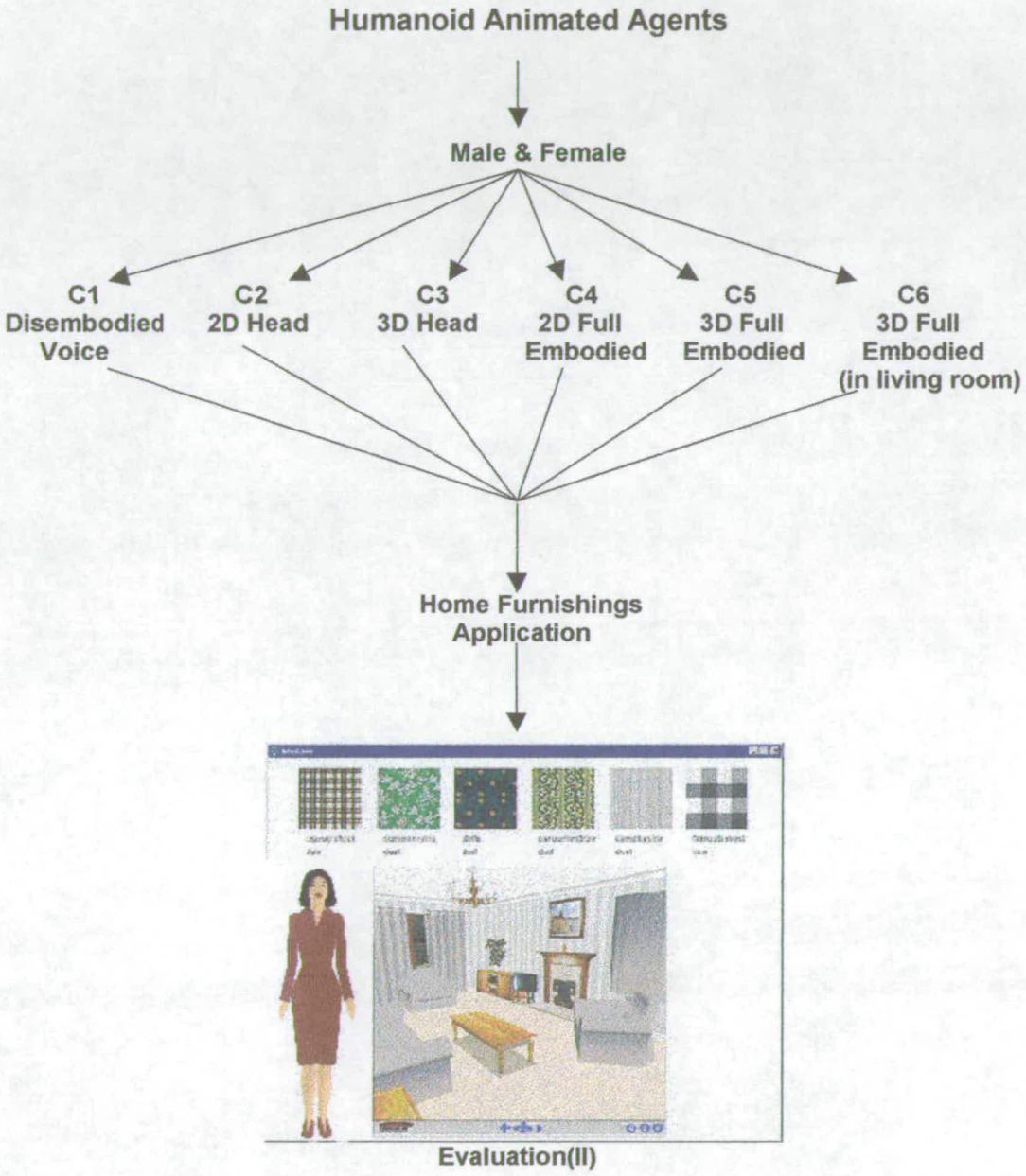


Figure 5.1 Illustration of Structure of Passive Viewing Evaluation II

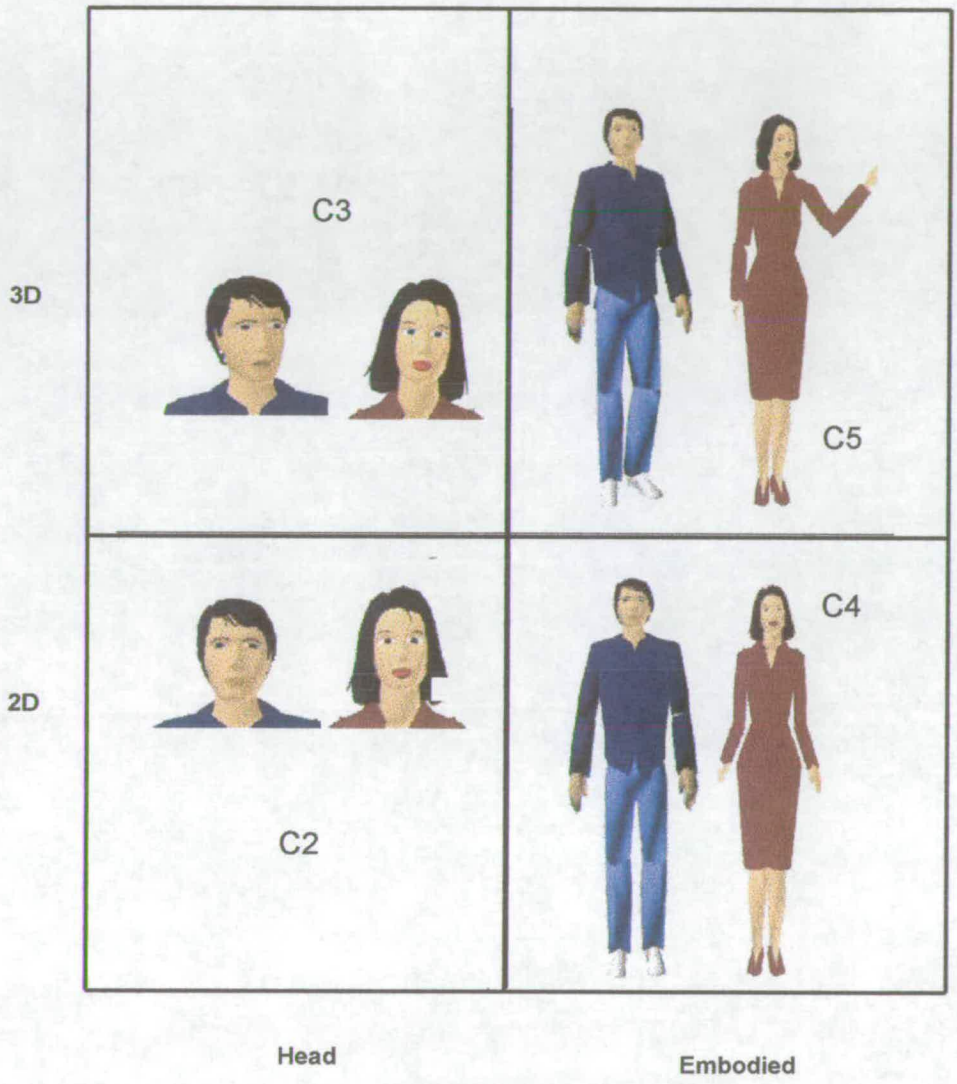


Figure 5.2 Array of Humanoid Animated Agents

5.3 Agent Implementation

The retail interface template that was designed and used for the evaluation of the humanoid photo-realistic agents (Chapter 4) was used again for the evaluation described in this chapter. The success of the passive viewing methodology for evaluating large numbers of agent types encouraged the re-use of the template for evaluating a large range of animated agents. The agent window was updated to cater for the addition of larger (embodied) agents and it increased in size from 4.5cm x 4.5cm to 4.5cm x 12cm. Using animation techniques a series of animated frames for each of the agents was created. Some of the agents were 2D realisations and these frames were sequenced using Macromedia Director. 3D Studio Max Character Studio was used to animate the 3D agent types that were included in the evaluation. Macromedia Director 5.6 was used again to combine the three components of the interface template: the agent window, the application window and the selection panel. A more in-depth discussion of the implementation for each of the agents is given next.

5.3.1 C1 – Disembodied voice

This agent was included as a control between the two casts of agents (photo-realistic and animated). Using an autocue to read scripts, one male and one female person played the role of the male and female assistants respectively. In the light of the negative attitude to the female voice of the photo-realistic agents in Chapter 4, different voices were used for the male and female agents in this evaluation. The audio file created, as with all previous agents, was sampled at 48kHz in stereo with a 16-bit resolution. Six similar but not identical agent dialogues (one per agent type) were created to avoid the monotony of repetitive dialogues that the participants reported feeling in the previous evaluation.

5.3.2 C2 – 2D talking head

Character profiles for both a male agent and female agent were selected (Figure 5.2). Adobe Photoshop 4.3, which can be described as a vector-based design and animation technology was used to create a series of different frames illustrating the movements of the face. Using Macromedia Director these individual frames were sequenced in order to produce an agent movie file in combination with the audio output file. These agents had

simplified lip-synchronisation corresponding to their verbal output. To create these visemes, five mouth movements were played in sequence to give the illusion of lip movement. The agents maintained gaze and looked straight ahead at all times during the conversation. Figure 5.3(i) uses the female version of the 2D talking head to illustrate individual frames showing facial movement. Figure 5.3(ii) again using the female version shows three mouth visemes.

Figure 5.3(i) 2D Facial Movements (Blinking, Smiling and Eyebrow Raising)

Figure 5.3(ii) 2D Mouth Visemes (Silence, O, EE)

5.3.3 C3 – 3D talking head

Using the same male and female character profiles that were created for the C2 agent, 3D Studio Max was used to model and animate the 3D agent. This software package has a rich set of operations and motion libraries for animating characters and using these operations it was possible to create a sequence, which depicted the agent in its 3D state (Figure 5.4). These 3D versions had the ability to nod at appropriate times during the conversation, as well as having the capability to display the head movements and mouth visemes of the corresponding 3D heads. They were also capable of displaying their three-dimensionality by turning their heads to give the user the impression they were looking at changes that were being made in the interface. Using the male agent as an example these features of the 3D head are illustrated in Figure 5.4(i) and Figure 5.4(ii).

3D Studio Max also contains a series of rendering operations to achieve high quality frames and animation clips in a variety of output formats and it was also possible to time the animations to the content of the audio files. The final format was played in Apple QuickTime and this agent component was combined with the Apple QuickTime movie of the main application window changes and the selection panel, using Macromedia Director to create a projector file.

Figure 5.4(i) 3D Head Turning

Figure 5.4(ii) 3D Head Nodding

5.3.4 C4 – 2D embodied

The male and female characters profiles that were chosen to create C2, were extended to have full bodies. As with the 2D head, this agent type was created using Adobe Photoshop 4.3 animation software and a series of different frames illustrating the movements of the face and the body were created. Using Macromedia Director these individual frames were sequenced in order to produce an agent movie file in combination with the audio output file recorded when creating C1, the disembodied voices. In addition to all of the facial features of C2 (mouth visemes, smiling, blinking) this agent type maintained gaze straight ahead and had deictic pointing gestures. A simple movement of the agent's arm could direct the user's attention to the selection panel at the

top of the GUI or to changes being made in the interface. This is illustrated using the female version of this agent type in Figure 5.5.

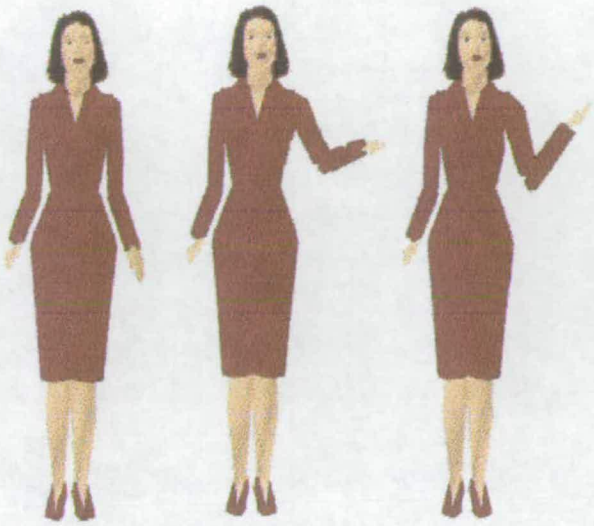


Figure 5.5 Series of Frames Illustrating 2D Female Embodied Movements

5.3.5 C5 – 3D embodied

This agent type can be described as a 3D fully embodied version of the 3D talking heads (C3). As with the 3D heads male and female versions were created using 3D Studio Max, which successfully modelled and animated the 3D characters. As with the 3D heads, this agent type could also nod. Gesturing also added to the realism of these agent and they could also turn around to give the impression that they were looking at changes that were being made in the interface. Figure 5.6 depicts the male 3D embodied agent displaying its capability to turn around. While speaking the agents displayed spontaneous gesturing. More in-depth discussion about gesturing is included later in this section. In combination with the audio file, the correct gestures could be displayed during the interaction. As with C3, the final format for this agent was Apple QuickTime which allowed the component to be combined easily with the other components of the retail interface template to create a single projector file.

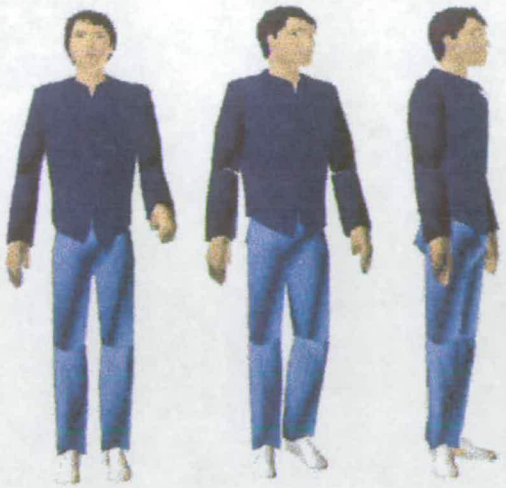


Figure 5.6 Series of Frames Illustrating 3D Male Turning Ability

5.3.6 C6 – 3D fully embodied in 3D environment

The 3D fully embodied animated agents (C5) actually appeared inside the 3D application environment as opposed to appearing at the left of the main application window. The male and female versions were identical to C5 and this agent type differed only in the position in the interface (Figure 5.7). To create the illusion that the agents appeared in the virtual space, 3D Studio Max was used to give the animation frames a background texture. This background was an image of the living space. The animation frames were created using 3D Studio Max, combined with the pre-recorded agent output file and finally exported into an Apple QuickTime movie file.

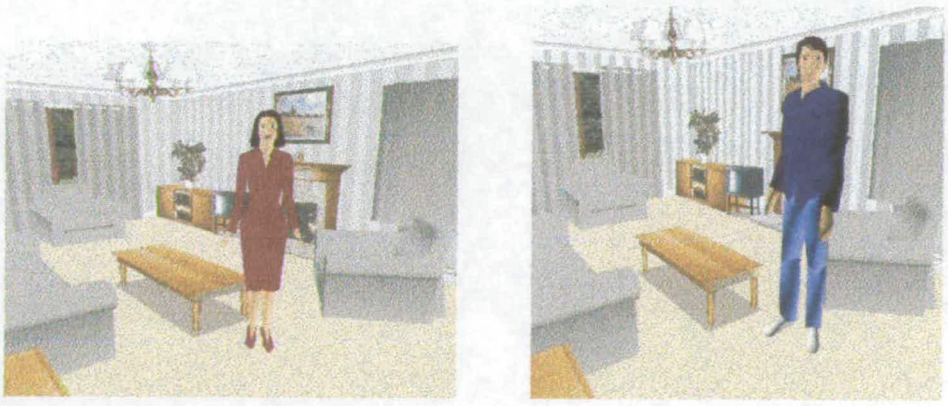


Figure 5.7 3D Embodied Agents in the 3D Room (C6)

5.3.7 Agents Verbal and Non-verbal Behaviour

All of the agents that appeared on the screen (C2-C6) were provided with facial expressions based on four categories defined by Cassell (2000):

- *Planning*: at the beginning of a turn speakers look at listeners more when they are sure about what they are going to say. They then look away to prevent information overload.

- *Comment*: head nods are used to emphasise linguistic items, eyebrow raising to synchronise with pitch accents.

- *Control*: during a conversation eye movement is used to regulate conversation; looking at the listener to request a response, or looking away to suppress the listener's response.

- *Feedback*: speakers look toward the listener when asking questions, listeners then establish gaze and/or nod.

The agents in this evaluation were given head nods to add emphasis although this was only evident in the 3D head and 3D embodied agents. Eyebrow raising was also included at appropriate pitch accents. For all agents, when asking questions, the agent raised the voice slightly, stressed the main word of the sentence and raised its eyebrows. During an affirmation the agent nodded (only 3D agents), raised its eyebrows and blinked at the end of the sentence. Further non-verbal facial feedback displays were included for all the agents. This ensured that the agents looked toward the customer during pauses in the conversation, when the customer was asking questions and when the agent was at the end of a turn. The 2D heads and 2D embodied agents maintained this gaze with the customer throughout the dialogue. However the 3D heads and 3D embodied agents were more mobile, and when they were speaking they had the ability to turn and look at the changes being made in the 3D room, but always returned to look at the customer at the end of an utterance.

Gesturing was introduced in a controlled manner for the 2D and 3D embodied agents. According to Cassell (2000) there are three categories of gesturing:

- *Emblematic:* culturally specific conscious actions than can be effective with and without the accompaniment of speech (e.g. shaking hands)

- *Propositional*: conscious actions found commonly where the physical world in which the conversation takes place is also the topic of conversation (e.g. Put-That-There, (Bolt, 1985)).

- *Spontaneous*: unconscious, unplanned, but co-verbal gestures (e.g. iconic, metaphoric, deictic or pointing).

The first two types of gestures are consciously included throughout conversation, but it is unconscious spontaneous gestures that constitute the vast majority of gestures and "are the gestural vehicles for our communicative intent with other humans, and potentially with our computer partners as well" (Cassell, 2000). In particular the nature of the evaluation dialogue supported the inclusion of deictic and beat gesturing. The 2D embodied agents only had deictic gesturing and pointed to the selection area at the top of the interface, thereby directing the user's gaze to that point. The functionality of deictic gesturing in animated agents has been assessed and results show that it is effective in directing and maintaining users' attention (Van Mulken, André & Müller, 1999). The three-dimensionality of C5 and C6 afforded the use of deictic and also beat gesturing, which was introduced to complement the agent's speech output. Beat gestures are pragmatic functions that can be described as "small baton-like movements that do not change in form the content of the accompanying speech", but rather they accompany the speech to sustain the conversational flow. It is important to note that the non-verbal behaviour for male and female agents of the same type was designed to be the same.

5.4 Experiment Predictions

1. It is predicted here that participants will respond positively to humanoid animated agents based on the findings presented in Chapter 4, which showed participants enjoy conversational interfaces, coupled with evidence that animated agents can enhance certain interfaces. However questions about the inclusion of humanoid animated agents in retail interfaces may be raised.

2. Attitude to the voices of the agents may differ between agent types, despite the fact that the same male voice was used for all male agents and the same female voice was used for all female agents. This can be stated since it was shown in Chapter 4 that the physical realisation of the agent can impact on the perception of the quality of the agent's voice and if the attitude to the particular agent is poor this can effect the perception of the agent's voice.

3. It was predicted that within agent types the male and female agents would be rated similarly as they were designed to have the same verbal and non-verbal behaviour, the only difference was their gender realisation. But it was predicted that attitudes might differ between the agent types.

5.5 Experiment Design

This experiment design was similar although not identical to the evaluation described in Chapter 4. After completion of the first passive viewing evaluation some improvements and alterations were made:

1. Only one application, the Home Furnishings application, was selected for use in the evaluation, as no outstanding application dependency issues arose during the evaluation of the humanoid photo-realistic agents.

2. The agent-customer dialogue in the first evaluation was identical for each agent and participants felt this was extremely monotonous. As a result six similar although not identical scripts were created and assigned to each of the six agent types in this evaluation. A sample dialogue can be examined in Table 4.1. The differences in the dialogues were variations in the fabric selections made by the customer.

3. Due to the negative reaction to the voices, especially the female voice, in the first passive viewing experiment (Chapter 4), new male and female voices were selected to record the output prompts for the male and female agents in this second passive viewing experiment.

4. Additional usability attributes were included in the questionnaires to enquire about attitudes to agents' facial movements and gestures. Extra usability attribute statements addressed agent politeness.

5. No focus groups followed the completion of the evaluation as it was concluded that no substantial additional information was gathered in the focus groups from the first evaluation.

To test the hypotheses a participant sample (N=36) was balanced for age and gender. Table 5.1 details the figures for each participant category. As stated in Chapter 3, cognitive walkthrough evaluation was completed for the experiment design. From this it was felt that no constraints should be placed over the participant sample as regards computing experience because the participants were only asked to listen and observe and were not asked to use the mouse or keyboard.

	Home Furnishings Evaluation		Total
Participant	**Male**	**Female**	
Age 18-35	6	5	11
Age 36-49	7	6	13
Age 50+	6	6	12
	19	17	36

Table 5.1 Analysis of Participants by Gender and Age Group

Participants first read a brief explanation of the purpose of the experiment (see Appendix 2.1), after which they were primed verbally to ensure they understood that they would be "eavesdropping" on a conversation between a customer (who was not visible) and an agent (who was sometimes visible: C2-C6). The customer was represented by a female disembodied voice for all interactions.

Each of the participants viewed 2-minute videos and these were presented in randomised order on a Pentium II PC. The videos showed the dialogue between the customer and an agent and the changes that were made to the interface corresponding to the dialogue. After each participant had witnessed a customer-agent interaction they were asked to complete a Likert questionnaire designed to retrieve quantitative data about their attitudes to the agent. Two content areas were chosen in the design of the questionnaire blueprint: agent embodiment and agent dimensionality. The manifestations were voice, personality and appearance. Not all dimensions of interest were relevant to all the

agents, therefore three questionnaires were used (see Appendix 2.2). These questionnaires correspond to the selected questionnaire statements listed in Table 5.2, where it is possible to also examine which usability attributes are relevant for particular agent types. The main categories of distinction are attitude to agents' voices, agents' competence, friendliness and politeness, agents' appearance, agents' facial movements and agents' gesturing.

Questionnaire Statements	C1	C2	C3	C4	C5	C6
1. I liked the assistant's voice.	*	*	*	*	*	*
2. The assistant's voice was annoying.	*	*	*	*	*	*
3. The assistant's voice was natural.	*	*	*	*	*	*
4. I felt the conversation was unnatural.	*	*	*	*	*	*
5. The assistant was competent.	*	*	*	*	*	*
6. The assistant was unfriendly.	*	*	*	*	*	*
7. The assistant was polite.	*	*	*	*	*	*
8. Being able to see the assistant was helpful.		*	*	*	*	*
9. The appearance of the assistant was unsuitable for the Home Furnishings application.		*	*	*	*	*
10. I liked the appearance of the assistant.		*	*	*	*	*
11. The lip movement was distracting.		*	*	*	*	*
12. The facial expressions made the assistant appear lifelike.		*	*	*	*	*
13. The facial expressions made the assistant appear unhelpful.		*	*	*	*	*
14. The facial expressions made the assistant appear unfriendly.		*	*	*	*	*
15. I liked the gestures the assistant made.				*	*	*
16. The gestures made the assistant appear lifelike.				*	*	*
17. The gestures made the assistant appear unhelpful.				*	*	*
18. The gestures made the assistant appear unfriendly.				*	*	*

Table 5.2 Questionnaire Statements

The dependent variables in the evaluation were the responses to the individual usability attributes on the questionnaire and the responses given during a closing interview, including an overall rating of each agent. The interview was designed to initiate discussion in order to retrieve information on the following topics:

1. Participant's view of the use of animated agents in an e-retail application.
2. The effective deployment of animated agents in the interface.
3. The expected characteristics of such agents.
4. The conversational possibilities with such agents in future applications.

To summarise, this evaluation of six types of humanoid animated agents was undertaken in the context of a Home Furnishings application. Participants in the evaluation completed questionnaires relating to each individual agent after which they took part in a one-to-one closing interview.

Title	Passive Viewing Evaluation II: Humanoid Animated Agents	
Design		One Independent Sample
Predictions	5.1	The deployment of humanoid animated agents will be accepted.
	5.2	Attitudes to voices of the same gender may differ for agent type, based on the acceptance of the realisation of the agent type.
	5.3	Within agent types male and female agents will be rated similarly; between agent types differences may occur.
Dependent Variables		Attitude Questionnaire Responses (1-7 Likert scale) Agent Ratings (1-10 scale)
Other Data		Interview Answers
(Experiment) Independent Variables:	1	Agent Type (6 levels)
	2	Agent Gender (2 levels)
(Participant) Independent Variables	1	Gender (2 levels)
	2	Age Group (3 levels)
Extraneous Variables:	Presentation Order	Agent presentation order randomised.
Location		Edinburgh - CCIR Premises, Central Edinburgh
Cohort		N = 36 50% male, 50% female
Remuneration		£10
Duration:		50 minutes

Table 5.3 Summary Table of Passive Viewing Evaluation II

5.6 Results

For each usability attribute, and also for the agent ratings, a series of repeated measures ANOVAs, taking agent gender and agent type as the independent variables, were completed. Each ANOVA table presents the results for within and between subject effects, together with interactions and results of the participant between subject variables of age and gender. For each usability attribute an explanation and discussion of any significant and non-significant results is presented.

5.6.1 Agent Ratings

After participants had witnessed each of the twelve agents in this evaluation they were asked to rate each one of a scale of 1 to 10, where 1 was low and 10 was high.

	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	6.762	5	1.352	.658	.656
A(Type) * P(Age)	14.190	10	1.419	.691	.732
A(Type) * P(Gender)	4.410	5	.882	.429	.828
Error(A(Type))	308.125	150	2.054		
A(Gender)	10.391	1	10.391	5.068	**.032**
A(Gender) * P(Age)	.227	2	.113	.055	.946
A(Gender) * P(Gender)	19.169	1	19.169	9.349	**.005**
Error(A(Gender))	61.514	30	2.050		
A(Gender) * A(Type)	18.595	5	3.719	3.124	**.010**
Error(A(Gender) * A(Type))	178.569	150	1.190		
Between Subject Effects					
P(Gender)	28.521	1	28.521	1.831	.186
P(Age)	78.838	2	39.419	2.531	.096
Error	467.292	30	15.576		

Table 5.4 Ratings ANOVA

The results of the repeated measures ANOVA (Table 5.4) taking agent gender and agent type as the independent variables and the mean rating scores as the dependent variable showed significant results for agent gender. The female agents were marginally significantly preferred to male agents, (mean female score = 6.20, mean male score = 5.89). In addition a highly significant interaction between agent type and agent gender showed that agent types C1, C2, C3, C4 and C5 of both genders were rated similarly, however there was a highly significant difference between the female and male versions of C6 ($p < 0.01$). The female 3D embodied agent (C6) that appeared in the 3D environment was rated significantly higher than the male counterpart. The mean results are displayed in Table 5.5.

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	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

Table 5.5 Agent Ratings Mean Scores for Agent Type and Gender

A second interaction showed an effect for agent gender and participant gender. The table of means indicates that the female participants in the evaluation rated the female agents significantly higher than the male agents (Table 5.6). The male participants however rated both the male and the female agents equally.

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	6.66	5.94
Male Participants	5.73	5.84

Table 5.6 Agent Ratings Mean Scores for Participant Gender and Agent Gender

5.6.2 Attitude to Voices

To provide a general impression of the results for this section, the mean scores for the independent variables of agent type and agent gender are presented in Figures 5.8.

5.6.2.1 Usability Attribute – "Liked voice"

The ANOVA table for this usability attribute shows a significant effect for agent gender (Table 5.7). In this evaluation the female voice was preferred to the male voice (see Figure 5.8(ii)). Again there is strong evidence of a greater preference from female participants for the voices of the female agents (Table 5.8). The male participants however liked the male and female voices equally. There was no effect for agent type, and so it can be stated that the voices of the varying agent types were liked equally.

I liked the assistant's voice	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	5.880	5	1.176	1.406	.225
A(Type) * P(Age)	5.995	10	.600	.717	.708
A(Type) * P(Gender)	2.880	5	.576	.689	.633
Error(A(Type))	125.444	150	.836		
A(Gender)	21.333	1	21.333	7.37	**.008**
A(Gender) * P(Age)	4.514	2	2.257	8.654	.431
A(Gender) * P(Gender)	25.037	1	25.037	9.602	**.004**
Error(A(Gender))	78.222	30	2.607		
A(Gender) * A(Type)	9.083	5	1.817	2.193	.058
Error(A(Gender) * A(Type))	124.278	150	.829		
Between Subject Effects					
P(Gender)	13.370	1	13.370	1.823	.187
P(Age)	4.394	2	2.197	.299	.743
Error	220.056	30	7.335		

Table 5.7 ANOVA for Usability Attribute "Liked Voice"

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	5.70	4.77
Male Participants	4.87	4.90

Table 5.8 Usability Attribute "Liked Voice"
Mean Scores by Participant Gender and Agent Gender

5.6.2.2 Usability Attribute – "Voice annoying"

Based on the results of the participants' attitudes to voice in the first evaluation, more usability attributes were included here to probe further participants' attitudes to the agents' voices. This usability attribute addressed whether participants felt the agents' voices were annoying. Statistical results (Table 5.9) showed highly significant effects for agent gender, where participants felt the female voice was also less annoying than the male voice.

The assistants voice was annoying	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	14.111	5	2.822	2.418	**.038**
A(Type) * P(Age)	9.292	10	.929	.796	.633
A(Type) * P(Gender)	3.380	5	.676	.579	.716
Error(A(Type))	175.056	150	1.167		
A(Gender	26.009	1	26.009	10.961	**.003**
A(Gender) * P(Age)	11.116	2	5.558	2.323	.115
A(Gender) * P(Gender)	21.333	1	21.333	8.916	**.006**
Error(A(Gender))	71.778	30	2.393		
A(Gender) * A(Type)	6.519	5	1.304	1.666	.146
Error(A(Gender) * A(Type))	117.389	150	.783		
Between Subject Effects					
P(Gender)	23.148	1	23.148	2.639	.115
P(Age)	21.181	2	10.590	1.208	.313
Error	263.111	30	8.770		

Table 5.9 ANOVA for Usability Attribute "Voice Annoying"

There were also significant effects for agent type (Table 5.10). As the male and female voices were the same for all male and female agents respectively, differences between agent types with respect to voice suggest the occurrences of cross-modal effects between the agent's physical realisation and speech output. Specifically, t-tests confirmed that the voices of the 3D embodied agents (C5 and C6) were significantly less annoying that all the other embodied agents C2, C3, C4 (all $p < 0.05$).

Humanoid Animated Agent Type	Mean Score
C1 (Disembodied voice)	4.96
C2 (2D Head)	4.93
C3 (3D Head)	4.72
C4 (2D Embodied)	4.64
C5 (3D Embodied)	5.19
C6 (3D Embodied in room)	4.99

Table 5.10 Usability Attribute "Voice Annoying"
Mean Scores by Agent Type

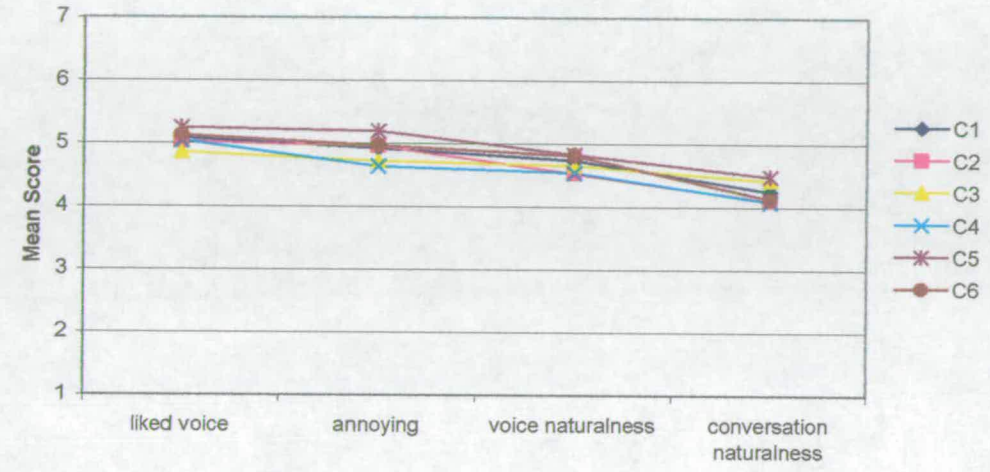


Figure 5.8(i) Usability Attributes for Agents' Voice by Agent Type

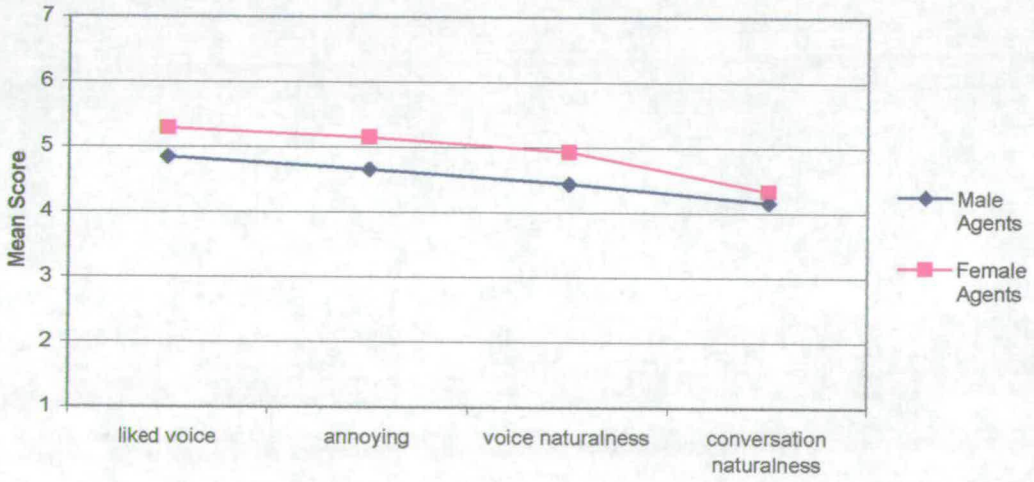


Figure 5.8(ii) Usability Attributes for Agents' Voice by Agent Gender

5.6.2.3 Usability Attribute – "Naturalness of Voice"

Participants were asked if the voice of the assistant was natural and the ANOVA table (Table 5.11) indicates highly significant differences for agent gender. There is a preference for the voices of the female agents and mean scores confirm that the female voice was perceived as being more natural (see Figure 5.8(ii)). No other significant results emerged.

The assistants voice was natural	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	6.407	5	1.281	1.303	.266
A(Type) * P(Age)	14.829	10	1.483	1.508	.142
A(Type) * P(Gender)	3.657	5	.731	.744	.592
Error(A(Type))	147.500	150	.983		
A(Gender)	26.009	1	26.009	13.537	**.001**
A(Gender) * P(Age)	1.699	2	.850	.439	.649
A(Gender) * P(Gender)	8.333	1	8.333	4.302	.057
Error(A(Gender))	58.111	30	1.937		
A(Gender) * A(Type)	3.935	5	.787	.997	.422
Error(A(Gender) * A(Type))	118.389	150	.789		
Between Subject Effects					
P(Gender)	17.120	1	17.120	1.556	.222
P(Age)	55.282	2	27.641	2.513	.098
Error	330.000	30	11.000		

Table 5.11 ANOVA for Usability Attribute "Naturalness of Voice"

5.6.2.4 Usability Attribute – "Naturalness of Conversation"

The results for this attribute show that the conversations with all the agents, regardless of their physical realisation or gender, were thought to be natural (grand mean = 4.24). So far the results show that the female voice was preferred; it was less annoying and more natural. This was the opposite finding to the evaluation reported in Chapter 4. Because of the poor perception of the female voice in the evaluation in Chapter 4, the female voice was changed to a more fluent conversational voice. Although the male voice was also changed, in comparison to the new female voice it was not thought to be as natural.

I felt the conversation was unnatural	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	10.799	5	2.160	1.532	.183
A(Type) * P(Age)	7.972	10	.797	.565	.840
A(Type) * P(Gender)	2.928	5	.586	.415	.838
Error(A(Type))	211.486	150	1.410		
A(Gender)	3.169	1	3.169	3.501	.071
A(Gender) * P(Age)	1.185	2	.593	.655	.527
A(Gender) * P(Gender)	.521	1	.521	.575	.454
Error(A(Gender))	27.153	30	.905		
A(Gender) * A(Type)	4.817	5	.963	1.152	.336
Error(A(Gender) * A(Type))	125.431	150	.836		
Between Subject Effects					
P(Gender)	14.447	1	14.447	.887	.354
P(Age)	24.000	2	12.000	.737	.487
Error	488.431	30	16.281		

Table 5.12 ANOVA for Usability Attribute "Naturalness of Conversation"

Finally, participants felt the conversations between the customer and the agents in the application were natural. This is an improvement on the previous experiment, where participants did not feel the conversations were natural because of their repetitive nature. The elimination of the repetitive dialogues is thought to have improved the results for this usability attribute.

5.6.3 Attitude to Competence, Friendliness and Politeness

In the evaluation of the photo-realistic agents (Chapter 4) qualitative findings indicated that politeness was an important trait for assistants, particularly in retail spaces. For this reason an additional usability attribute addressing agents' politeness was included in the attitude questionnaires. Figures 5.9(i) and 5.9(ii) present the mean scores for the independent variables of agent type and agent gender with respect to these three usability attributes.

5.6.3.1 Usability Attribute – "Competence"

The ANOVA showed that all twelve agents in the evaluation were thought to be equally competent with no effects evident for agent gender or for agent type.

The assistant was competent	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	9.259E-03	1	9.259E-03	.031	.862
A(Type) * P(Age)	7.407E-02	2	3.704E-02	.124	.884
A(Type) * P(Gender)	3.704E-02	1	3.704E-02	.124	.727
Error(A(Type))	8.972	30	.299		
A(Gender)	2.602	5	.520	1.082	.373
A(Gender) * P(Age)	1.898	10	.190	.395	.947
A(Gender) * P(Gender)	2.685	5	.537	1.117	.354
Error(A(Gender))	72.139	150	.481		
A(Gender) * A(Type)	3.491	5	.698	1.229	.298
Error(A(Gender) * A(Type))	85.194	150	.568		
Between Subject Effects					
P(Gender)	3.704	1	3.704	1.165	.289
P(Age)	7.907	2	3.954	1.244	.303
Error	95.361	30	3.179		

Table 5.13 ANOVA for Usability Attribute "Competence"

5.6.3.2 Usability Attribute – "Friendliness"

The assistant was friendly	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	15.102	5	3.020	3.15	**.006**
A(Type) * P(Age)	3.204	10	.320	.359	.962
A(Type) * P(Gender)	.611	5	.122	.137	.983
Error(A(Type))	133.750	150	.892		
A(Gender)	.454	1	.454	.583	.451
A(Gender) * P(Age)	.130	2	6.481E-02	.083	.920
A(Gender) * P(Gender)	1.333	1	1.333	1.712	.201
Error(A(Gender))	23.361	30	.779		
A(Gender) * A(Type)	13.491	5	2.698	2.657	**.025**
Error(A(Gender) * A(Type))	152.306	150	1.015		
Between Subject Effects					
P(Gender)	12.000	1	12.000	3.106	.088
P(Age)	6.685	2	3.343	.865	.431
Error	115.917	30	3.864		

Table 5.14 ANOVA for Usability Attribute "Friendliness"

Participants' attitudes to the friendliness of the agents significant effects for agent types were recorded (Table 5.14). The mean scores show that C5 and C6 were deemed to be

most friendly (Table 5.15). In fact, t-tests show that the 3D embodied agents C5 and C6 were significantly friendlier than the talking heads C2 and C3, all at $p < 0.01$.

Humanoid Animated Agent Type	Mean Rating Female Agents
C1 (Disembodied voice)	5.48
C2 (2D Head)	5.18
C3 (3D Head)	5.19
C4 (2D Embodied)	5.39
C5 (3D Embodied)	5.63
C6 (3D Embodied in room)	5.65

Table 5.15 Usability Attribute "Friendliness"
Mean Scores by Agent Type

A significant interaction between agent gender and agent type showed significant differences between the male and female agents in the C6 agent case ($p < 0.01$), where the male agent was significantly less friendly than the its female counterpart (Table 5.16). It has already been shown that this agent was not rated as highly as other agents and further in the chapter during a discussion of attitudes to gesturing an explanation is provided for this lower perception of this male agent type.

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	5.64	5.33
C2 (2D Head)	5.19	5.16
C3 (3D Head)	4.89	5.30
C4 (2D Embodied)	5.53	5.25
C5 (3D Embodied)	5.58	5.67
C6 (3D Embodied in room)	5.90	5.42

Table 5.16 Usability Attribute "Friendliness"
Mean Scores by Agent Type and Agent Gender

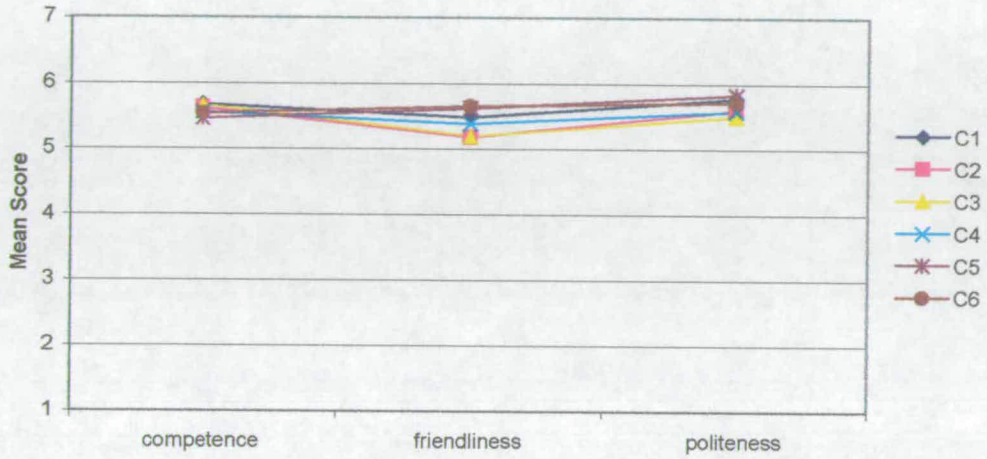


Figure 5.9(i) Usability Attributes for Agents' Personality by Agent Type

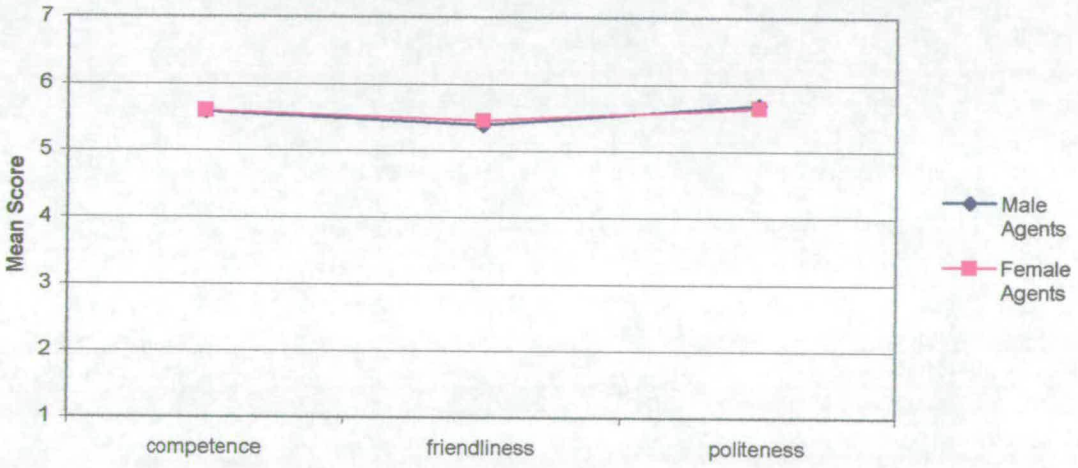


Figure 5.9(ii) Usability Attributes for Agents' Personality by Agent Gender

5.6.3.3 Usability Attribute – "Politeness"

The assistant was polite	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	5.956	5	1.191	5.17	**.009**
A(Type) * P(Age)	6.162	10	.616	1.084	.378
A(Type) * P(Gender)	1.604	5	.321	.565	.727
Error(A(Type))	85.236	150	.568		
A(Gender)	.187	1	.187	.269	.608
A(Gender) * P(Age)	.681	2	.340	.488	.618
A(Gender) * P(Gender)	6.021	1	6.021	9.15	**.006**
Error(A(Gender))	20.903	30	.697		
A(Gender) * A(Type)	9.826	5	1.965	2.571	**.029**
Error(A(Gender) * A(Type))	114.681	150	.765		
Between Subject Effects					
P(Gender)	7.521	1	7.521	2.994	.094
P(Age)	15.005	2	7.502	2.987	.066
Error	75.347	30	2.512		

5.17 ANOVA for Usability Attribute "Politeness"

Participants felt that all agents were polite but the ANOVA table shows significant differences for agent type (Table 5.17). C5 was significantly more polite that C2, C3 and C4, all at $p < 0.05$. Overall, the disembodied voice agents (C1) and the 3D fully embodied agents (C5, C6) were thought to be more polite than the 2D or 3D heads (C2, C3) suggesting that, as with friendliness, the 3D embodied agents could play an important role in participants' perceptions of politeness (Table 5.18).

Humanoid Animated Agent Type	Mean Rating
C1 (Disembodied voice)	5.78
C2 (2D Head)	5.59
C3 (3D Head)	5.50
C4 (2D Embodied)	5.58
C5 (3D Embodied)	5.83
C6 (3D Embodied in room)	5.72

Table 5.18 Usability Attribute "Politeness"
Mean Scores by Agent Type

An interaction was evident between the two main repeated measures variables of agent type and agent gender. The mean results (Table 5.19) and pair-wise comparisons show

that this effect was caused by a greater difference in participants' perceptions of the politeness of the 3D embodied agent type who appeared in the 3D environment (C6).

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	5.69	5.86
C2 (2D Head)	5.53	5.67
C3 (3D Head)	5.22	5.47
C4 (2D Embodied)	5.58	5.58
C5 (3D Embodied)	5.94	5.72
C6 (3D Embodied in room)	5.92	5.53

Table 5.19 Usability Attribute "Politeness"
Mean Scores by Agent Type and Agent Gender

An interaction between participant gender and agent gender 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 5.20).

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	5.71	5.59
Male Participants	5.39	5.68

Table 5.20 Usability Attribute "Politeness"
Mean Scores by Participant Gender and Agent Gender

The results from this section show that all the agents were thought to be equally competent, however the 3D embodied agents and the disembodied voices were thought to be more polite and also friendlier.

5.6.4 Attitude to Appearance

The mean scores for usability attributes relating to participants' perceptions of the appearance of the agents are graphically presented in Figure 5.10.

5.6.4.1 Usability Attribute - "Helpfulness"

Participants were asked if they thought 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. No significant effect for agent type emerged, and results indicate it was helpful to see all the agent types (Table 5.21).

Being able to see the assistant was helpful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	7.928	4	1.982	1.294	.276
A(Type) * P(Age)	13.622	8	1.703	1.111	.360
A(Type) * P(Gender)	7.694	4	1.924	1.255	.291
Error(A(Type))	183.867	120	1.532		
A(Gender)	9.344	1	9.344	6.410	**.017**
A(Gender) * P(Age)	2.822	2	1.411	.968	.391
A(Gender) * P(Gender)	.544	1	.544	.373	.546
Error(A(Gender))	43.733	30	1.458		
A(Gender) * A(Type)	3.517	4	.879	1.119	.351
Error(A(Gender) * A(Type))	94.267	120	.786		
Between Subject Effects					
P(Gender)	1.111	1	1.111	.077	.784
P(Age)	94.489	2	47.244	3.265	.052
Error	434.133	30	14.471		

Table 5.21 ANOVA for Usability Attribute "Helpfulness"

5.6.4.2 Usability Attribute – "Appearance suitable"

There were significant effects for agent gender with respect to this usability attribute and the appearance of the female agents was 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 also emerged (Table 5.23). T-tests showed that the male 3D embodied agent (C6) was less suitable for the application that the female counterpart, ($p < 0.01$). The presentation of the qualitative data further in the chapter provides reasoning why participants significantly felt that the male 3D embodied agent (C6) was unsuitable for the application. It does not appear to be simply a direct effect of the agent gender, but conclusions drawn from comments steer toward reasoning which suggests the male 3D embodied agent in the 3D environment was too large, making it difficult for participants to see changes that were being made in the interface.

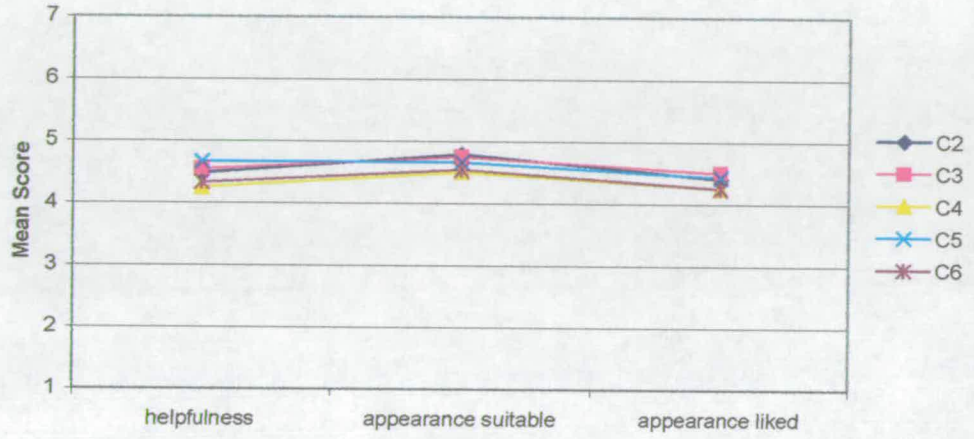


Figure 5.10(i) Usability Attributes for Agents' Appearance by Agent Type

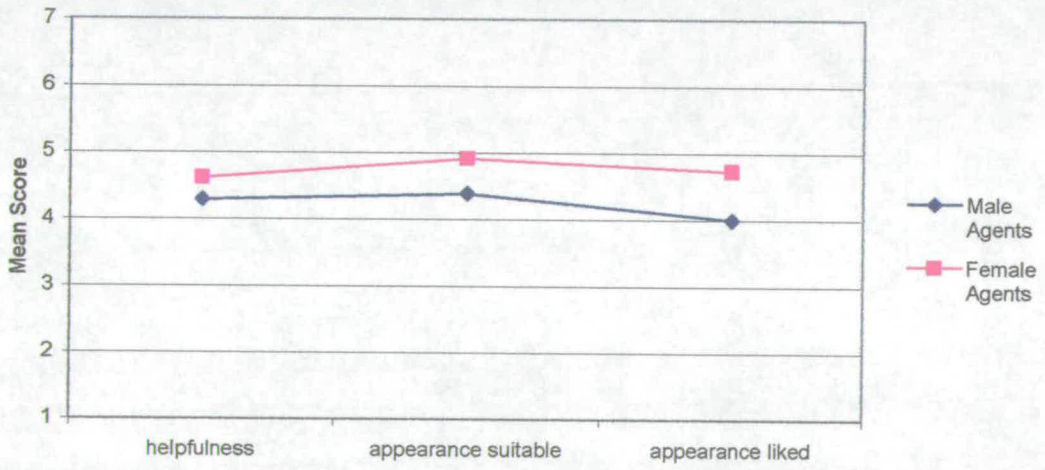


Figure 5.10(ii) Usability Attributes for Agents' Personality by Agent Gender

The appearance of the assistant was unsuitable for the application	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	4.650	4	1.162	.845	.499
A(Type) * P(Age)	18.183	8	2.273	1.653	.117
A(Type) * P(Gender)	1.350	4	.338	.245	.912
Error(A(Type))	165.000	120	1.375		
A(Gender)	24.544	1	24.544	16.522	**.000**
A(Gender) * P(Age)	8.672	2	4.336	3.233	.054
A(Gender) * P(Gender)	10.678	1	10.678	7.962	**.008**
Error(A(Gender))	40.233	30	1.341		
A(Gender) * A(Type)	8.872	4	2.218	3.04	**.045**
Error(A(Gender) * A(Type))	105.933	120	.883		
Between Subject Effects					
P(Gender)	1.344	1	1.344	.143	.708
P(Age)	63.150	2	31.575	3.365	.058
Error	281.500	30	9.383		

Table 5.22 ANOVA for Usability Attribute "Appearance suitable"

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	NA	NA
C2 (2D Head)	4.86	4.72
C3 (3D Head)	4.89	4.62
C4 (2D Embodied)	4.72	4.27
C5 (3D Embodied)	5.16	5.16
C6 (3D Embodied in room)	4.91	4.16

Table 5.23 Usability Attribute "Appearance suitable" by Agent Type and Agent Gender

A highly significant interaction between agent gender and participant gender follows earlier trends, where the female participants significantly preferred the female agents and no such gender distinctions occurred for the male participants (Table 5.24).

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	5.02	4.15
Male Participants	4.80	4.63

Table 5.24 Usability Attribute "Appearance suitable" by Participant Gender and Agent Gender

5.6.4.3 Usability Attribute – "Appearance liked"

I liked the appearance of the assistant	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	3.628	4	.907	.644	.632
A(Type) * P(Age)	17.906	8	2.238	1.589	.135
A(Type) * P(Gender)	11.483	4	2.871	2.038	.093
Error(A(Type))	169.033	120	1.409		
A(Gender)	48.400	1	48.400	22.021	**.000**
A(Gender) * P(Age)	17.817	2	8.908	5.525	.069
A(Gender) * P(Gender)	15.211	1	15.211	9.435	.064
Error(A(Gender))	48.367	30	1.612		
A(Gender) * A(Type)	29.294	4	7.324	8.619	**.000**
Error(A(Gender) * A(Type))	101.967	120	.850		
Between Subject Effects					
P(Gender)	3.211	1	3.211	.275	.604
P(Age)	25.317	2	12.658	1.083	.351
Error	350.633	30	11.68		

Table 5.25 ANOVA for Usability Attribute "Appearance liked"

The appearances of the female agents were significantly preferred to the male agents, and a significant interaction between agent type and agent gender showed that this was specifically the case for agent types C4, C5 and C6 (2D and 3D fully-embodied agents), where the female appearance was significantly preferred to the male (all $p < 0.01$), see Table 5.26.

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	NA	NA
C2 (2D Head)	4.36	4.45
C3 (3D Head)	4.58	4.42
C4 (2D Embodied)	4.86	3.64
C5 (3D Embodied)	5.03	3.83
C6 (3D Embodied in room)	4.83	3.67

Table 5.26 Usability Attribute "Appearance liked" by Agent Type and Agent Gender

These results showed a stronger attitude to the female agents and so not only was the female voice preferred, its appearance was also preferred in the retail application. However, attitudes to the male 3D embodied agent that appeared in the virtual world

(C6) were thought to be significantly poorer than its female counterpart. This poor score could also have impacted on the attitude to the appearance of male agents as a whole.

5.6.5 Attitude to Facial Movements

Based on feedback from the previous evaluation additional usability attributes relating to the agent's facial movements were included during this analysis of humanoid animated agents. The mean scores for agent type and agent gender are presented in Figure 5.11.

5.6.5.1 Usability Attribute – "Expressions distracting"

The lip movement was distracting	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	39.667	4	9.917	3.996	**.000**
A(Type) * P(Age)	19.883	8	2.485	2.012	.070
A(Type) * P(Gender)	5.822	4	1.456	1.179	.324
Error(A(Type))	148.200	120	1.235		
A(Gender)	.803	1	.803	.502	.484
A(Gender) * P(Age)	4.272	2	2.136	1.336	.278
A(Gender) * P(Gender)	1.225	1	1.225	.766	.388
Error(A(Gender))	47.983	30	1.599		
A(Gender) * A(Type)	10.322	4	2.581	3.184	.056
Error(A(Gender) * A(Type))	97.267	120	.811		
Between Subject Effects					
P(Gender)	3.025	1	3.025	.220	.643
P(Age)	49.950	2	24.975	1.814	.180
Error	413.050	30	13.768		

Table 5.27 ANOVA for Usability Attribute "Expressions distracting"

A significant effect for agent type emerged and a series of pair-wise comparisons showed that the lip movements of C2, C3 and C4 were more distracting than C5 and C6 (Table 5.28). In the interviews, participants said the lip movement of the 2D and 3D talking heads was more noticeable and looked artificial.

Figure 5.11(i) Usability Attributes for Agents' Facial Expressions by Agent Type

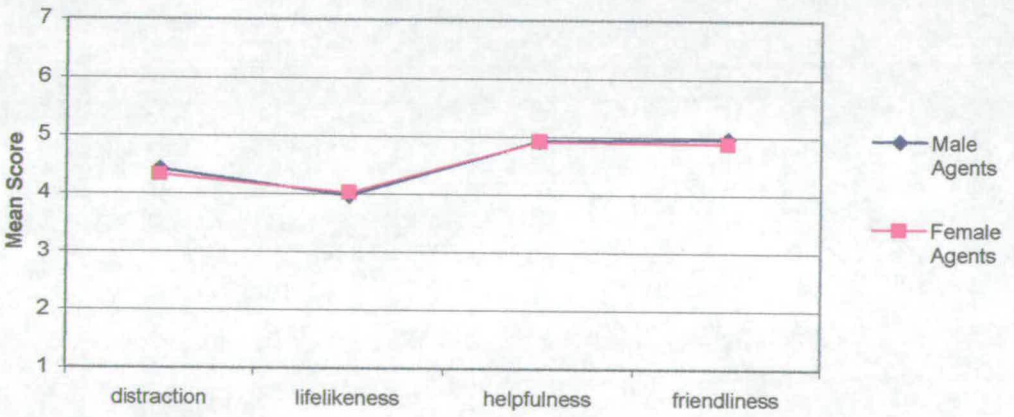


Figure 5.11(ii) Usability Attributes for Agents' Facial Expressions by Agent Gender

Humanoid Animated Agent Type	Mean Rating
C1 (Disembodied voice)	NA
C2 (2D Head)	3.89
C3 (3D Head)	4.14
C4 (2D Embodied)	4.39
C5 (3D Embodied)	4.67
C6 (3D Embodied in room)	4.79

Table 5.28 Usability Attribute "Expressions distracting" Mean Scores for Agent Type

5.6.5.2 Usability Attribute – "Expressions lifelike"

The facial expressions made the assistant appear lifelike	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	28.850	4	7.213	3.692	**.001**
A(Type) * P(Age)	6.617	8	.827	.568	.803
A(Type) * P(Gender)	2.572	4	.643	.441	.778
Error(A(Type))	174.833	120	1.457		
A(Gender)	.278	1	.278	.426	.519
A(Gender) * P(Age)	.839	2	.419	.643	.533
A(Gender) * P(Gender)	5.878	1	5.878	9.012	**.005**
Error(A(Gender))	19.567	30	.652		
A(Gender) * A(Type)	9.639	4	2.410	2.659	.056
Error(A(Gender) * A(Type))	108.767	120	.906		
Between Subject Effects					
P(Gender)	4.011	1	4.011	.322	.574
P(Age)	30.939	2	15.469	1.244	.303
Error	373.167	30	12.439		

Table 5.29 ANOVA for Usability Attribute "Expressions lifelike"

Significant results (Table 5.30) emerged for agent type with respect to the lifelikeness of the agents' facial expression. The facial expressions of C3 and C5 (the 3D head and 3D fully-embodied agent) appeared to be the most lifelike. The mean scores show that C3 and C5 had similar mean scores and t-tests showed that the facial movements of C3 were significantly more lifelike than C2, C4 and C6, ($p < 0.01$) indicating that the 3D agents were perceived as being more lifelike.

Humanoid Animated Agent Type	Mean Rating
C1 (Disembodied voice)	NA
C2 (2D Head)	3.89
C3 (3D Head)	4.48
C4 (2D Embodied)	3.68
C5 (3D Embodied)	4.12
C6 (3D Embodied in room)	3.80

Table 5.30 Usability Attribute "Expressions lifelike" by Agent Type

Again an effect for agent gender and participant gender emerged, with female participants having more positive attitudes to the female agents (Table 5.31)

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	4.25	3.94
Male Participants	3.78	3.98

Table 5.31 Usability Attribute "Expressions lifelike" by Participant Gender and Agent Gender

5.6.5.3 Usability Attribute – "Expressions helpful"

The facial expressions made the assistant appear unhelpful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	10.094	4	2.524	1.870	.120
A(Type) * P(Age)	8.539	8	1.067	.791	.611
A(Type) * P(Gender)	8.928	4	2.232	1.654	.165
Error(A(Type))	161.900	120	1.349		
A(Gender)	2.778E-03	1	2.778E-03	.003	.956
A(Gender) * P(Age)	1.356	2	.678	.765	.474
A(Gender) * P(Gender)	2.669	1	2.669	3.013	.093
Error(A(Gender))	26.583	30	.886		
A(Gender) * A(Type)	10.372	4	2.593	2.422	.052
Error(A(Gender) * A(Type))	128.500	120	1.071		
Between Subject Effects					
P(Gender)	9.669	1	9.669	1.708	.201
P(Age)	5.489	2	2.744	.485	.621
Error	169.850	30	5.662		

Table 5.32 ANOVA for Usability Attribute "Expressions helpful"

No significant interaction emerged for this usability attribute (Table 5.32), showing that the facial expressions for each of the visible agents (C2-C6) did not make the agents appear unhelpful (grand mean = 4.92).

5.6.5.4 Usability Attribute – "Expressions friendly"

The facial expressions made the assistant appear unfriendly	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	3.178	4	.794	.622	.647
A(Type) * P(Age)	6.922	8	.865	.678	.710
A(Type) * P(Gender)	2.456	4	.614	.481	.750
Error(A(Type))	153.167	120	1.276		
A(Gender)	.711	1	.711	.780	.384
A(Gender) * P(Age)	.106	2	5.278E-02	.058	.944
A(Gender) * P(Gender)	1.344	1	1.344	1.474	.234
Error(A(Gender))	27.367	30	.912		
A(Gender) * A(Type)	7.844	4	1.961	1.685	.158
Error(A(Gender) * A(Type))	139.633	120	1.164		
Between Subject Effects					
P(Gender)	5.378	1	5.378	.758	.391
P(Age)	1.272	2	.636	.090	.914
Error	212.833	30	7.094		

Table 5.33 ANOVA for Usability Attribute "Expressions friendly"

This usability attribute, which probed participants about their attitudes as to whether the facial expressions of the agents made them appear more friendly, showed no significant effects (Table 5.33), indicating the perceived friendliness of the facial expressions for each agent were similar (grand mean = 4.93).

5.6.6 Attitude to Gesturing

Usability attributes relating to the gesturing of the full-embodied agents (C4, C5 and C6) were also included. The mean scores are presented in Figure 5.12 and a discussion of the main findings follows.

5.6.6.1 Usability Attribute – "Gestures liked"

I liked the gestures the assistant made	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	9.231	2	4.616	2.431	.097
A(Type) * P(Age)	9.019	4	2.255	1.187	.326
A(Type) * P(Gender)	5.287	2	2.644	1.392	.256
Error(A(Type))	113.944	60	1.899		
A(Gender)	11.574	1	11.574	8.292	**.004**
A(Gender) * P(Age)	7.676	2	3.838	3.300	.051
A(Gender) * P(Gender)	6.000	1	6.000	5.159	**.030**
Error(A(Gender))	34.889	30	1.163		
A(Gender) * A(Type)	13.787	2	6.894	6.402	**.003**
Error(A(Gender) * A(Type))	64.611	60	1.077		
Between Subject Effects					
P(Gender)	2.241	1	2.241	.272	.606
P(Age)	50.454	2	25.227	3.057	.062
Error	247.556	30	8.252		

Table 5.34 ANOVA for Usability Attribute "Gesture liked"

Participants significantly preferred the gesturing of the female agents to the male gestures (mean female = 3.89, mean male = 3.69). The interaction between agent type and agent gender shows significantly that there exist differences between agents' gender for the 3D embodied agent in the room (C6). The mean results and pair-wise comparisons show the male 3D embodied agent (C6) in the 3D room had gesturing that did not appeal to the majority of the participants (Table 5.35).

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	NA	NA
C2 (2D Head)	NA	NA
C3 (3D Head)	NA	NA
C4 (2D Embodied)	4.00	3.84
C5 (3D Embodied)	3.94	4.19
C6 (3D Embodied in room)	4.14	3.36

Table 5.35 Usability Attribute "Gesture liked" by Agent Type and Agent Gender

Figure 5.12(i) Usability Attributes for Agents' Gesturing by Agent Type

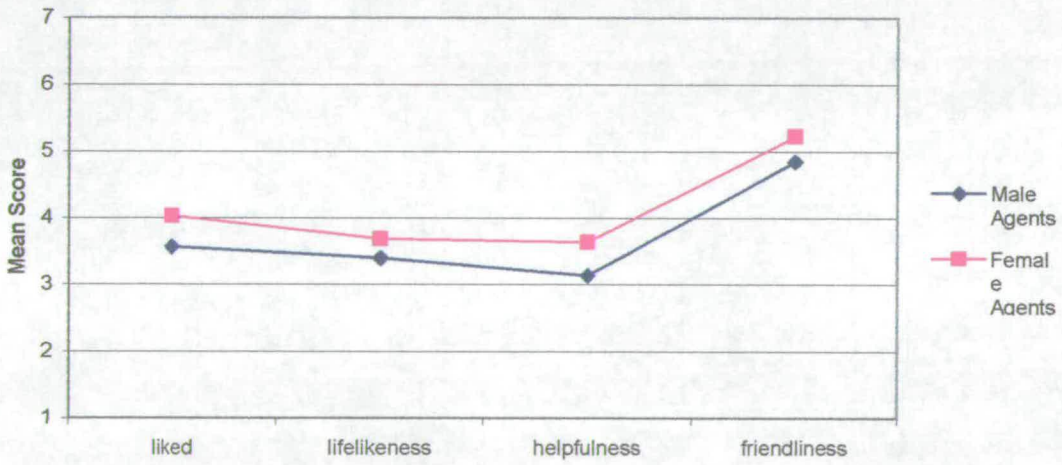


Figure 5.12(ii) Usability Attributes for Agents' Gesturing by Agent Gender

5.6.6.2 Usability Attribute – "Gestures lifelike"

The gestures made the assistant appear lifelike	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	11.444	2	5.722	3.78	**.017**
A(Type) * P(Age)	13.778	4	3.444	2.646	.052
A(Type) * P(Gender)	.778	2	.389	.299	.743
Error(A(Type))	78.111	60	1.302		
A(Gender)	4.449	1	4.449	3.180	.085
A(Gender) * P(Age)	7.676	2	3.838	2.743	.081
A(Gender) * P(Gender)	4.449	1	4.449	3.180	.085
Error(A(Gender))	41.972	30	1.399		
A(Gender) * A(Type)	4.037	2	2.019	2.665	.078
Error(A(Gender) * A(Type))	45.444	60	.757		
Between Subject Effects					
P(Gender)	5.042	1	5.042	.576	.454
P(Age)	34.361	2	17.181	1.962	.158
Error	262.639	30	8.755		

Table 5.36 ANOVA for Usability Attribute "Gestures lifelike"

Participants significantly felt that the gesturing of certain embodied agents was more lifelike than others (Table 5.37). Upon closer examination of the mean scores it is concluded that the 3D embodied agent outside the 3D space (C5) was perceived as being more lifelike than the 3D embodied agent who appeared inside the 3D space (C6). Pair-wise comparisons showed that C5 was significantly more lifelike than C4 and C6, both at $p < 0.01$.

Humanoid Animated Agent Type	Mean Rating
C1 (Disembodied voice)	NA
C2 (2D Head)	NA
C3 (3D Head)	NA
C4 (2D Embodied)	3.23
C5 (3D Embodied)	3.79
C6 (3D Embodied in room)	3.59

Table 5.37 Usability Attribute "Gestures lifelike" by Agent Type

5.6.6.3 Usability Attribute – "Gestures helpful"

The gestures made the assistant appear unhelpful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	10.037	2	5.019	15.78	**.003**
A(Type) * P(Age)	5.574	4	1.394	1.822	.136
A(Type) * P(Gender)	3.370	2	1.685	2.203	.119
Error(A(Type))	45.889	60	.765		
A(Gender)	7.782	1	7.782	5.75	**.025**
A(Gender) * P(Age)	5.574	4	1.394	1.822	.136
A(Gender) * P(Gender)	1.042	1	1.042	.747	.394
Error(A(Gender))	41.861	30	1.395		
A(Gender) * A(Type)	9.481	2	4.741	3.85	**.014**
Error(A(Gender) * A(Type))	62.222	60	1.037		
Between Subject Effects					
P(Gender)	7.782	1	7.782	2.087	.159
P(Age)	1.287	2	.644	.173	.842
Error	111.861	30	3.729		

Table 5.38 ANOVA for Usability Attribute "Gestures helpful"

Participants were asked if the agents' gesturing appeared unhelpful. The ANOVA shows significant results for agent gender, agent type and an interaction between agent type and agent gender (Table 5.38). The gesturing of the 3D embodied agents was perceived as being more helpful than the 2D fully embodied agent, and this was significantly the case for female agents who appeared outside the 3D room (C5), see Table 5.39.

Humanoid Animated Agent Type	Mean Rating Female Agents	Mean Rating Male Agents
C1 (Disembodied voice)	NA	NA
C2 (2D Head)	NA	NA
C3 (3D Head)	NA	NA
C4 (2D Embodied)	4.80	4.72
C5 (3D Embodied)	5.53	5.25
C6 (3D Embodied in room)	5.33	4.55

Table 5.39 Usability Attribute "Gestures helpful"
Mean Scores by Agent Type and Agent Gender

5.6.6.4 Usability Attribute – "Gestures friendly"

The gestures made the assistant appear unfriendly	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
A(Type)	5.731	2	2.866	3.592	**.043**
A(Type) * P(Age)	1.185	4	.296	.320	.864
A(Type) * P(Gender)	.565	2	.282	.305	.738
Error(A(Type))	55.611	60	.927		
A(Gender)	2.667	1	2.667	3.229	.082
A(Gender) * P(Age)	2.778	2	1.389	1.682	.203
A(Gender) * P(Gender)	8.963	1	8.963	10.852	**.003**
Error(A(Gender))	24.778	30	.826		
A(Gender) * A(Type)	3.528	2	1.764	2.128	.128
Error(A(Gender) * A(Type))	49.722	60	.829		
Between Subject Effects					
P(Gender)	4.741	1	4.741	1.210	.280
P(Age)	5.481	2	2.741	.699	.505
Error	117.556	30	3.919		

Table 5.40 ANOVA for Usability Attribute "Gestures friendly"

The final usability attribute reveals significant effects for agent type when asked if the gestures made the assistants appear unfriendly. In fact C5 and C6 were significantly friendlier than C4, $p < 0.01$.

Humanoid Animated Agent Type	Mean Rating
C1 (Disembodied voice)	NA
C2 (2D Head)	NA
C3 (3D Head)	NA
C4 (2D Embodied)	4.85
C5 (3D Embodied)	5.23
C6 (3D Embodied in room)	5.15

Table 5.41 Usability Attribute "Gestures friendly"
Mean Scores by Agent Type

An interaction between participant gender and agent gender indicated once again that the female participants had more positive attitudes toward the female agents (Table 5.42). A similar usability attribute ("The facial expressions made the assistant appear unfriendly") showed no effects for friendliness due to the facial movements of the agents. This

usability attribute ("The gestures made the assistant appear unfriendly") however does show that the friendliness of the agents may be promoted through gesturing.

	Mean Rating Female Agents	Mean Rating Male Agents
Female Participants	5.54	4.90
Male Participants	4.84	5.02

Table 5.42 Usability Attribute "Gestures friendly"
Mean Scores by Participant Gender and Agent Gender

The final section of usability attributes relating to the attitudes to the agent's facial expressions and gestures revealed a positive attitude toward the 3D embodied agents. Although there were poor attitudes recorded for the 3D male embodied agent who appeared in the virtual world because of its dimensions, overall the facial expressions of the 3D embodied agents were more lifelike, the gesturing made them appear more lifelike and friendly and finally they were also perceived as being more helpful than the other agents.

5.6.7 Agent Ranking

To investigate which of the agents the participants would like to interact with if they were to use the service themselves, they were shown short excerpts of each of the agents and then asked to place in order the top three assistants. The table of results is given in Table 5.43.

Rank	C1 (Dis-embodied voice)		C2 (2D Head)		C3 (3D Head)		C4 (2D Fully Embodied)		C5 (3D Embodied)		C6 (3D Embodied in room)	
	F	M	F	M	F	M	F	M	F	M	F	M
1	6	5	0	1	2	2	1	0	6	5	6	2
2	3	4	1	2	3	2	1	2	5	6	4	3
3	2	2	3	3	4	5	4	4	3	4	1	1

Table 5.43 Participants' Rankings for Humanoid Animated Agents

First preference votes went primarily to the disembodied voices agents (C1) and the 3D fully embodied agents, (C5 and C6). The majority of second preference votes also went to these three agent types again. As with the ranking results for the humanoid photo-realistic agents detailed in Chapter 4, the second preference trend from individual

participants was to vote for the opposite gender of the same agent type. In fact, for agent types, C1, C5 and C6, half or more of the participants who voted for a particular agent type and agent gender, allocated their second preference vote to the same agent type, but to the opposite agent gender. For example, of the six participants who voted for the female disembodied voice, three allocated their second preference votes to the male disembodied voice. This result accentuates the preferences for particular agent types, regardless of agent gender. These first and second preference correlations are explained further in Table 5.44. The figures in parentheses show the number of votes that came from participants who voted for the opposite gender of the same agent type for their first preference. The table illustrates this correlation for the highest scoring agent types.

Rank	C1 (Disembodied voice)		C5 (2D Head)		C6 (3D Head)	
	F	M	F	M	F	M
1	6	5	6	5	6	2
2	3(3)	4(3)	5(4)	6(4)	4(2)	3(3)

Table 5.44 First and Second Preferences Correlations

5.6.8 Interview Feedback

Overall participants considered that the conversational capabilities of the application actually enhanced the service and participants expressed a desire to use the service interactively. One participant voiced that although it may depend on the application, conversational agents could improve the shopping experience by adding social benefits. The overall response to the humanoid animated agents was good and 25 participants in the sample (N= 36) felt that being able to see the assistants improved the experience. This is an improvement to the results found in Chapter 4 where approximately half the participants in the sample felt that the visual presence of the agents enhanced the service. A selection of comments from the participants is presented in Table 5.45 below.

Home Furnishings
"Just hearing the assistant was dull"
"Seeing the assistant adds character, giving the feeling of interaction"
"It was possible to get conversational cues from seeing them [the assistants] making it more fun",
"Having the agents present on the screen makes it feel like you are actually shopping".

Table 5.45 Participants' Comments about the Agents

To probe participants about their opinions of the 3D fully embodied agents they were asked if they preferred the agent inside or outside the 3D environment. Twenty-two participants preferred to see the assistant outside the living room. Some of the comments offered explaining this choice were as follows: *"I couldn't see the living room", "the guy looked big inside the room", "I could concentrate more on the changes in the room when the assistant wasn't in the room"*. Comments from other participants explain the poor acceptance of the embodied agents in the 3D environment, in particular the male agent. Post evaluation analysis indicated the physical presence of the male agent in the environment was in fact larger than the female counterpart. Because of this the agent was perceived as being *"dominating"*. The larger physical realisation of this male agent in the environment was not intended, however the results nevertheless produced interesting information. Namely, if intending to place 3D embodied agents within 3D environments it should firstly be ensured that the dimensions of the agent are in proportion with the environment and secondly that the agent does not block the user's view of any changes that may be occurring in the environment. Encouragingly the participants who did accept the 3D embodied agent in the room they stated that *"it made the room easier to look at", "it seemed more complete and more natural to have the assistant in the room", "it's more like real life", and "it added to the realism"*.

In an attempt to investigate preferences for humanoid animated agents and humanoid photo-realistic agents, participants were asked if they had a choice which type of assistant they would rather interact with an animated assistant or an assistant that appeared as a video of a human. Eight said that it would depend on what they were buying and that it would probably be better to do more serious tasks with human-like assistants. Those in favour of photo-realistic video supported their opinion by offering the following comments: *"it may be more realistic", "perhaps it would be more believable"*. These comments echo the results presented in Chapter 4 in support of photo-realistic agents that behave in a believable, realistic manner. Those in favour of humanoid animated agents commented: *"animated worlds should have animated characters"*. Another participant commented: *"I would feel less pressured with animations, it's less serious"*. It can be said that from the results of this evaluation there is evidence to support the deployment of animated agents in retail interfaces.

5.7 Discussion

Prior to the evaluation three experiment predictions were made. Firstly it was claimed that the humanoid animated agents would be received positively in the Home Furnishings application (Prediction 5.1). This claim was supported and both the quantitative and qualitative analysis confirmed that participants did think that deploying animated agents as conversational assistants in Internet retail applications was a good idea. Greater support for the use of animated agents was evident from the participant sample, as two-thirds of the sample stated they preferred the presence of a visible agent was an actual enhancement. This is a marked improvement on the support demonstrated for the deployment of photo-realistic agents in the same Home Furnishings application, when only fifty percent of the sample thought the visual presence of an agent enhanced the interface and it was suggested that this photo-realistic agent must have sophisticated human-like verbal and non-verbal behaviour at all times during the interaction. The physical realisation of the agents as disembodied voices was still popular, confirming that speech interfaces would be usable, useful and comfortable mediums to complete a variety of retail tasks.

The second prediction (Prediction 5.2) was that attitude differences with respect to the agents' voices may occur, as user acceptance of the physical realisation of the agent can cause the occurrence of cross-modal effects. For instance, if the appearance of the agent does not appeal to the user this could impact on users' perception of the quality of the agent's speech output as was discovered in Chapter 4. So despite the fact that the same male voice was used for all male agents and the same female voice for all female agents, attitude differences to the various agent types of the same gender may occur. This claim was not supported and no differences emerged with respect to perceptions of the voices between agent types, however in the results for the attributes relating to participants' attitudes to the agents' voices, it was shown that the voices of the 3D embodied agents were thought to be less annoying than the other visible agents.

The results also showed the general preference for the female voice over the male voice. This was the opposite finding to that reported in Chapter 4. Despite the fact that new voices, with a more fluent and conversational style were used for the male and female agents' speech output, gender differences still emerged. Although both of the people that made the recordings had experience in the creation of speech output for a variety of

speech interfaces it was significantly felt by the participant sample that the male voice was monotonous. This result, combined with similar evidence in Chapter 4, strongly indicates that if and when selecting human voices for agents it is essential to pre-evaluate the voices on a sample of the potential user group.

The third experiment prediction (Prediction 5.3) stated that within agent types the male and female agents would be rated similarly and that attitudes may differ between the agent types. The first part of this prediction was not supported when it was discovered that there were significant attitude differences to the male and female agents of C6. The result showed that the male 3D fully embodied agent who appeared in the application environment was not as well received as well as the female counterpart. Participants felt that the 3D embodied agent in the 3D room was poorly accepted because the agent was perceived as being dominating and large and as a result participants found it difficult to concentrate on the changes that were being made in the interface. The poor perception of the male 3D embodied agent also offers a plausible explanation for the preference for the gesturing of the female embodied agents.

Between the agent types a number of differences emerged. The ratings, rankings and responses to the majority of usability attributes confirmed the popularity of the 3D embodied agents over all other visible agent types. In the interviews it was confirmed that the majority of participants preferred the 3D agents to the 2D agents and that the 2D and 3D fully embodied agents were preferred to 2D and 3D heads. In particular the 3D embodied agents were thought to be friendlier and more polite. Although the deictic pointing gestures of the 2D embodied agents were thought to be helpful in agreement with van Mulken et al.(1999), the integration of deictic and beat gesturing together with the other non-verbal behaviour in a 3D embodiment (i.e. nodding and turning appropriately) promoted the general acceptance of that agent type. In fact usability attributes relating to gesturing demonstrated that the gesturing of 3D embodied agents can promote the perceived lifelikeness, friendliness and helpfulness of the agents as assistants.

Other results listed earlier in the chapter show a stronger preference from the female participants toward the female agents and that no such gender differences occurred for the male participants in the sample. It is difficult to assess why this gender division occurred in the participant sample and it is even more difficult to offer possible explanations given the results favouring the female agents firstly because of the general

preference for the female voice and also because of the significant preference for the female 3D embodied agent in the environment (C6) due to the fact the male 3D agent was perceived as being dominating. As the Home Furnishings application is not used in the evaluations that follow it is not possible to definitely state that these gender differences were application dependent. However, based on the findings reported in Chapter 4, where no application dependency issues arose between the Home Furnishings application and a CD Service, and also that no participant gender differences emerged such as those reported in this chapter, it is concluded here that the findings in Chapter 4 support the possibility that the gender difference was not based on the subject matter of the application.

Only further experiments can clearly explain why this gender difference occurred. Detailed in the chapters that follow are evaluations where participants had the opportunity to interactively communicate with 3D embodied conversational agents that were immersed in a variety of 3D retail applications. The negative attitudes to voices were eliminated through pre-evaluation testing and it was also ensured that the dimensions of the 3D male agent in the 3D environments were in proportion. As the participant samples were again balanced for age and gender it was possible to analyse whether such gender differences from the female participants would re-occur.

5.8 Summary

In summary this chapter has presented an empirical evaluation that aimed to assess the effectiveness of a range of humanoid animated agents in a retail application. An interface template was used in order to employ evaluation by observation methods. A detailed technical description of this graphical user interface template was provided in Chapter 4 and in this chapter a technical description of the technology used to create six agent types was provided. A series of images presented throughout the text assists in the description of the functionality and capability of the animated agents.

The experiment predictions and procedure were detailed, including an explanation of necessary improvements that had to be made to the experiment design. Both quantitative and qualitative findings were then documented. The results of this chapter show significant support for the inclusion of humanoid animated agents in retail applications.

In particular the presence of a fully embodied agent is preferred to the realisation of the agent as a head. Extending this, users seem more satisfied when the agent is described in three dimensions.

In the following chapter a more in-depth analysis of 3D embodied agents is detailed. In order to complete a comprehensive evaluation of this agent type an interactive experiment platform was designed and constructed. The interactive system allowed embodied agents to appear as conversational assistants in a series of retail applications. By inviting participants to *take part* in controlled interactive experiments rather than simple observe, further empirical evidence is gathered about the effectiveness of this agent type as an ECA in retail applications.

Chapter 6

Constructing Contrasting Interactive Retail Applications Inhabited by 3D VRML Embodied Conversational Agents

6.1 Introduction

This work positions itself within the socially intelligent agents (SIA) design community, where research is aiming to enhance the way in which people communicate with computers. Developing agents (software or robotic) that behave socially in an interaction has become the principal goal for many interdisciplinary researchers involved with the development of intelligent communicative systems. As presented in Chapter 2 where an in-depth overview was given, a variety of interactive lifelike characters are being successfully used within SIA research to address many challenging issues. These research issues range from the exploring the affordances of embodiment, to discovering the possibilities of building and developing a social interaction between computers and users through an agent. The construction of 'social relationships' between agents and users is also at present an insightful domain within this research community, as is the search to understand the cognitive and emotional needs of human users in order to create more effective and affective communicative systems.

Developments in speech systems engineering and computer graphics are making it possible for ECA to appear in interactive applications. However for interactions in these applications to be effective the user must be able to complete tasks with the agent without difficulty and without reduction in expectations toward the agent. In addition, for the system to be affective, the agent must appropriately satisfy and respond to users' desires and intentions with respect to a particular application and adapt to the users' changing cognitive state over time. To begin to construct effective and affective socially intelligent agents, empirical evidence needs to be documented in order to discover what are users' expectations, desires and intentions toward the agents.

The research reported in this chapter forms a natural progression from the findings detailed in Chapter 5, which demonstrated a largely positive attitude to the conversational capabilities of embodied animated agents, but showed that there are

numerous unanswered questions with respect to the use of 3D humanoid animated agents inhabiting 3D retail environments. In an attempt to address these unanswered questions an interactive system was created where participants could actually converse with the agents. This scenario will allow more in-depth research to be completed than the previously discussed passive viewing evaluations and importantly it will allow empirical evidence to be documented regarding users' expectations, desires and intentions toward agents' behaviour, appearance and personality which will form the basis for creating firstly effective and possibly affective interactions with ECA in retail applications.

Churchill et al. (2000) stated that to create rich interactions with ECA the agents should be designed with "conscious consideration of social rules of engagement". To do this Churchill charted out the most important issues to consider when designing socially responsive interactive embodied agents. These issues, illustrated in Figure 6.1, include the agent's role, and its domain and interactional competencies; the agent's self-presentation in terms of appearance and motion; and the agent's personality.

Regardless of the role the ECA will play it must display domain competence for the interaction with the user to be credible. It is not only through verbal behaviour that the agent can display competence, but as Churchill et al. (2000) and Cassell (2000) explain non-verbal behaviour plays an essential role in conveying competence and so consideration must be given to the agent's direction of gaze, facial movement and gesturing as demonstrated in Chapter 5. Together with verbal displays of competence through intonational cues for example, using these multiple modes of communication the result may be a better social interaction and the user may then perceive the agent to be more comfortable in its role. In this evaluation the agents were conversational assistants in contrasting retail applications and their verbal and non-verbal behaviour, as described later in the chapter, reflected Churchill's guide in order to create credible agents. The applications in which the agents were assessed were a virtual cinema box-office, a virtual travel agency and a virtual bank. They were chosen to represent popular telephone activities for which the expansion of Internet technologies, the likelihood of greater bandwidth and the growing evidence that embodiment can enhance speech interfaces, it is probable that in the future, tasks in electronic retail applications may successfully be completed in collaboration with embodied agents. The application interfaces were realised in three dimensions using VRML and the 3D ECA appearing in these interfaces were immersed in the 3D environment.

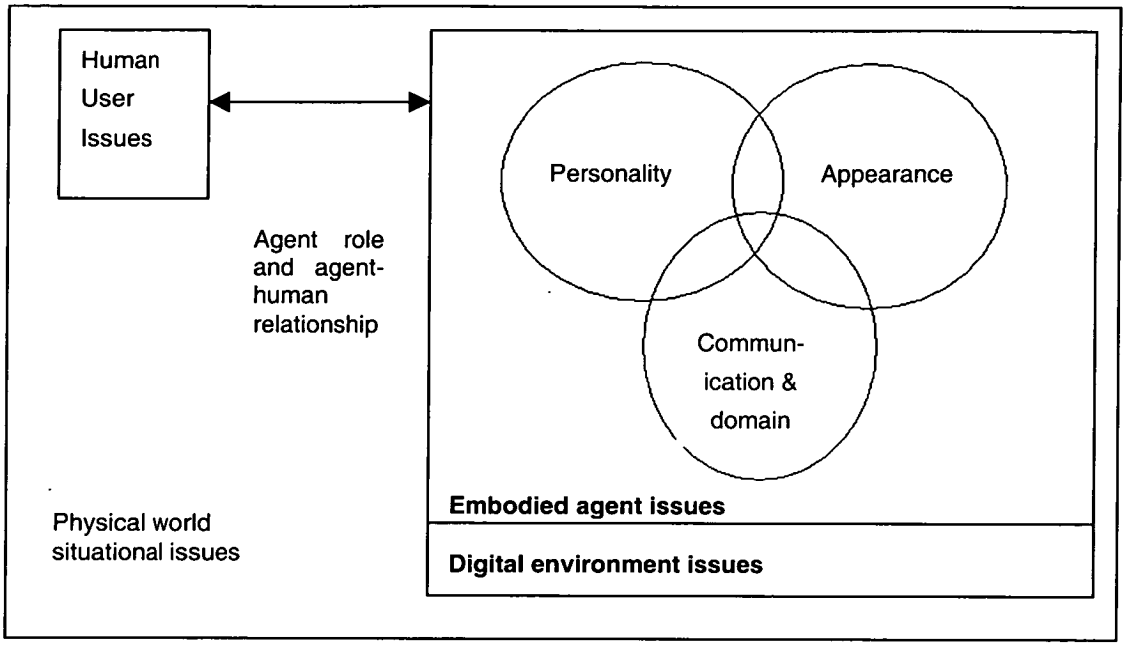


Figure 6.1 Issues in the Design of an Embodied Agent (Churchill et al, 2000)

To create a rich social interaction, according to Churchill the second important issue in the design of embodied agents is the agent's appearance. Research has confirmed that agents should be designed to engage the user, however Churchill describes this design feature as an unspecific goal and states that further attention should be given to the surface appearance of the agents. Beliefs about others are commonly based on the "stereotypical interpretation of surface appearance" and it is these beliefs that can shape the user's expectations of the agent and intentions and actions toward the agent. The evaluation in Chapter 5 showed that 3D humanoid animated embodied agents were preferred to 2D humanoid heads and 2D embodied agents, and in an attempt to extend the research into the perception of the appearance of the agents, stereotypical assistant appearances were created. The agents (male and female) appeared either formally dressed or casually dressed in all three applications, making it possible to uncover firstly user preferences with respect to the agents in the applications and secondly users' behaviour toward the agent in the application, e.g. whether they perceive formally dressed agents in more serious applications such as banking to be more trustworthy.

The personality of the agent is the final issue to consider in the design of embodied agents for rich social interactions. Between applications participants may desire agents to project different personalities and in order to begin to create such personalities it is constructive to first assess users' expectations toward agents' personalities before

creating fully functional interactive agents. Typically assistants in retail domains need to be competent, as discussed previously. In addition, these assistants should also be polite, agreeable and sociable and in certain scenarios the agents might be liked better if they are cheerful and friendly. Finally, although trustworthiness is not a personality trait, it is an essential dimension for users to perceive in agents in order to establish a social relationship with them (Cassell & Bickmore, 2001). The level of expected agent trustworthiness may vary depending on the context of the application and also by the agent's appearance and behaviour in the application and so it is important to assess this aspect of the agent during the evaluation.

The contents of this chapter describe the construction of an interactive communicative system designed to be used as an experiment platform to assess 3D embodied conversational agents who appear as assistants in virtual retail environments, where the agents' role, appearance and personality (including trustworthiness) were assessed. The development of the agents and the three VRML retail applications is also described together with the experiment predictions, procedure and results.

6.2 Interactive System Design

Good interface design requires good notation to describe the operation of the system and diagrammatic notations are used frequently in the design of dialogue systems. Flowcharts were used successfully here, with the advantage of being simple to read and edit, and they are also convenient guides from which to program the dialogues. Figure 6.2 shows a simplified section from a cinema box-office application flowchart.

The user is given a specific task (e.g. book two tickets for Casablanca on Saturday at 3pm). Starting at the top of the flowchart, the system prompts the user for the information and it is likely that the user will present some or all of the required information to the agent. The system processes this information and fills natural language slots (e.g. movie, time, date) accordingly. If the system does not receive all the required information, it requests the remainder. After it receives all the required information the system proceeds to the next state, confirmation. If the confirmation is affirmative, the state changes and the system proceeds to the payment section. If the response is negative, the state changes, this time correcting information. These

flowcharts were used as the initial starting point for the development of the application dialogues. The dialogue manager combined these dialogues with the code to describe the VRML application environments and the 3D embodied agents. This system architecture is described in more detail below.

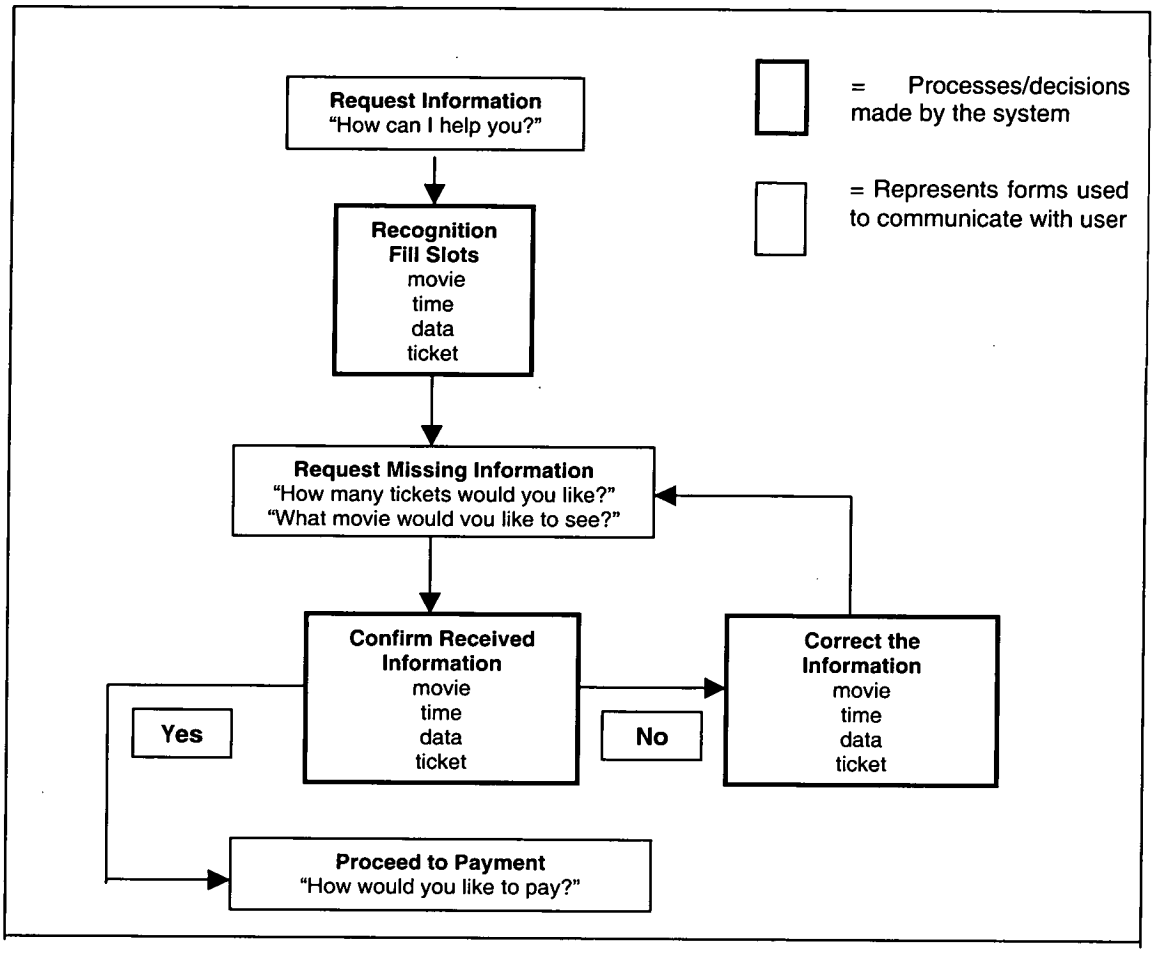


Figure 6.2 Extract from Cinema Application Flowchart

6.2.1 System Architecture

The system architecture is illustrated in Figure 6.3. It centres on a multi-user client-server architecture, which is an open source Java-based application. The code to describe the VRML application environments (cinema, travel, bank) is stored on the server PC. Networked to the server is the user client PC, which runs a Java applet with a VRML browser. The speech input to the system is captured on this client machine via a Nuance 7.0 speech recogniser and relayed to another PC, through the multi-user server. It is this second PC that runs the embodied agent applet. The applet links the dialogue

manager and a series of dialogue extensions, which are specific to the multi-user, DeepMatrix (2001) software (e.g. agent proximity sensors in the VRML world). The applet also has a multi-user client extension with specific network handling functions.

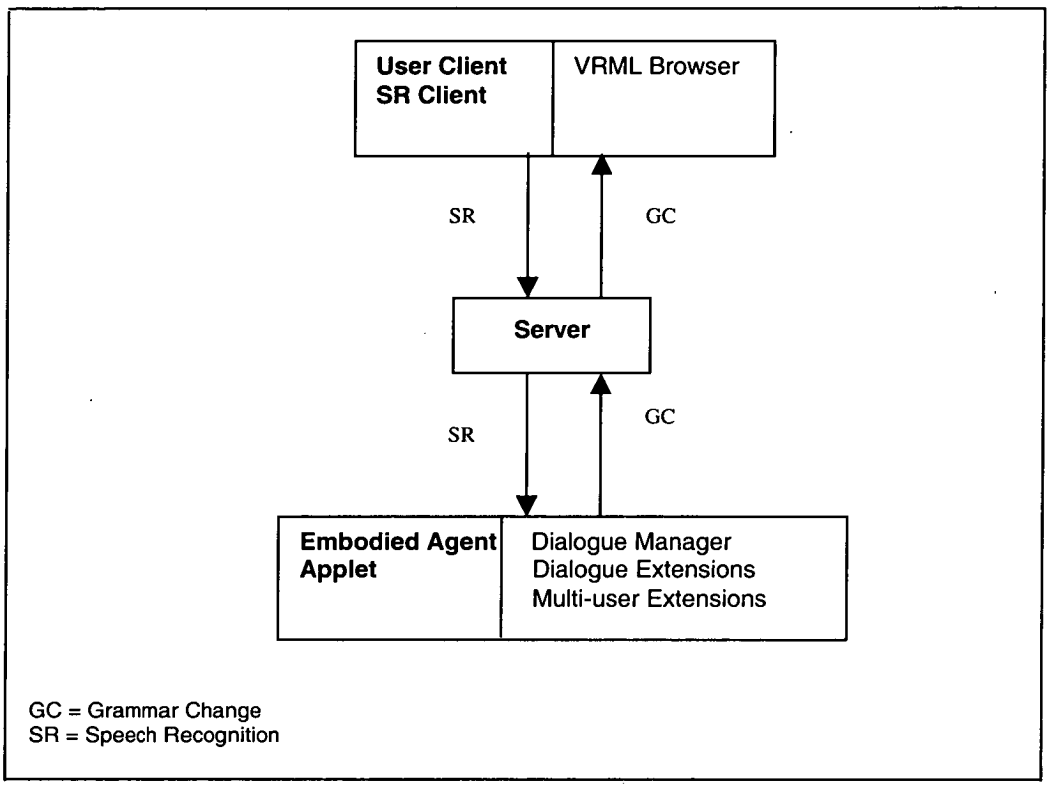


Figure 6.3 Illustration of System Architecture

6.2.2 Dialogue Manager Functionality

Dialogue models by their nature are logical and structured in design. Finite state machines are used to build such structured systems, where each finite state is an interaction that occurs over a pre-defined set of stable states. Shadbolt (1989) highlights some points for the development of planned dialogues. He states, in order to develop such a system it must be capable of the following:

- Goal interaction – once the goal of the conversation has been established the system must be able to deal with these goals changing.

- Top down planning – check first if a plan is feasible before commencing, since there is no point in interacting with a system that does not have the ability to complete a certain task.

- Using meta-knowledge in planning – this concept involves controlling the dialogue of a system. This means limiting the user within reason to the available vocabulary and grammar of a system.

- Intelligent backup on failure – if the system is unable to complete a task it must have the ability to explain this to the user.

- Non-monotonic logic – the system must be able to deal with information that is incomplete or inconsistent.

The functionality of the dialogue manager centres on condition-result pairs. If a condition is satisfied, the resulting action takes place. The client accepts the speech input from the user and passes this information to the dialogue manager, through the server PC. If a condition is satisfied there is a grammar change, which is relayed back to the multi-user client PC, again through the server, and the dialogue progresses to a new state. The finite state dialogues are written in a Dialogue Editor, which is Java-based (Figure 6.4), where the condition-result pairs are entered and parsed. Integral to the dialogue manager functionality are a series of predefined Java functions that can be programmed into the code. These include functions relating to the gesturing of the agent, speech concatenation functions and functions to change the grammar to the corresponding state.

Using the dialogue flowcharts as guides, individual Nuance™ grammar files were created for each application (cinema.grammar, travel.grammar, bank.grammar). Each grammar file contains a number of top-level grammars (e.g. Date, Time, Movie). Each of these top-level grammars is capable of calling one or more sub-grammars (e.g. DayofWeek, TimeofDay, MovieName), which in turn list the natural-language slots for that particular sub-grammar (DayofWeek = [monday, tuesday, wednesday, etc.]). For a condition to be satisfied and the resulting action to be taken, these natural language slots must be filled. Once the action has been completed, a grammar change takes place, a new sub-grammar is called on and the dialogue moves to a new state. This grammar change is initiated by the dialogue manager and relayed back to the multi-user client via the server. In the new state the user is prompted to give a reply that should fill new slots relevant to this new state.

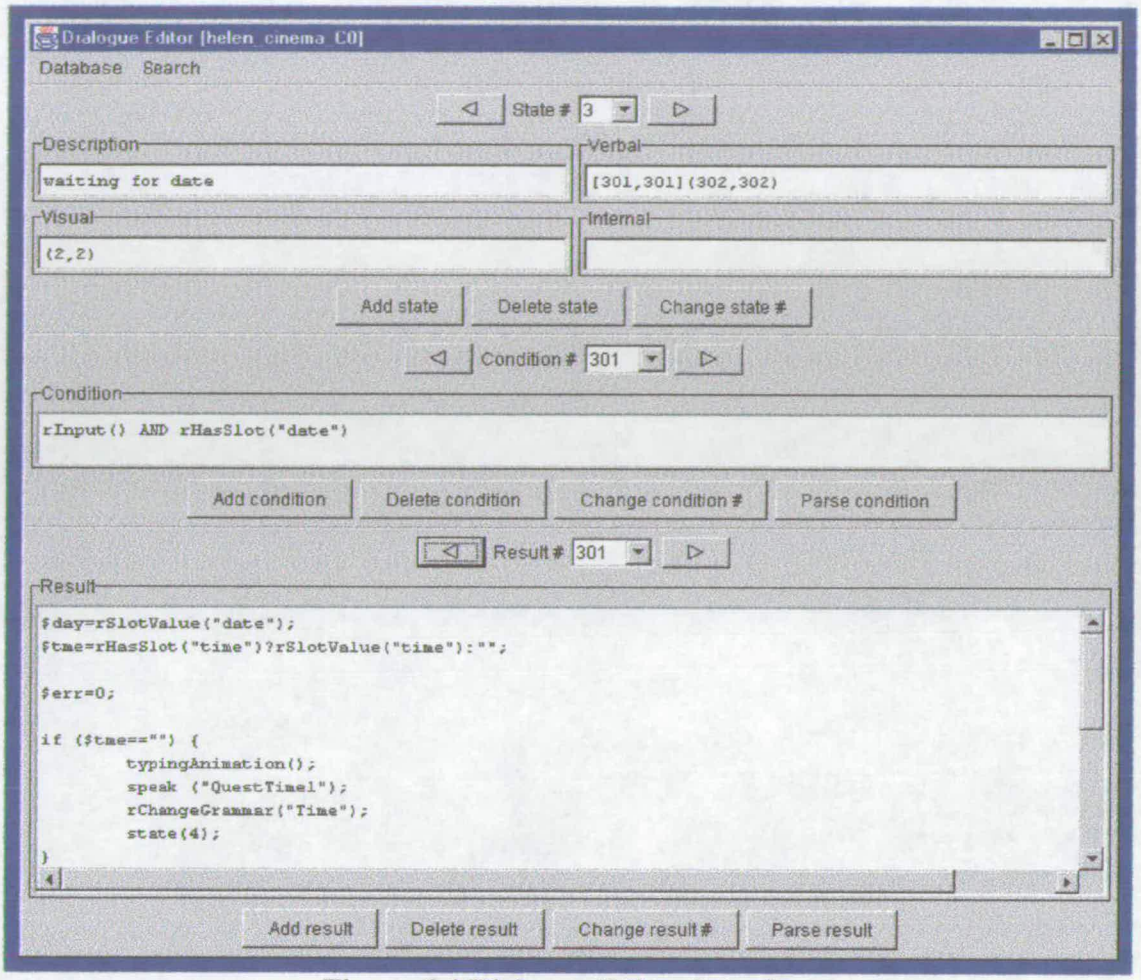


Figure 6.4 Dialogue Editor Interface

6.2.2.1 Example

Table 6.1 illustrates a section of the code for the cinema application and an explanation of the functionality of this code follows. The dialogues for all three applications were created in this way. In the cinema application some of the top-level grammars are called Movie, Date and Time.

If the dialogue manager is in State(3), "wait for day", it is waiting for Condition 301 or Condition 302 to be satisfied, i.e. it is waiting for the day of the week the user would like to see the movie. For Condition 301 to be satisfied, the user must have filled the natural language slot, "date". If this happens the system then checks to see how much more information it needs to continue. It may need the TimeOfDay ("time") or the NumberOfTickets ("tik"). If the system needs to get the "time", the user is asked: "At what time would you like to see the movie?".

State	Condition	Result

State [3]
 	
description: wait for day
verbal: [301](302)
	Condition 301
	
	rInput()
	AND
	rHasSlot("date")
	
	Condition 302
	
	rInput ()
		Result 301
		
		$day=rSlotValue("date");
		$tme=rHasSlot("time")?rSlotValue("time"):"";
		
		if ($tme=="") {
		say ("At what time would you like to see the movie?");
		rChangeGrammar("Time");
		state(4);
		}
		else if ($tik=="") {
		say ("How many tickets would you like?");
		rChangeGrammar("Tickets");
		state(5);
		}
		else {
		say("Certainly. You want "+$tik+" tickets, for the movie +$mvi+", at "+$tme+" o clock, on "+$day+". Is this correct?");
		rChangeGrammar("Confirm");
		state(6);
		}
		Result 302
		
		if ($err<2) {
		say("I'm sorry, I didn't understand that. Please select a day");
		shruggingAnimation();
		$err=$err+1;
		}
		else {
		say("Please call the Cinema Helpline for assistance");
		state(0);
		}

Table 6.1 Example of Condition-Result Pair (Cinema Application)

Following this, there is a top-level grammar change to Time, which calls the sub-grammar TimeOfDay and the state is changed to State(4). The system now waits to receive a natural language slot called "time". Similarly if the system already has the "time" but not the number of tickets, it asks the user: "How many tickets would you like?" The top-level grammar is changed to Tickets, which calls the sub-grammar NumberOfTickets and the state is changed to State(5). The system now waits to receive a natural language slot called "tik".

In the third instance the system may have received all the information it needs and then confirms this with the user by saying: "You would like two tickets, for the movie Casablanca, at three o'clock, on Saturday. Is this correct?" The system then changes the top-level grammar to Confirm and moves to State(6), to wait for confirmation from the user.

In the event of Condition 302 being satisfied this means the input utterance has not been understood and the system then responds: "I'm sorry, I didn't understand that. Please select a day." If after three attempts the utterances are still not understood, the system displays graceful degradation by advising the user to obtain assistance from the Cinema Helpline.

6.2.2.2 Application Tasks

Both the travel application grammar and the banking application grammar have been designed in an identical manner to the cinema grammar. In all three applications the dialogues have been built to satisfy specific tasks in each application. Table 6.2 describes the tasks and the information needed for a complete interaction in each application.

6.2.2.3 Dialogue Features

The application dialogues contained a number of features to ensure a natural interaction between the agents and the participants. As mentioned earlier the dialogue manager consists of a set of states and transitions between them. At its simplest the transition occurs due to a condition result pair, (i.e. a question-answer unit). However, users frequently over-answer (Hagen, 2000), providing more information than they are specifically asked for, but it is important that the system stores any of this extra information in the event that it may be needed later in the interaction and the system displays increased efficiency if it does not ask for information if it has already been given. The dialogue manager functionality facilitates such over-answering. For instance, top-level grammars allow many sub-grammars to be called and this makes it possible to capture natural language slots associated with any specified sub-grammar, within the active top-level grammar. This information may not necessarily be related to the top-level grammar in question, but is relevant to the completion of the task (Table 6.3).

Application	Example Task	Information Required (per sub-grammar)
Cinema	Book 4 tickets for the movie Casablanca on Wednesday at 6pm	MovieName DayOfWeek TimeOfDay NumberOfTickets PaymentMethod SecurityNumber
Travel	Book a flight from Edinburgh to London, departing on August 16th, 6pm, returning on August 19th, 10pm.	DepartCity DestinationCity DepartDate ReturnDate PaymentMethod SecurityNumber
Bank	Transfer £50 from saving account to current account on September 2nd.	SecurityNumber Amount FromAccount ToAccount DateofTransfer

Table 6.2 Application Tasks Used in the Evaluation

Typical Response			
		Slots Required	**Optional Slots**
Agent	"How many tickets would you like?"	Tickets	Movie Time Day
User	"Five tickets please."	Tickets = filled	

Over-Answered Response			
		Slots Required	**Optional Slots**
Agent	"How many tickets would you like?"	Tickets	Movie Time Day
User	"Five tickets please for Casablanca."	Tickets = filled	Movie = filled

Table 6.3 Example of Over Answering

Typical Response			
		Slots Required	**Optional Slots**
Agent	"You would like five tickets for the movie Casablanca. Is this correct?"	Confirm	Tickets Movie
User	"No."	Confirm = filled	
Agent	"I'm sorry. How many tickets would you like."	Tickets	Tickets Movie
User	"Four tickets please."	Tickets = filled	

Mixed Initiative Response			
		Slots Required	**Optional Slots**
Agent	"You would like five tickets for the movie Casablanca. Is this correct?"	Confirm	Tickets Movie
User	"No, I didn't say five tickets. I would like four tickets please."	Confirm = filled	Tickets = filled
Agent	"I'm sorry. You would like four tickets for the movie Casablanca. Is this correct?"	Confirm	Tickets Movie
User	"Yes, it is"	Confirm = filled	

Table 6.4 Example of Mixed Initiative Dialogue

Because the system is not only waiting for the information it asked the user for, it can also accept extra information. This can result in a flexible interaction in the event of over-answering occuring. However it is important to point out that "users still have to be guided very carefully into making utterances that the speech recognition software has a good chance of being able to understand. Poor speech recognition will result in poor quality applications, no matter how flexible the dialogue manager" (Bohlin, 1999).

Mixed-initiative dialogue allows a user to over-answer, correct, reject and accept information. At the confirmation stage of the interaction, the user has the option to accept the information as correct and then proceed to the payment stage. If the information is incorrect the user can reject the confirmation by simply saying "No". It frequently happens that the user rejects the confirmation and corrects the incorrect information in the same utterance. By programming the dialogue manager to handle this type of utterance it prevents the system from asking questions to correct the information, when the user has just done so (Table 6.4).

6.2.3. Creating the VRML Environments

VRML is an excellent software tool for the creation of interactive simulations that incorporate animation, and real-time, multi-user participation. The 3D worlds used in this research were created using VRML97, the international standard file format for describing interactive 3D multimedia in 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 6.5 and 6.6 illustrates the appearance of the environments with the embodied agents.

6.2.4 Creating the Embodied Agents

6.2.4.1 Agents' Appearance

MetaCreations Poser 4.0, a character animation software tool was used to create male and female animated agents. With this software tool it is a straightforward process to design different agents as clothes and colours can be selected from a large database. 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 polygon count of the agent files was reduced from 1.4 MB to 400 KB. Specifically triangular polygons were merged to form quadrangular polygons, parallel lines were merged and since the agents did not need to walk in the virtual worlds the entire leg section of the each of the agents could be merged. Little compromise was made with respect to the visual appearance of the agents when they were optimised, but when the agents appeared in the VRML environments, the reduction in polygon count did result in improved efficiency and the reduction in computational power did improve the real-time mobility.

After optimisation was complete the files were exported to VRML97. While in this format the geometry of the agent (e.g. upperarm, lowerleg etc) could be altered further. This was necessary as some polygons lost during optimisation had to be replaced. The code was finally altered to fit into the H-Anim specification which is 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." (Humanoid Animation, 2001).

6.2.4.2 Agents' Non Verbal Behaviour

Head nodding and eyebrow raising were also included at appropriate pitch accents. When asking questions the agent raised its voice slightly, stressed the main word of the sentence and raised its eyebrows. During an affirmation the agent nodded, raised its eyebrows and blinked at the end of the sentence.

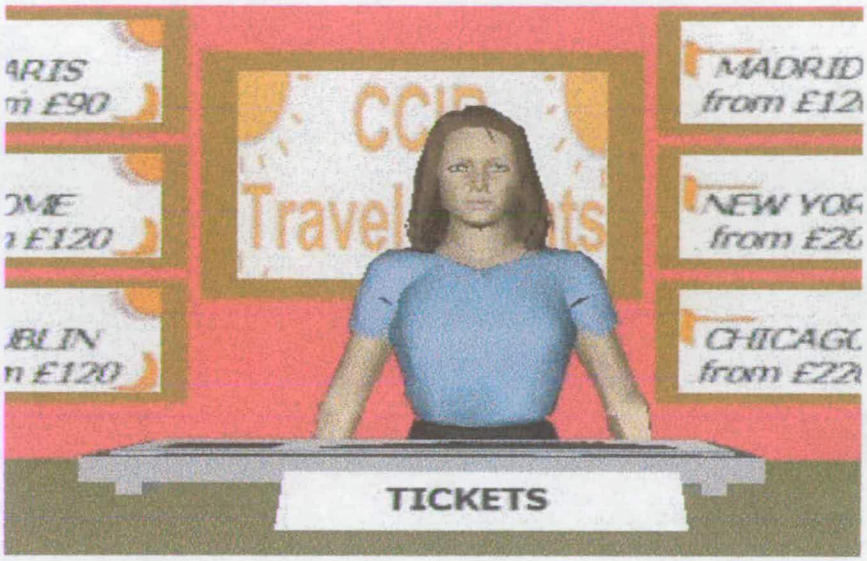


Figure 6.5(i) Informal Female

Figure 6.5(ii) Formal Female

Figure 6.6(i) Informal Male

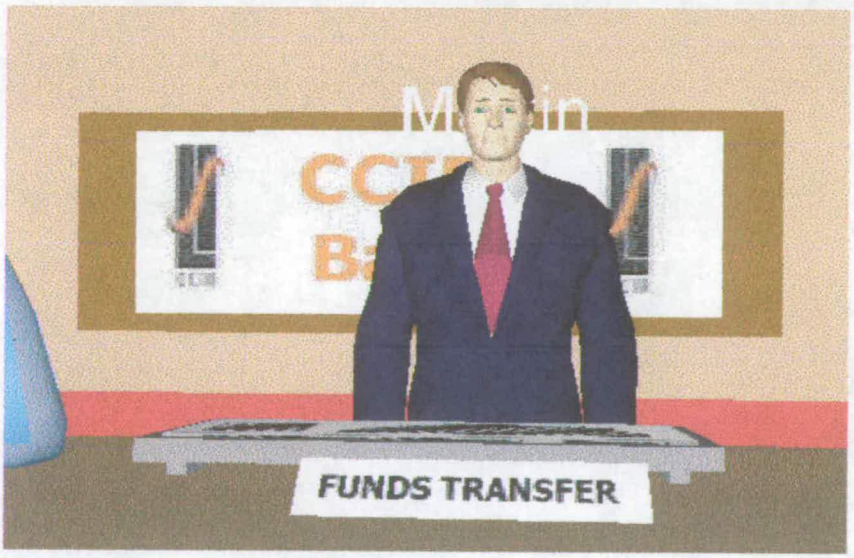


Figure 6.6(ii) Formal Male

The agents directed gaze towards the user during pauses in the conversation, when the customer was asking questions and when the agent was at the end of a turn. A more in-depth discussion of this non-verbal behaviour is presented in Chapter 4. When the agents were processing information they turned to look at the virtual computer, which appeared on the desk in the virtual world. They also had a typing gesture, which gave the impression that they were typing in the relevant information. The agent always returned to look at the customer at the end of an utterance.

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. The H-Anim specification defines abstractions for these segments and joints to allow a human body to be described in a structured and standardized way. For example:

```
DEF hanim_l_shoulder Joint {
            name "l_shoulder"
            center 0.167 1.36 -0.0518
      children   [
   DEF hanim_l_elbow Joint {
                     name "l_elbow"
                     center 0.196 1.07 -0.0518
              children      [
      DEF hanim_l_wrist Joint {
                              name "l_wrist"
                              center 0.213 0.811 -0.0338
           children   [
        DEF hanim_l_hand Segment {
                                        name "l_hand"
         ... }
          ]
        }
      DEF hanim_l_forearm Segment {  name "l_forearm"
       ...
      }
     ]
   }
   DEF hanim_l_upperarm Segment {  name "l_upperarm"
    ...
   }
 ]
}
```

Using this specification it was possible to obtain access to the joints of the agent and alter the joint angles to create particular gestures. Only four gestures were created (nodding, waving, shrugging and typing) for the embodied agents. An in-depth discussion about the large field of character animation is beyond the scope of this thesis, but it can be stated that to create any gestures it is important to understand the dynamics and physical movement of how the human joints work (Balder, 1993; Ratner, 1998). The gesture animations for this research were defined as action functions in the dialogue manager and these were displayed in parallel with particular speech utterances. For example, if the agent says: *"I'm sorry, I didn't understand"*, it also displays the shrugging gesture to accompany the speech and this further indicates mis-recognition. In addition to this the agents also displayed beat gesturing (Cassell, 2000) during the conversation. Mouth movements or visemes were also created to give the illusion of lip movement.

6.2.4.3 Agents' Voices

One male and one female voice recorded the necessary spoken prompts for the male and female agents respectively. These voices were pre-evaluated before using them as the official voice outputs for the agents. Ten participants (5 male, 5 female) listened to a selection of recorded example sentences after which they were asked to comment on the naturalness, clarity and likeability of both the male and female voices. The result of this pre-evaluation was positive with no indication that differences existed between the selected male and female voices. Both were thought to be clear, natural and likeable. These voices were then used as the voice output for the male and female agents and the prompts corresponding to each application were recorded as .WAV files. Varying intonational prompts were recorded for questions and statements (question = rising intonation; statement = falling intonation). When an entire output utterance was to be constructed, the dialogue manager called the file that stored the corresponding prompts. The relevant prompts were concatenated in a particular order to produce a plausible output sentence.

6.3 System Evaluation

New design and evaluation techniques are continually under development. These new techniques focus on users' involvement at the design stage and at the evaluation stage. It is necessary to understand their motives, their interests and their needs (Norman, 1998). Large participant samples, balanced for age and gender, have provided important information with respect to the design of embodied agents. These types of user-centred evaluations highlight the users' likes and dislikes with respect to the system leading to better designed systems in the future, which are responsive to user attitudes.

However, before the participants were invited to evaluate and interact with the 3D embodied agents, the system itself had to be evaluated. As explained in Chapter 3 there are three types of system evaluation: expert, non-expert and cognitive walkthrough. Combinations of all three types of evaluation were used to identify any problems with this interactive system.

Throughout the design of the system frequent reference has been made to the evaluation heuristics of Molich & Nielsen (1990) as presented in Chapter 3. An expert evaluation of the system was conducted to assess the system's functionality with respect to the experiment, to assess the interface and to identify any specific problems. Sanders and Scholtz (2000) defined five possible metrics for the evaluation of embodied conversational agents. These metrics were used as a guide during the expert evaluation and expanded on further for the actual evaluation of the 3D embodied agents.

- *Unconscious embodied mechanics of conversation*: ensuring the agents generate turn-taking behaviours and ensuring the user understands the embodied behaviours (nods, beat gestures).

- *Creating accurate user expectations:* creating accurate expectations about what the user can say.

- *Individuation (the perceived character or personality of the agent):* ensuring the agent's personality is consistent and its appearance is relevant.

- *Use of gesture to communicate content:* investigating if the agent and user can talk about objects in the domain of discourse with gesture and ensuring this occurs at the appropriate times.

- *Functionality as an ally of the user:* making sure the agent has the ability to guide the user to complete the task.

Using these five points as a guide, separate groups of experts evaluated the system. Four dialogue engineering experts debugged the speech systems ensuring that the experimental tasks could be completed. Three interface designers evaluated the VRML application environments, ensuring their geometric descriptions were correct and the presence of the embodied agents in the applications looked appropriate. This included an evaluation of the agent's verbal and non-verbal behaviour.

Finally the entire experiment procedure was evaluated by two cognitive psychologists, where particular attention was paid to the impact the interaction has on the user, the cognitive processes required by the user and any learning problems the may occur during the interaction. After changes and corrections were made to the system it was deemed ready to be used as an experiment platform. A repeated measures experiment was designed where participants were invited to interact with all four agents (male, female; formal, informal) in all three applications. The experiment predictions, procedure and results are detailed below. For clarity the experiment implementation is illustrated in Figure 6.7.

6.4 Experiment Predictions

1. Participants would encourage the presence of the agents in the retail applications. This prediction was made based on the results of previous experiments, where customers passively viewed conversational agents in retail spaces and expressed a desire to actually converse with them to complete tasks.

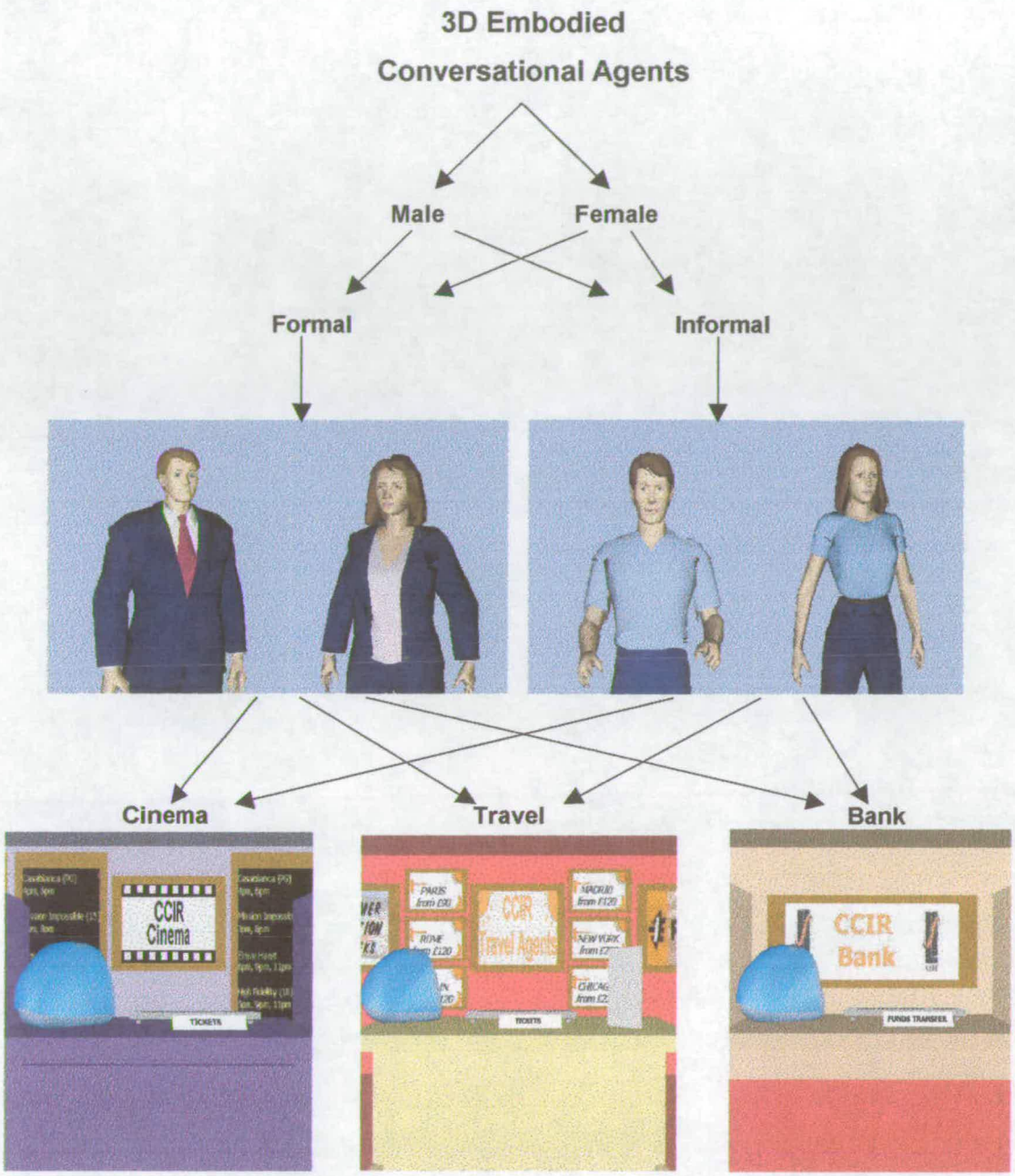


Figure 6.7 Illustration of Structure of Interactive Evaluation I

2. Participants would enjoy speaking to the agents equally in all three applications. This predication 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 completing tasks.

3. 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.

4. Responses would be similar for both the male and female agents. This can be stated given that the issues, which may have caused significant gender differences in Chapter 5, were eliminated. Firstly differences in attitudes for the male and female voices and secondly differences due to the gesturing of the male and female agents were eliminated through pre-evaluations. It must also be stated that although Nass et al. (1997) 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, this was prevented as both the male and female agents had the same output utterances (scripts) and these were by design as neutral as possible, with no intentional inclusion of utterances that were regarded as being more masculine than feminine or vice versa.

5. The perception of aspects of the agents' personalities that were examined would be similar within applications as the agents (male and female; formal and casual) were by design similar and all agents had identical output utterances. Between applications differences might emerge.

6.5 Experiment Design

To test these hypotheses a repeated measures experiment was designed. Two types of male and female agents were assessed as assistants in three VRML retail application environments. The first was a smartly dressed formal agent, and the second a casual

informally dressed agent. The dependent variables in the experiment were the responses to the individual statements in the application questionnaire; the assistant questionnaire; the application comparisons; and the responses given during a post experiment interview. The independent variables were agent gender, agent appearance (formal, informal), 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 order of presentation of the three applications was balanced amongst the participants. Four similar tasks were created for each application, which were randomised amongst the agents. Examples of the task for each application were provided earlier in the chapter (Table 6.2). The difference in the tasks within the applications simply involved changes in key information provided (e.g. MovieName, DayOfWeek, ReturnDate).

A total of 36 participants took part in the experiment, distributed according to gender and age as shown in Table 6.5. Effects of between-subject variables of age and gender were investigated.

	Interactive Evaluation I Evaluation		Total
Participant	**Male**	**Female**	
Age 18-35	6	6	12
Age 36-49	6	6	12
Age 50+	6	6	12
	18	18	36

Table 6.5 Analysis of Participants by Gender and Age Group

The experimental procedure required participants first of all to read an information sheet regarding the application they were about to see (Appendix 3.1). In all cases the participants were asked to observe the assistant and the application carefully. They were also told that they might be asked for security number information, which was presented to them before the interaction began. Questionnaires were designed using the blueprint method and an initial exhaustive list was created with content areas defined as application and agent and manifestations defined as behaviour, functionality and appearance. As the participants actually conversed with the agents it was possible to include usability attributes that directly assessed aspects of the agent's personalities. This was not possible in previous evaluations.

After the conversation participants were asked to complete a questionnaire relating to the assistant. They were also asked to rank each assistant on a scale of 1-10 (low-high). The usability attributes were described as 7-point Likert attitude questionnaire statements as shown in Table 6.6. Within the questionnaire, statements were balanced for polarity (equal number of positively and negatively worded stimulus statements). An equal number of statements were asked of all agent types. When participants had seen all four agents in an application they completed a short attitude questionnaire (again 7-point attitude statements) relating to the application. During the cognitive walkthrough evaluation of the experiment design (Chapter 3) it was felt that no constraints should be placed over the participant sample as regards computing experience since the participants were only asked to converse with the agents and were not required to use the mouse or keyboard.

Questionnaire Statements	
Applications	1. I would use this service myself.
	2. I felt this service was difficult to use.
	3. I did not think this service was a good idea.
	4. I think this service is convenient.
Agents' Voices	5. I liked the assistant's voice
	6. I did not like speaking to the assistant.
	7.The assistant's voice was annoying.
	8. The assistant spoke naturally.
	9. I felt the assistant understood me.
Agents' Personality	10. The assistant was polite.
	11. The assistant was friendly.
	12. The assistant was competent.
	13. The assistant was unsociable.
	14. The assistant was cheerful.
	15. The assistant was agreeable.
	16. The assistant was trustworthy.
Agents' Appearance	17. I did not like the appearance of the assistant.
	18. The assistant was not dressed appropriately for this service.
	19. The assistant appeared lifelike.

Table 6.6 Questionnaire Statements

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.

Title	**Interactive Evaluation I: 3D ECA VRML Retail Applications**	
Design		One Independent Sample
Predictions	6.1	The deployment of 3D ECA would be accepted in all applications.
	6.2	Participants would enjoy speaking to the agents in all applications.
	6.3	The stereotypes created (formal, informal) would be suited to different applications
	6.4	No differences would emerge based solely on the gender of the agents (including differences with respect to voice)
	6.5	Within applications attitudes to agents' personalities would be similar.
Dependent Variables		Applications Attitude Questionnaire Responses (1-7 Likert scale)
		Agent Attitude Questionnaire Responses (1-7 Likert scale)
		Application Comparisons
Other Data		Interview Answers
(Experiment) Independent Variables:	1	Agent Type (2 levels)
	2	Agent Gender (2 levels)
	3	Application (3 levels)
(Participant) Independent Variables	1	Gender (2 levels),
	2	Age Group (3 levels),
Extraneous Variables:	Order	Agent presentation randomised within applications.
		Task randomised within applications.
		Applications balanced amongst participants
Location		Edinburgh - CCIR Premises, Central Edinburgh
Cohort		N = 36 50% male, 50% female
Remuneration		£20
Duration:		90 minutes

Table 6.7 Summary Table of Interactive Evaluation I

6.6 Results

A series of repeated measures ANOVA, taking agent gender, agent type and application as the independent variables, were completed for the usability attributes in the application questionnaire and the agent questionnaire. Each ANOVA table presents the

results for within subject effects, together with interactions and results of the participant between subject variables of age and gender. For each usability attribute an explanation and discussion of any significant results are presented and the ANOVA tables are also listed.

It should first be noted that in total 432 interactions took place during the course of the experiment: 36 participants interacted with 4 individual agents in 3 different applications. Upon analysis of these interactions, data performance information was retrieved. A total of 77 interactions were in some way disrupted. These disruptions included attempting to start the conversation more than once; correcting information during the confirmation stage; or repeating details that the system did not capture at the time of user input. Examples of these disruption can be found in Table 6.8. The data performance analysis also showed that 8 interactions were incomplete. Incomplete interactions occurred when the system mis-recognised an input statement more than two times. Analysis indicated that no significant effects for these disrupted or incomplete interactions emerged between or within variables.

Type of Disruption	Example
More than one start	Assistant: "Hello, how can I help you" User: "Three tickets for Casablanca" Assistant: "I'm sorry. I didn't understand that. How can I help you"
Correcting during confirmation	Assistant: "You want three tickets for Casablanca on Friday at 6pm. Is that correct?" User: "No" Assistant: "I'm sorry. You would like to see Casablanca, is that correct?"
Repeating details	User: "I would like three tickets please" Assistant: "I'm sorry. I didn't understand that, how many tickets would you like?"

Table 6.8 Examples of Disruptions

6.6.1 Application Ratings

After interacting with all four agents in an application, participants were asked to rate the application on a scale of 1-10 (low-high). The mean rating scores suggest a moderately positive response to all three applications. The ANOVA (Table 6.9), taking experiment application as the independent variable showed no significant effects for applications.

	Sum of Squares	df	Mean Square	F	p
Application	2.574	2	1.287	.979	.382
Application * P(Age)	1.759	4	.440	.335	.854
Application * P(Gender)	2.389	2	1.194	.908	.409
Error(Application)	78.889	60	1.315		

Table 6.9 Applications Rating ANOVA

No interactions for the between subject effects of age and gender were evident. Table 6.10 presents the mean scores for each of the applications; the cinema was rated the highest, followed by the travel agency and thirdly the bank.

Application	Mean Rating Score (max 10)
Cinema	6.56
Travel Agency	6.46
Bank	6.12

Table 6.10 Mean Rating Scores for Application

6.6.2 Attitude to Applications

Participants also expressed their opinions about the application using a 4-item 7-point Likert statement questionnaire.

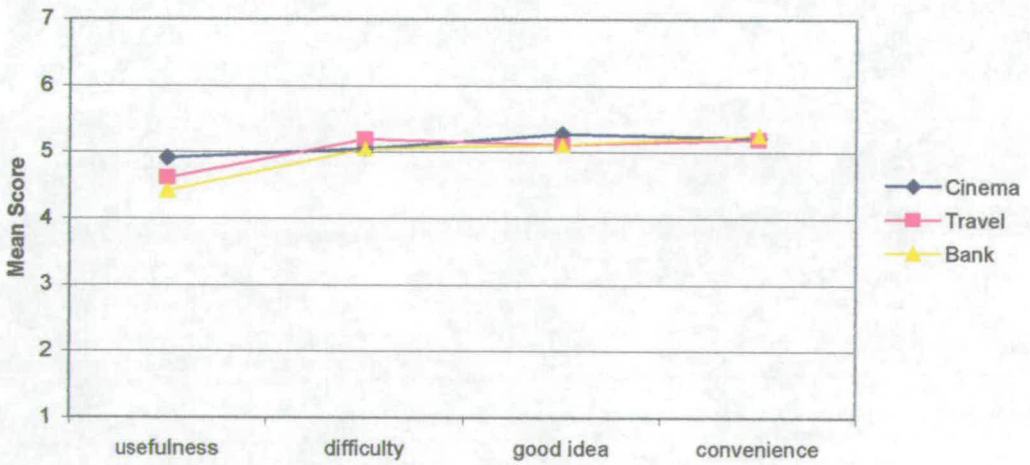


Figure 6.8 Usability Attributes for Applications

6.6.2.1 Usability Attribute – "Usefulness"

A largely positive response, with no significant differences between applications, was found when participants were asked if they would use the service themselves (Table 6.11).

I would use this service myself	Sum of Squares	df	Mean Square	F	p
Application	4.574	2	2.287	2.465	.094
Application * P(Age)	4.759	4	1.190	1.282	.287
Application * P(Gender)	2.722	2	1.361	1.467	.239
Error(Application)	55.667	60	.928		

Table 6.11 ANOVA for Usability Attribute "Usefulness"

6.6.2.2 Usability Attribute – "Difficulty"

Overall the participants felt that the applications were not difficult to use and no significant differences or interactions emerged (Table 6.12).

I felt this service was difficult to use	Sum of Squares	df	Mean Square	F	p
Application	.463	2	.231	.374	.689
Application * P(Age)	3.926	4	.981	1.587	.189
Application * P(Gender)	3.389	2	1.694	2.740	.073
Error(Application)	37.111	60	.619		

Table 6.12 ANOVA for Usability Attribute "Difficulty"

6.6.2.3 Usability Attribute – "Good idea"

This usability attribute showed that participants felt the three applications were good ideas with no significant difference evident (Table 6.13).

I did not think this service was a good idea	Sum of Squares	df	Mean Square	F	p
Application	.667	2	.333	.371	.692
Application * P(Age)	1.167	4	.292	.325	.860
Application * P(Gender)	.222	2	.111	.124	.884
Error(Application)	53.889	60	.898		

Table 6.13 ANOVA for Usability Attribute "Good idea"

6.6.2.4 Usability Attribute – "Convenience"

Encouragingly the results showed that participants felt all three applications were equally convenient (Table 6.14).

I think this service is convenient	Sum of Squares	df	Mean Square	F	p
Application	5.556E-02	2	2.778E-02	.029	.971
Application * P(Age)	1.611	4	.403	.427	.788
Application * P(Gender)	2.019	2	1.009	1.071	.349
Error(Application)	56.556	60	.943		

Table 6.14 ANOVA for Usability Attribute "Convenience"

6.6.3 Application Comparisons

After participants had interacted with all four agents in all 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. In disagreement with the statistical evidence from the application questionnaire where all three applications were rated equally, these comparisons produced strong indications that the cinema application was the preferred choice *in comparison* to the other applications.

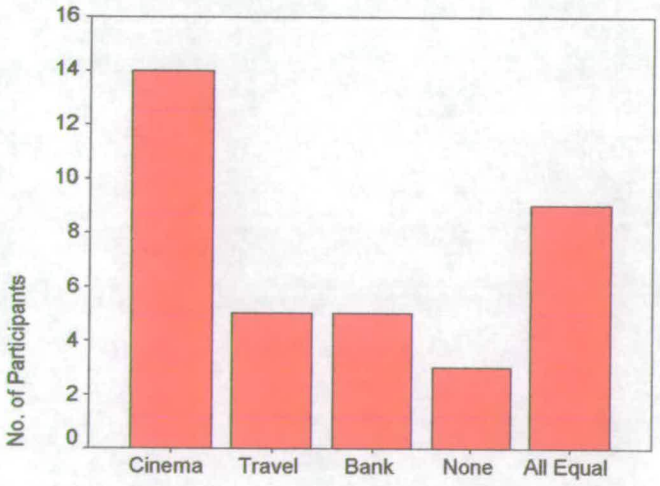


Figure 6.9 Application Comparisons

Some 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% liked all applications equally. A chi-square test showed that the cinema application was significantly preferred to the travel and banking applications (both at $p < 0.05$).

6.6.4 Qualitative Comments about the Applications

Qualitative comments explain the preference for the cinema application over the travel and banking applications. Of the 14 participants who declared a preference for the cinema application, 11 provided comments for their choice, which indicated that 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"*. With respect to the cinema in comparison to the travel agency the issue of confidence was restated when one participant said he would *"rather miss a movie, than a flight"* in the event of a system error occurring. Participants 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"*. In support of the use of embodied agents one participant commented that the systems were 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"*. Some individual comments made with respect to each of the applications are presented next.

6.6.4.1 Cinema Application

One participant was recorded as saying the following about the cinema application: *"I liked the service instead of having to wait in line and it was convenient to book in advance"*. Participants commented that more visual content would be an improvement. This included adding a text interface to provide additional feedback. Participants did experience delayed responses from the system as it was processing information and the general impression was that if the delayed responses in the system could be eliminated, this application would indeed be successful.

6.6.4.2 Travel Application

Although this application was thought to be a good idea and seemed easy to use, participants were concerned with delays in the system's responses. Due to these delays participants seemed to lack confidence in the system and did not think it would be able to handle more complicated flight information. For more *"critical"* tasks, lack of confidence can greatly reduce users' perceptions of that application (see Section 6.6.3). The inclusion of text in the interface was again suggested in order to provide extra feedback.

6.6.4.3 Bank Application

Participants felt uncertain about security, confidentiality and reliability with respect to this application and felt they should have the opportunity to use the keyboard to enter security numbers and amounts of money. Similar comments about the delays in the system were made with regard to this application.

The results from the application questionnaire and ratings showed no significant differences between the applications and participants found them all equally useful, easy to use, good idea and convenient. When asked to directly compare the three applications, there was a significant response for the cinema application. This finding was echoed in the qualitative analysis as participants explained they found they had less confidence in the capabilities of the more serious banking application. The results of the agents' questionnaires are presented next detailing attitudes to the agents' voices, appearance, and aspects of personality and also perceived trustworthiness. Although it is already known that the cinema application was preferred in comparison to the other applications, the following results will show if this had any impact on attitudes toward the agents in the contrasting applications.

6.6.5 Attitude to Agents' Voices & Conversation

A series of 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 analysis participants' attitudes to the attributes relating to the embodied agents as assistants. Attitudes toward

the assistants' voices and the mean scores for these attributes are presented in Figure 6.10.

6.6.5.1 Usability Attribute – "Liked voice"

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 (Table 6.15).

I liked the assistant's voice	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.727	2	.363	.333	.718
Application * P(Age)	3.481	4	.870	.797	.532
Application + P(Gender)	.727	2	.363	.333	.718
Error(Application)	65.556	60	1.093		
A(Type)	.113	1	.113	.101	.752
A(Type) * P(Age)	1.407	2	.704	.629	.540
A(Type) * P(Gender)	.280	1	.280	.251	.620
Error(A(Type))	33.542	30	1.118		
A(Gender)	4.280	1	4.280	2.847	.102
A(Gender) * P(Age)	1.907	2	.954	.634	.537
A(Gender) * P(Gender)	2.225	1	2.225	1.480	.233
Error(A(Gender))	45.097	30	1.503		
Application * A(Type)	1.532	2	.766	1.682	.195
Error(Application * A(Type))	27.333	60	.456		
Application * A(Gender)	1.116	2	.558	.914	.406
Error(Application * A(Gender))	36.611	60	.610		
A(Type) * A(Gender)	.280	1	.280	.443	.511
Error(A(Gender)*A(Type))	18.986	30	.633		
Between Subject Effects					
P(Age)	13.685	2	6.843	.947	.399
P(Gender)	33.891	1	33.891	4.693	.038
Error	216.653	30	7.222		

Table 6.15 ANOVA for Usability Attribute "Liked voice"

6.6.5.2 Usability Attribute – "Liked speaking"

Participants were also asked if they liked speaking to the assistant. Encouragingly the results showed that participants did like speaking to the assistants in all the applications, with no significant differences for agent type or for application (Table 6.16).

I did not like speaking to the assistant	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.616	2	.308	.237	.790
Application * P(Age)	9.245	4	2.311	1.777	.145
Application * P(Gender)	4.866	2	2.433	1.871	.163
Error(Application)	78.028	60	1.300		
A(Type)	.521	1	.521	.563	.459
A(Type) * P(Age)	3.014	2	1.507	1.630	.213
A(Type) * P(Gender)	3.521	1	3.521	3.808	.060
Error(A(Type))	27.736	30	.925		
A(Gender)	5.787E-02	1	5.787E-02	.043	.837
A(Gender) * P(Age)	.782	2	.391	.291	.749
A(Gender) * P(Gender)	.669	1	.669	.498	.486
Error(A(Gender))	40.292	30	1.343		
Application * A(Type)	3.792	2	1.896	1.686	.194
Error(Application * A(Type))	67.472	60	1.125		
Application * A(Gender)	.449	2	.225	.222	.801
Error(Application * A(Gender))	60.583	60	1.010		
A(Type) * A(Gender)	1.447	1	1.447	3.770	.062
Error(A(Gender)*A(Type))	11.514	30	.384		
Between Subject Effects					
P(Age)	19.088	2	9.544	.556	.579
P(Gender)	61.502	1	61.502	3.584	.068
Error	514.847	30	17.162		

Table 6.16 ANOVA for Usability Attribute "Liked speaking"

6.6.5.3 Usability Attribute – "Voice annoying"

Participants were asked is they thought the assistants' voices were annoying (Table 6.17). For this usability attribute a marginally significant effect was found for the agent gender where the male voice was perceived to be marginally more annoying that the female voice (mean female = 5.35; mean male = 5.02). This result was surprising given the fact that the pre-evaluation tests did not signify any gender differences. Later in this result section an explanation is offered for this unexpected occurrence.

The assistant's voice was annoying	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.060	2	.530	.604	.550
Application * P(Age)	4.509	4	1.127	1.284	.287
Application * P(Gender)	.431	2	.215	.245	.783
Error(Application)	52.694	60	.878		
A(Type)	9.259E-03	1	9.259E-03	.007	.933
A(Type) * P(Age)	.421	2	.211	.162	.851
A(Type) * P(Gender)	8.333E-02	1	8.333E-02	.064	.802
Error(A(Type))	39.056	30	1.302		
A(Gender)	11.343	1	11.343	5.523	**.026**
A(Gender) * P(Age)	1.144	2	.572	.278	.759
A(Gender) * P(Gender)	1.120	1	1.120	.546	.466
Error(A(Gender))	61.611	30	2.054		
Application * A(Type)	1.394	2	.697	1.312	.277
Error(Application * A(Type))	31.861	60	.531		
Application * A(Gender)	.421	2	.211	.363	.697
Error(Application * A(Gender))	34.806	60	.580		
A(Type) * A(Gender)	.333	1	.333	.259	.615
Error(A(Gender)*A(Type))	38.667	30	1.289		
Between Subject Effects					
P(Age)	18.699	2	9.350	1.014	.375
P(Gender)	21.333	1	21.333	2.314	.139
Error	276.556	30	9.219		

Table 6.17 ANOVA for Usability Attribute "Voice annoying"

6.6.5.4 Usability Attribute – "Voice natural"

Extending from this highly significant effects were also found for agent gender when participants were asked if the assistant spoke naturally (Table 6.18). Overall, participants perceived the female voice to be significantly more natural than the male voice (mean female = 5.05; mean male = 4.44).

Figure 6.10(i) Usability Attributes for Agents' Voice by Application

Figure 6.10(ii) Usability Attributes for Agents' Voice by Agent Gender

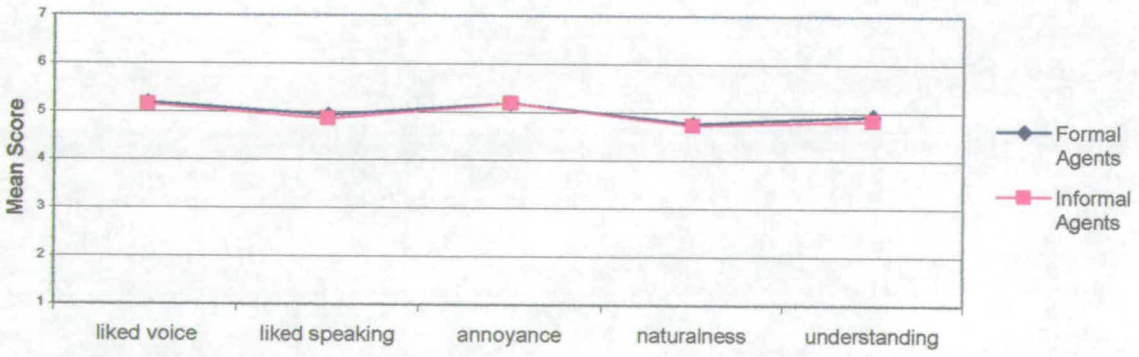


Figure 6.10(iii) Usability Attributes for Agents' Voice by Agent Type

The assistant spoke naturally	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	4.505	2	2.252	1.595	.211
Application * P(Age)	7.912	4	1.978	1.401	.245
Application * P(Gender)	6.449	2	3.225	2.284	.111
Error(Application)	84.722	60	1.412		
A(Type)	.148	1	.148	.138	.713
A(Type) * P(Age)	4.630E-03	2	2.315E-03	.002	.998
A(Type) * P(Gender)	3.704E-02	1	3.704E-02	.034	.854
Error(A(Type))	32.222	30	1.074		
A(Gender)	40.333	1	40.333	19.604	**.000**
A(Gender) * P(Age)	3.431	2	1.715	.834	.444
A(Gender) * P(Gender)	1.333	1	1.333	.648	.427
Error(A(Gender))	61.722	30	2.057		
Application * A(Type)	1.005	2	.502	.603	.550
Error(Application * A(Type))	49.944	60	.832		
Application * A(Gender)	3.014	2	1.507	1.651	.201
Error(Application * A(Gender))	54.778	60	.913		
A(Type) * A(Gender)	1.120	1	1.120	1.315	.261
Error(A(Gender)*A(Type))	25.556	30	.852		
Between Subject Effects					
P(Age)	35.199	2	17.600	1.735	.194
P(Gender)	15.565	1	15.565	1.535	.225
Error	304.278	30	10.143		

Table 6.18 ANOVA for Usability Attribute "Spoke naturally"

6.6.5.5 Usability Attribute – "Understanding"

Participants felt the agents understood them during the course of the interaction and no significant effects, for application, agent gender or agent type was found (Table 6.19). The quantitative results with respect to the agents' voices and the conversations the participants had with the agents, showed no attitude differences with respect to the conversations and all were not thought natural. Between the agent types the female voices were thought to be less annoying and more natural.

I felt the assistant understood me	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.449	2	.725	.564	.572
Application * P(Age)	12.495	4	3.124	2.432	.057
Application * P(Gender)	4.866	2	2.433	1.894	.159
Error(Application)	77.083	60	1.285		
A(Type)	1.120	1	1.120	.495	.487
A(Type) * P(Age)	.241	2	.120	.053	.948
A(Type) * P(Gender)	4.898	1	4.898	2.163	.152
Error(A(Type))	67.944	30	2.265		
A(Gender)	1.565	1	1.565	.614	.440
A(Gender) * P(Age)	.796	2	.398	.156	.856
A(Gender) * P(Gender)	3.343	1	3.343	1.311	.261
Error(A(Gender))	76.500	30	2.550		
Application * A(Type)	6.588	2	3.294	1.483	.235
Error(Application * A(Type))	133.306	60	2.222		
Application * A(Gender)	.755	2	.377	.162	.851
Error(Application * A(Gender))	140.083	60	2.335		
A(Type) * A(Gender)	.454	1	.454	.347	.560
Error(A(Gender)*A(Type))	39.222	30	1.307		
Between Within Subject Effects					
P(Age)	32.519	2	16.259	2.567	.094
P(Gender)	22.231	1	22.231	3.510	.071
Error	190.000	30	6.333		

Table 6.19 ANOVA for Usability Attribute "Understanding"

6.6.6 Attitude to Agents' Personality

Many aspects of the agents' personalities were assessed. The usability attributes reflect the most important characteristics that should be evident from agents who play the role of assistants in retail environments. The mean results are presented in Figure 6.11.

6.6.6.1 Usability Attribute – "Politeness"

An interaction between agent type and application was found with respect to the politeness of the agents (Table 6.20).

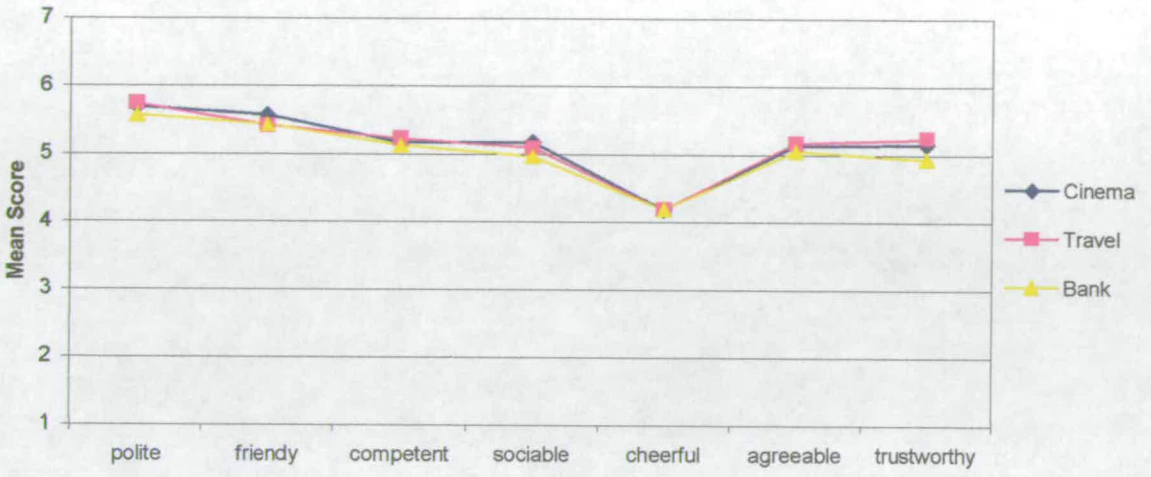


Figure 6.11(i) Usability Attributes for Agents' Personality by Application

Figure 6.11(ii) Usability Attributes for Agents' Personality by Agent Gender

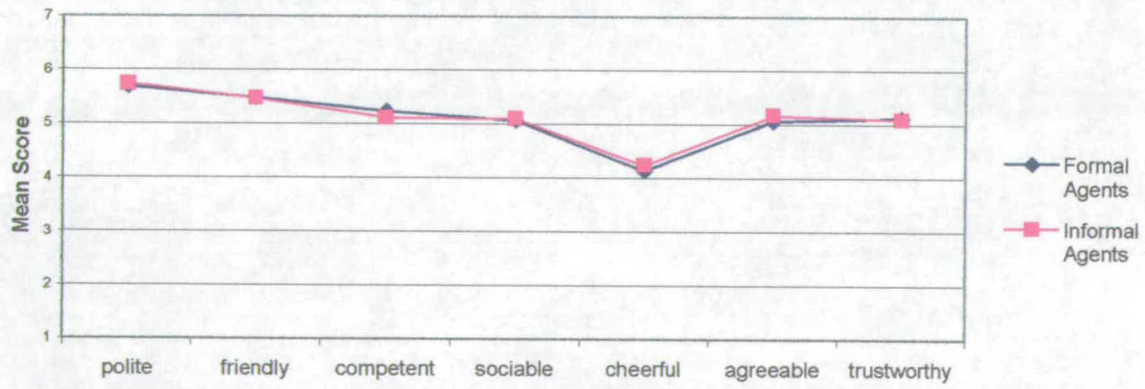


Figure 6.11(iii) Usability Attributes for Agents' Personality by Agent Type

The assistant was polite	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.199	2	9.954E-02	.241	.787
Application * P(Age)	1.481	4	.370	.896	.472
Application * P(Gender)	.421	2	.211	.510	.603
Error(Application)	24.806	60	.413		
A(Type)	.231	1	.231	.499	.485
A(Type) * P(Age)	.338	2	.169	.364	.698
A(Type) * P(Gender)	.593	1	.593	1.277	.267
Error(A(Type))	13.917	30	.464		
A(Gender)	1.815	1	1.815	5.429	**.027**
A(Gender) * P(Age)	8.796E-02	2	4.398E-02	.132	.877
A(Gender) * P(Gender)	9.259E-03	1	9.259E-03	.028	.869
Error(A(Gender))	10.028	30	.334		
Application * A(Type)	2.671	2	1.336	3.260	**.045**
Error(Application * A(Type))	24.583	60	.410		
Application * A(Gender)	.421	2	.211	.439	.647
Error(Application * A(Gender))	28.806	60	.480		
A(Type) * A(Gender)	8.333E-02	1	8.333E-02	.254	.618
Error(A(Gender)*A(Type))	9.861	30	.329		
Between Subject Effects					
P(Age)	1.560	2	.780	.254	.777
P(Gender)	20.454	1	20.454	6.656	.015
Error	92.194	30	3.073		

Table 6.20 ANOVA for Usability Attribute "Politeness"

The mean results for agent type and application are provided in Table 6.21. Post-hoc t-tests showed that this effect was caused by the high perception of politeness from the informally dressed agents in the travel agency application. This agent type was significantly more polite than the formally dressed version in the same application ($p < 0.05$).

Application	Mean Rating Formal Agents	Mean Rating Informal Agents
Cinema	5.78	5.62
Travel Agency	5.64	5.85
Bank	5.65	5.75

Table 6.21 Usability Attribute "Politeness"
Mean Scores by Application and Agent Type

6.6.6.2 Usability Attribute – "Friendliness"

The results for this usability attribute show that all the agents were perceived as being friendly, regardless of their gender, type or the application in which they appeared (Table 6.22).

The assistant was friendly	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	2.171	2	1.086	1.108	.337
Application * P(Age)	.718	4	.179	.183	.946
Application * P(Gender)	1.144	2	.572	.584	.561
Error(Application)	58.778	60	.980		
A(Type)	9.259E-03	1	9.259E-03	.007	.936
A(Type) * P(Age)	.727	2	.363	.259	.774
A(Type) * P(Gender)	.148	1	.148	.105	.748
Error(A(Type))	42.167	30	1.406		
A(Gender)	.454	1	.454	.390	.537
A(Gender) * P(Age)	.199	2	9.954E-02	.085	.918
A(Gender) * P(Gender)	.148	1	.148	.127	.724
Error(A(Gender))	34.944	30	1.165		
Application * A(Type)	1.699	2	.850	1.062	.352
Error(Application * A(Type))	48.000	60	.800		
Application * A(Gender)	3.255	2	1.627	1.569	.217
Error(Application * A(Gender))	62.222	60	1.037		
A(Type) * A(Gender)	1.588	2	.794	1.016	.374
Error(A(Gender)*A(Type))	23.444	30	.781		
Between Within Subject Effects					
P(Age)	4.505	2	2.252	.465	.632
P(Gender)	9.481	1	9.481	1.959	.172
Error	145.222	30	4.841		

Table 6.22 ANOVA for Usability Attribute "Friendliness"

The evaluation presented in Chapter 5 showed that there were significant differences with respect to the friendliness of the 3D agents who appeared inside the 3D room (C6), caused by the poor perception of the male 3D agent because it appeared to be larger and more dominating than the female agent. While designing the 3D agents for the evaluation discussed in this chapter it was mentioned that great care was taken to ensure that both male and female agents were in proportion to the 3D world in which they inhabited. The results from this usability attribute, showing no perceptual differences to

friendliness due to agent type and indicate that the suggestion offered in Chapter 5 was possibly correct and that the improvement was beneficial.

6.6.6.3 Usability Attribute – "Competence"

The assistant was competent	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.782	2	.391	.335	.717
Application * P(Age)	6.606	4	1.652	1.415	.240
Application * P(Gender)	1.699	2	.850	.728	.487
Error(Application)	70.028	60	1.167		
A(Type)	1.565	1	1.565	.799	.378
A(Type) * P(Age)	.977	2	.488	.249	.781
A(Type) * P(Gender)	9.259E-03	1	9.259E-03	.005	.946
Error(A(Type))	58.750	30	1.958		
A(Gender)	.926	1	.926	.461	.502
A(Gender) * P(Age)	.588	2	.294	.147	.864
A(Gender) * P(Gender)	2.370	1	2.370	1.181	.286
Error(A(Gender))	60.194	30	2.006		
Application * A(Type)	6.421	2	3.211	1.645	.202
Error(Application * A(Type))	117.083	60	1.951		
Application * A(Gender)	1.671	2	.836	.601	.552
Error(Application * A(Gender))	83.472	60	1.391		
A(Type) * A(Gender)	.375	2	.187	.192	.827
Error(A(Gender)*A(Type))	29.361	30	.979		
Between Within Subject Effects					
P(Age)	28.949	2	14.475	2.640	.088
P(Gender)	23.148	1	23.148	4.222	**.049**
Error	164.472	30	5.482		

Table 6.23 ANOVA for Usability Attribute "Competence"

This usability attribute showed that the male and female agents were thought to be equally competent in all three applications regardless of agent type (Table 6.23). This result agrees with the findings of the passive viewing evaluation, which included 3D agents (Chapter 5) where they were also perceived to be competent regardless of gender or type.

6.6.6.4 Usability Attribute – "Sociability"

The results from this usability attribute, as seen in the ANOVA table above show that all the agents were perceived as being sociable, in all three applications (Table 6.24).

The assistant was unsociable	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	3.394	2	1.697	1.349	.267
Application * P(Age)	6.218	4	1.554	1.236	.305
Application * P(Gender)	2.681	2	1.340	1.066	.351
Error(Application)	75.444	60	1.257		
A(Type)	.113	1	.113	.080	.779
A(Type) * P(Age)	.907	2	.454	.320	.728
A(Type) * P(Gender)	.521	1	.521	.368	.549
Error(A(Type))	42.486	30	1.416		
A(Gender)	2.083E-02	1	2.083E-02	.028	.868
A(Gender) * P(Age)	1.056	2	.528	.715	.497
A(Gender) * P(Gender)	.280	1	.280	.379	.543
Error(A(Gender))	22.153	30	.738		
Application * A(Type)	1.005	2	.502	.986	.379
Error(Application * A(Type))	30.556	60	.509		
Application * A(Gender)	.431	2	.215	.435	.650
Error(Application * A(Gender))	29.722	60	.495		
A(Type) * A(Gender)	5.787E-02	1	5.787E-02	.061	.807
Error(A(Gender)*A(Type))	28.431	30	.948		
Between Within Subject Effects					
P(Age)	36.074	2	18.037	2.051	.146
P(Gender)	2.083E-02	1	2.083E-02	.002	.961
Error	263.764	30	8.792		

Table 6.24 ANOVA for Usability Attribute "Sociability"

6.6.6.5 Usability Attribute – "Cheerfulness"

From on from the other usability attribute relating to aspects of personality this usability attribute showed that all the agents were equally cheerful in all three applications (Table 6.25). The mean result is just above the neutral marking on the Likert scale, suggesting that the agents could perhaps have been more cheerful during the interaction.

The assistant was cheerful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.852E-02	2	9.259E-03	.011	.989
Application * P(Age)	1.065	4	.266	.322	.862
Application * P(Gender)	1.241	2	.620	.749	.477
Error(Application)	49.667	60	.828		
A(Type)	1.120	1	1.120	.858	.362
A(Type) * P(Age)	2.171	2	1.086	.831	.445
A(Type) * P(Gender)	4.481	1	4.481	3.430	.074
Error(A(Type))	39.194	30	1.306		
A(Gender)	4.083	1	4.083	2.707	.110
A(Gender) * P(Age)	.542	2	.271	.180	.837
A(Gender) * P(Gender)	3.704	1	3.704	2.455	.128
Error(A(Gender))	45.250	30	1.508		
Application * A(Type)	3.352	2	1.676	3.170	**.049**
Error(Application * A(Type))	31.722	60	.529		
Application * A(Gender)	1.500	2	.750	.742	.481
Error(Application * A(Gender))	60.667	60	1.011		
A(Type) * A(Gender)	1.514	2	.757	1.264	.297
Error(A(Gender)*A(Type))	17.972	30	.599		
Between Subject Effects					
P(Age)	62.310	2	31.155	2.641	.088
P(Gender)	22.231	1	22.231	1.884	.180
Error	353.917	30	11.797		

Table 6.25 ANOVA for Usability Attribute "Cheerfulness"

6.6.6.6 Usability Attribute – "Agreeable"

Another important trait for assistants to have is to be agreeable. All the agents in the evaluation regardless of type, gender or application were thought to be equally agreeable (Table 6.26).

The assistant was agreeable	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.032	2	.516	.859	.429
Application * P(Age)	4.162	4	1.041	1.732	.155
Application * P(Gender)	1.005	2	.502	.836	.438
Error(Application)	36.056	60	.601		
A(Type)	1.225	1	1.225	1.186	.285
A(Type) * P(Age)	.977	2	.488	.473	.628
A(Type) * P(Gender)	1.021	1	1.021	.988	.328
Error(A(Type))	30.986	30	1.033		
A(Gender)	1.021	1	1.021	1.971	.171
A(Gender) * P(Age)	1.625	2	.812	1.568	.225
A(Gender) * P(Gender)	5.787E-02	1	5.787E-02	.112	.741
Error(A(Gender))	15.542	30	.518		
Application * A(Type)	.699	2	.350	.574	.566
Error(Application * A(Type))	36.556	60	.609		
Application * A(Gender)	2.264	2	1.132	1.254	.293
Error(Application * A(Gender))	54.167	60	.903		
A(Type) * A(Gender)	.187	1	.187	.340	.564
Error(A(Gender)*A(Type))	16.542	30	.551		
Between Subject Effects					
P(Age)	11.894	2	5.947	1.000	.380
P(Gender)	14.447	1	14.447	2.430	.129
Error	178.319	30	5.944		

Table 6.26 ANOVA for Usability Attribute "Agreeable"

6.6.6.7 Usability Attribute – "Trustworthiness"

This ANOVA table shows significant results with respect to application (Table 6.27). Although just beyond significance there is a marginal interaction between application and agent gender.

The assistant was trustworthy	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	6.838	2	3.419	3.520	.036
Application * P(Age)	2.870	4	.718	.739	.569
Application * P(Gender)	1.838	2	.919	.946	.394
Error(Application)	58.278	60	.971		
A(Type)	.113	1	.113	.115	.737
A(Type) * P(Age)	.310	2	.155	.157	.855
A(Type) * P(Gender)	.836	1	.836	.846	.365
Error(A(Type))	29.625	30	.988		
A(Gender)	1.021	1	1.021	1.509	.229
A(Gender) * P(Age)	1.347	2	.674	.996	.381
A(Gender) * P(Gender)	.113	1	.113	.168	.685
Error(A(Gender))	20.292	30	.676		
Application * A(Type)	4.310	2	2.155	2.821	.051
Error(Application * A(Type))	45.833	60	.764		
Application * A(Gender)	.264	2	.132	.191	.827
Error(Application * A(Gender))	41.500	60	.692		
A(Type) * A(Gender)	5.787E-02	1	5.787E-02	.089	.767
Error(A(Gender)*A(Type))	19.403	30	.647		
Between Subject Effects					
P(Age)	41.866	2	20.933	2.534	.096
P(Gender)	6.502	1	6.502	.787	.382
Error	247.847	30	8.262		

Table 6.27 ANOVA for Usability Attribute "Trustworthiness"

The mean results for application are presented in Table 6.28 and graphically in Figure 6.12. T-tests confirm that the agent in the banking application was perceived as significantly less trustworthy than either of the other two applications ($p < 0.05$).

Application	Mean Rating Score
Cinema	5.15
Travel Agency	5.24
Bank	4.94

Table 6.28 Usability Attribute "Trustworthiness" - Mean Scores by Application

T-tests also show that both the formal and informal agents in the banking application were the least trustworthy and both were significantly less trustworthy than the informal agent in the travel agency ($p < 0.01$).

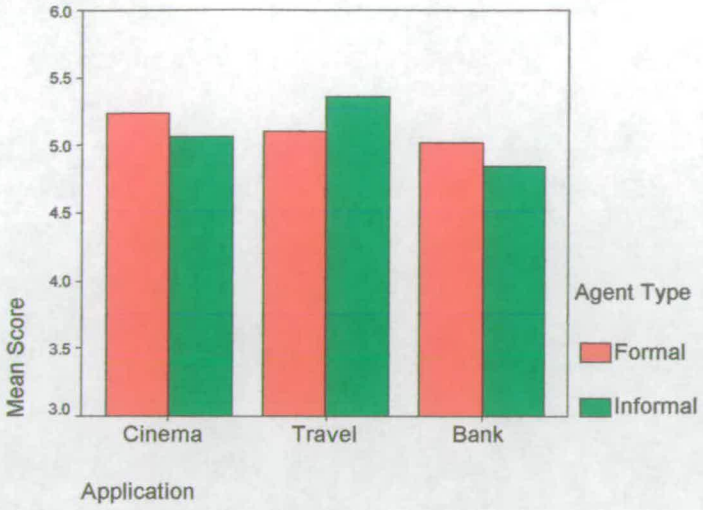


Figure 6.12 Usability Attribute "Trustworthiness" - Mean Scores by Application and Agent Type

The results of this section show that the 3D embodied conversational agents were thought to be friendly, competent, sociable, cheerful and agreeable regardless of their gender or type. However, in the banking applications agent were thought to be less trustworthy.

6.6.7 Attitude to Agents' Appearance

The mean scores for agent gender, agent type and application relating to the participants' attitudes towards the appearance of the agents are presented in Figure 6.13.

6.6.7.1 Usability Attribute – "Liked appearance"

The mean results are presented graphically in Figure 6.14. Post-hoc t-tests show that participants significantly preferred the formal agents to the informal agents in the banking application ($p < 0.01$). A significant interaction effect for agent type and application emerged (Table 6.29).

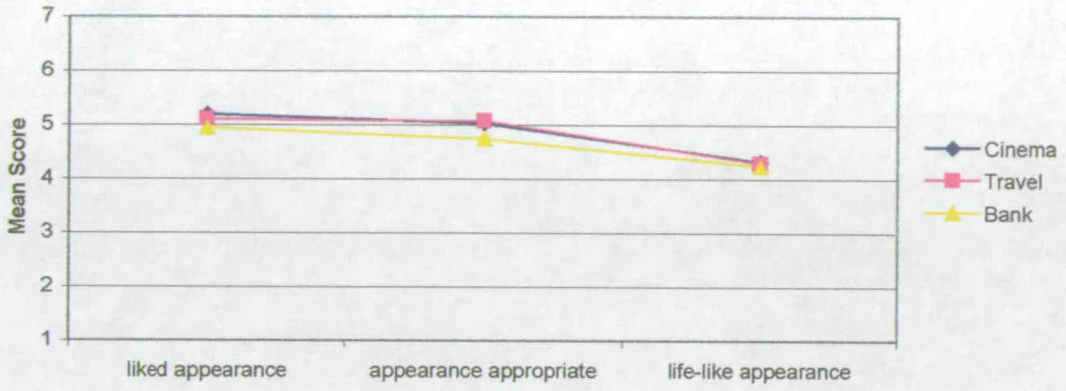


Figure 6.13(i) Usability Attributes for Agents' Appearance by Application

Figure 6.13(ii) Usability Attributes for Agents' Appearance by Agent Gender

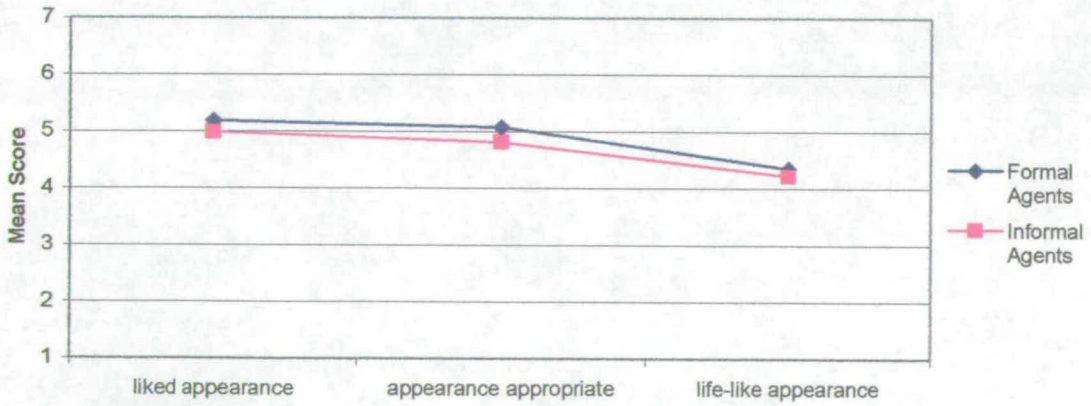


Figure 6.13(iii) Usability Attributes for Agents' Appearance by Agent Type

I did not like the appearance of the assistant	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	4.667	2	2.333	2.074	.135
Application * P(Age)	5.361	4	1.340	1.191	.324
Application * P(Gender)	1.185	2	.593	.527	.593
Error(Application)	67.500	60	1.125		
A(Type)	4.083	1	4.083	1.055	.313
A(Type) * P(Age)	18.722	2	9.361	2.418	.106
A(Type) * P(Gender)	3.000	1	3.000	.775	.386
Error(A(Type))	116.139	30	3.871		
A(Gender)	25.037	1	25.037	19.467	**.000**
A(Gender) * P(Age)	.130	2	6.481E-02	.050	.951
A(Gender) * P(Gender)	1.565	1	1.565	1.217	.279
Error(A(Gender))	38.583	30	1.286		
Application * A(Type)	14.000	2	7.000	8.915	**.000**
Error(Application * A(Type))	47.111	60	.785		
Application * A(Gender)	2.907	2	1.454	1.610	.208
Error(Application * A(Gender))	54.167	60	.903		
A(Type) * A(Gender)	1.056	2	.528	.246	.784
Error(A(Gender)*A(Type))	64.472	30	2.149		
Between Subject Effects					
P(Age)	2.722	2	1.361	.269	.766
P(Gender)	98.231	1	98.231	19.398	**.000**
Error	151.917	30	5.064		

Table 6.29 ANOVA for Usability Attribute "Liked appearance"

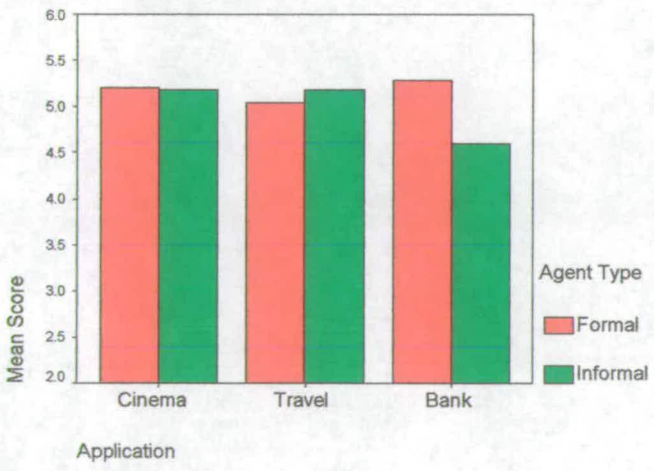


Figure 6.14 Usability Attribute "Liked appearance"
Mean Scores by Application and Agent Type

6.6.7.2 Usability Attribute – "Dressed appropriately"

Participants were asked if the assistants were dressed appropriately for the applications. A significant interaction (Table 6.30) between application and agent type showed that participants felt the agents in the cinema application should be dressed informally and agents should be dressed formally in the banking application.

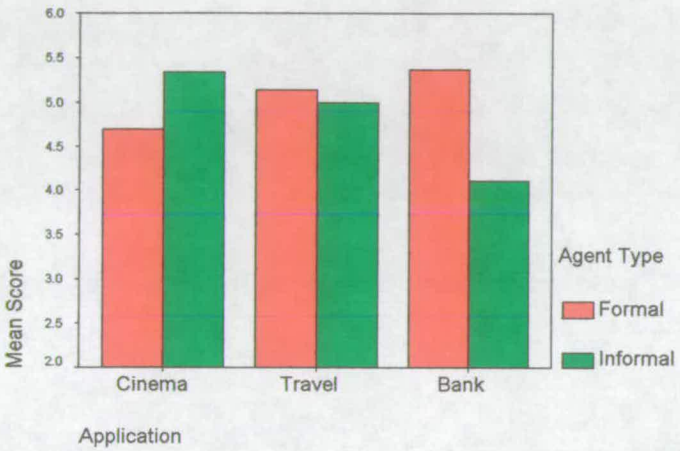


Figure 6.15 Usability Attribute "Dressed appropriately" Mean Scores by Application and Agent Type

In the banking application, male and female formal agents were thought to be more appropriate than their informal counterparts. T-Tests also showed that the male formal agent was also more appropriate in the bank than the female informal agent, and the female formal agent was also more appropriate than the male informal agent (all at $p < 0.01$).

The assistant was not dressed appropriately for the service	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	8.931	2	4.465	1.999	.144
Application * P(Age)	15.694	4	3.924	1.756	.150
Application * P(Gender)	.366	2	.183	.082	.921
Error(Application)	134.028	60	2.234		
A(Type)	6.750	1	6.750	2.063	.161
A(Type) * P(Age)	5.681	2	2.840	.868	.430
A(Type) * P(Gender)	1.333	1	1.333	.408	.528
Error(A(Type))	98.139	30	3.271		
A(Gender)	6.750	1	6.750	2.887	.100
A(Gender) * P(Age)	3.597	2	1.799	.769	.472
A(Gender) * P(Gender)	.333	1	.333	.143	.708
Error(A(Gender))	70.139	30	2.338		
Application * A(Type)	66.792	2	33.396	17.369	**.000**
Error(Application * A(Type))	115.361	60	1.923		
Application * A(Gender)	1.847	2	.924	.755	.474
Error(Application * A(Gender))	73.361	60	1.223		
A(Type) * A(Gender)	5.514	2	2.757	1.945	.161
Error(A(Gender)*A(Type))	42.528	30	1.418		
Between Subject Effects					
P(Age)	4.167E-02	2	2.083E-02	.003	.997
P(Gender)	34.454	1	34.454	5.722	**.023**
Error	180.639	30	6.021		

Table 6.30 ANOVA for Usability Attribute "Dressed appropriately"

6.6.7.3 Usability Attribute – "Lifelikeness"

Finally, participants were asked if they thought the agents appeared lifelike. The male agents appeared equally lifelike, as did the female agents (Table 6.31). The grand mean is just above the neutral mark of 4 on the Likert scale indicating that there is room for improvement with respect to the lifelikeness of the agents.

With respect to the quantitative data focusing specifically on the agents' appearance in the application, it was found the informally dressed agents were more appropriate for informal applications and formally dressed agents were more appropriate for formal applications.

The assistant appeared lifelike	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.519	2	.259	.225	.800
Application * P(Age)	1.051	4	.263	.228	.922
Application * P(Gender)	.130	2	6.481E-02	.056	.945
Error(Application)	69.278	60	1.155		
A(Type)	1.815	1	1.815	1.972	.171
A(Type) * P(Age)	.616	2	.308	.335	.718
A(Type) * P(Gender)	2.083	1	2.083	2.264	.143
Error(A(Type))	27.611	30	.920		
A(Gender)	4.481	1	4.481	6.402	.057
A(Gender) * P(Age)	1.838	2	.919	1.313	.284
A(Gender) * P(Gender)	9.259E-03	1	9.259E-03	.013	.909
Error(A(Gender))	21.000	30	.700		
Application * A(Type)	3.241	2	1.620	3.065	.054
Error(Application * A(Type))	31.722	60	.529		
Application * A(Gender)	.130	2	6.481E-02	.173	.842
Error(Application * A(Gender))	22.500	60	.375		
A(Type) * A(Gender)	5.787	1	5.787	9.645	**.004**
Error(A(Gender)*A(Type))	18.000	30	.600		
Between Subject Effects					
P(Age)	15.477	2	7.738	.410	.668
P(Gender)	45.370	1	45.370	2.402	.132
Error	566.722	30	18.891		

Table 6.31ANOVA for Usability Attribute "Lifelikeness"

6.6.8 Qualitative Comments about the Assistants

Participants were given the option to comment on their experiences. Most of the comments concerned ways to improve certain aspects of the system and the main themes that emerged are presented below.

- *Gesturing:* at times the agents' gesturing seemed to be inappropriate. The shrugging gestures were not perceived as natural human gestures, they seemed "a bit awkward".

- *Tone of voice*: during the interactions some participants were more sensitive to the fact that the speech output was concatenated pre-recorded speech. One participant felt the voices sounded *"a bit robotic"* at times during the

conversation. This was particularly the case for the male voice and the concatenated nature of the male voice did not appear as natural to the participants. Unfortunately this gender difference did not emerge in the pre-evaluation voice test as the participants in that test listened to entire recorded sentences from the male and female voices and not concatenated sentences created from individual recorded prompts. It can then be stated that the poorer response to the male voice was caused by less fluid concatenations.

- *Delays in responses:* due to real-time technological restraints, some of the output responses were slow; some participants found these delays off-putting, annoying and sometimes gave the impression that the assistant was unsure.

- *Eye contact*: some participants felt slightly confused with the assistants' eye-contact and stated that during the confirmation stage of the interaction the assistants should focus their attention on the virtual computer screen in the virtual world and then restore eye-contact with the user. This would give a better impression that the agent was processing the information.

6.6.9 Interview Feedback

All participants in the experiment took part in a closing interview to probe further issues with regard to their experience.

6.6.9.1 Suggested Improvements to the Assistants

With nine participants commenting on the assistants' gestures, this was the main aspect with regard to improving the assistants. Secondly, the speed of processing caused concern. Participants who experienced difficulties during the interaction stated that the recognition should be improved.

6.6.9.2 Enjoyment Factor

Two thirds of the participants (24 of 36) thought the assistants enhanced the services and also thought it was a good idea to talk to this type of assistant and that there is a need for such assistants in computing applications. Three participants stated that the assistants particularly enhanced the cinema application. Participants encouragingly enjoyed

speaking to the assistants and interestingly one participant is quoted as saying: "Yes, I enjoyed talking to the assistants, I was even polite to them".

6.6.9.3 Most Important Characteristics

The assistants should be polite and cheerful and should demonstrate competence during the interaction. They should smile and provide appropriate verbal feedback. It is also important for the assistants to have friendly voices.

6.6.9.4 Most Annoying Characteristics

The exaggerated gesturing, especially the shrugging gesture, seemed to be the most annoying and inappropriate feature of the system. Also included in this list are the slow responses of the system. Three participants felt the exiting comment from the assistants "Have a nice day", at the end of the interactions was the most annoying feature.

6.6.9.5 General Comments

Thirty-four participants said they would use the services themselves, with six participants expressing a greater interest in the cinema application. Participants were asked to suggest other applications that would benefit from having an interactive assistant. The list included on-line shopping scenarios, including supermarkets, bookstores and pizza delivery. Also mentioned were educational services whereby the assistants would be effective learning devices for children. Travel information was also mentioned, including train and flight information services.

6.7 Discussion

The first prediction, which stated that the participants would respond positively to the deployment of embodied agents in the applications, was supported. The results support the claim that 3D embodied conversational agents have a role to play as assistants in interactive VRML retail application environments. This can be said given the positive responses recorded in the quantitative questionnaires, the qualitative comments and also the post experiment interviews. Usability attributes showed that all the applications were easy to use and convenient. Participants stated they would use the applications

themselves and that all were good ideas. When asked to make a selection between the applications the majority of participants did select the cinema application in comparison to the other applications. Although participants were positive toward all three applications, they seemed to have more confidence in the cinema applications and without improvements in real-time interaction they would be more concerned about the deployment of the agents in more serious domains such as banking.

Despite the fact that the participants enjoyed speaking to the agents in all three applications, supporting the second prediction, as mentioned previously the cinema application was more popular in comparison to the other applications. Participants felt it was more entertaining than the travel agency and banking application. In addition, results show that the participants preferred to speak to the female agents. It was discovered that despite pre-evaluation tests on the voices of the agents, the participants did not rate the male voice was highly as the female due to the unnaturalness of the concatenated nature of the voice. This resulted in the participants claiming that they would prefer to speak to the female agents.

The third claim (Prediction 6.3) was also supported when it was discovered that formally dressed agents were preferred in the banking application and informally dressed agents were preferred in the cinema application, showing that the stereotypes created are suited to different environments and should the agents appearance not correspond to the participant's expectations, the participant's behaviour can change, as predicted by Churchill et al. (2000), and manifest itself in the form of a negative attitude toward the agent, or expression of unwillingness to re-use the service.

Previous evaluations (Chapter 4) indicated significant differences with respect to the voices of various agents, despite the fact that the same, recorded voice was used for all agents of the same gender. It was again important to test for such cross-modal effects. 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 this experiment, participants liked both the male and female voices for both types of agent. However, issues did arise due to the concatenated nature of the output utterances from the agents, in particular the male agents and 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 intonational recordings to ensure that the transitions between recorded prompts in every utterance sound correct.

All four agents were perceived as having similar personalities, with respect to politeness, friendliness, competence, cheerfulness, sociability and agreeability; all traits important for assistants in retail spaces. Differences did emerge with respect to the trustworthiness of the agents between the applications. This did not then support the fifth prediction. The qualitative results showed that participants were less likely to trust the agents to complete tasks correctly particularly in the banking application. During the interviews, participants stated that if they could be convinced that the assistant would understand every input utterance correctly, their confidence in the system would probably increase and they would be more likely to use the application.

Establishing trust between the agent and the user is of importance, and on-going research (Bickmore & Cassell, 2000) is exploring the construction of a social relationship to assist with establishing this trust. In particular the use of 'small-talk' is being investigated as an effective mechanism for establishing the social interaction. Other approaches investigating the trustworthiness of (van Mulken, André & Müller, 1999) have shown that system competence has a direct relationship with the trust of agents. This is in agreement with the result reported in this chapter, that unless users are confident that the system can understand and process information correctly they are less likely to trust the agent. However, in the study by Van Mulken et al (1999) the results also showed that the personification of interfaces does not appear to be sufficient for raising trustworthiness, raising the question of whether there are other methods for establishing trust.

Burgoon et al. (2000) investigated the credibility, understanding and influence of a variety of interfaces, including interfaces ranging from text-only output to face-to-face communication, and showed that the nature of the interface had no effect on user perception of credibility, understanding or influence. This conflicts with information gathered in the experiment reported here, where participants stated that the use of text output in the interface could help with feedback, thus raising user confidence in the system and therein, according to the results by Van Mulken et al. (1999), have an effect on users' perception of agents' trust. The issue of raising trust in the interface, in particular with respect to more serious retail applications such as banking is explored in the final evaluation of this thesis, which is presented in the next chapter. For if a social

interaction is to be established in any application some level of trust needs to be reached to encourage a successful first interaction and to prompt further interactions.

6.8 Summary

This chapter has reported the design, development and implementation of an interactive experiment platform where 3D ECA could be evaluated in contrasting VRML retail applications. Overall results from the participant sample that were invited to converse with the agents were encouraging with respect to the deployment of embodied agents in retail environments and with improvements in real-time technology the interaction could improve further. It was discovered that real-life stereotypes do transfer to virtual retail environments and casually dressed agents are more suitable in informal, entertaining domains and formally dressed agents are more suitable in more serious application domains. Finally as mentioned it was shown that the trustworthiness of agents can vary in contrasting domains and participants are less likely to trust agents in more serious applications, such as banking, at least until they have more confidence in the capabilities of the recognition system. The development of trustworthy interfaces is the research topic of the next chapter, which details the final experiment in the evolution of empirical research detailed in this thesis. Using the interactive experiment interface designed and presented in this chapter, a multi-modal approach to interacting with agents is explored, in particular the use of text input and output combined with speech input and output is investigated as methods to promote the trustworthiness of both the agents and the interfaces in which they appear.

Chapter 7

Adding Multi-Modal Features to Interactive Retail Interfaces as Means to Improving Agent Trustworthiness

7.1 Introduction

The Cheskin Report (Ecommerce Trust Study, 1999) is an excellent example of recent research which demonstrates, through the study of user perceptions, the importance of creating trustworthy e-commence applications. Specifically, the report highlights the finding that it is essential to maintain a feeling of security and trust in the interface where on-line transactions are involved. Meech and Marsh (2000) reinforce these conclusions but go further when they point out that as on-line buying and selling tasks become increasingly more complex, the use of interpersonal and social communications skills can help, and it is argued in this thesis that use of embodied conversational agents can thereby help to promote this type of communication and thus infer a greater sense of confidence and trust in eCommerce environments generally.

The results presented in Chapter 6 are in general agreement with Meech and Marsh's conclusions. They demonstrate that not all applications require the same level of confidence and trust from the user for successful interactions to take place: some require the user to have more confidence and trust than others. This is especially true of eCommerce applications where users are asked to disclose financial information and the agents that appear in these banking applications are thought to be less trustworthy. These results also agree with the findings of Bickmore and Cassell (2000) who demonstrated that users are initially reluctant to disclose personal and financial information to an agent but that by establishing a social relationship with a conversational agent users can be brought to feel more engaged in a trusting relationship which then leads to a more successful interaction.

Sproull et al. (1996) found that users attributed social responses to agents represented by faces. Importantly for this chapter, they also found that users themselves showed social responses such as trust and co-operation. Van Mulken et al. (1999) found, however, that

anthropomorphising an interface with lifelike agents was not sufficient in itself to *maintain* trust and confidence in the interface. This is an important finding since it is this maintenance of a consistent level of trust that encourages further interactions and provides users with an environment in which they can confidently disclose information. In further support of the need for the type of investigation presented in this chapter, Meech and Marsh (2000), like Van Mulken et al. (1999), argue that it is no longer sufficient to create "the illusion of life" by simply inhabiting spaces with ECA. It is the purpose of this chapter to evaluate multi-modal interface features, which are designed to improve user confidence so that trustworthy interactions between agents and users may take place in retail environments where eCommerce transactions occur.

Alternative research strategies for investigating trust in more risky environments have been pursued by Kim and Moon (1997) who investigated designing user interfaces for eCommerce that would evoke target feelings in the customer such as trustworthiness. Results showed that to arouse trustworthiness in the customer in a cyber-banking context the system design should be three-dimensional, covering half the screen. It should use cool, pastel colours, which would elicit a greater feeling of trustworthiness. The significance of the user being able to manipulate the interface became evident in Chapter 6 when participants suggested that providing the user with text input might improve user confidence in the functionality of the system, thus promoting a more trustworthy interaction. Consequently, the research reported in this chapter introduces mechanisms that may arouse, maintain and improve user's confidence and thus improve the perceived trustworthiness of the agents in these applications.

The experiment described here introduced text output to VRML speech-based retail interfaces and aimed to investigate the effects of this text output to examine if it assists the user throughout the conversation with an agent by providing extra feedback. In addition to this the user was given the option of entering details with the keyboard instead of via speech. It was considered that this could improve user confidence in the system from two perspectives. Firstly, from a security and privacy perspective, by not vocalising the information users may have more confidence in the system. Secondly, in the event of users not being confident with the system's speech recognition capabilities, text entry may be preferred. Two of the applications (cinema and bank) were selected from the experiment platform described in detail in Chapter 6 for purposes of the evaluation in this chapter. In Chapter 6 it was shown that significant differences

emerged between these two applications, where the cinema agent was perceived as being significantly more trustworthy than the bank agent. These applications provide a medium in which improvements in the level of trust between the applications can be carefully examined.

In order to design an experiment focusing primarily on perceived trustworthiness it is important to understand its meaning, however, the elusive concept of trust can make it difficult to analyse associated attributes such as reliability, dependability, security, confidence and credibility. As Corritore et al. (Corritore, Kracher & Wiedenbeck, 2000) discuss, trust is a multi-dimensional topic of research in management, psychology, philosophy, and sociology and more recently it has become a salient feature within the interdisciplinary field of HCI. In the field of sociology it is thought to be a social lubricant inherent to communication. Within philosophy, trust is thought to be an essential element of societal operations. Psychologists focus on individual traits and how such personal traits manifest into trusting behaviour. Within management, trust is being researched primarily to enable successful interaction within risky or uncertain situations. In marketing, both online and offline, trust is essential for successful business transactions. Although between the various disciplines there is little agreement on the actual concept of trust, the following list contains many of the points that are generally agreed upon and can be transferred to the field of human computer interaction and in the context of the interdisciplinary investigation of trust in this chapter, these points help to explain the difficulty in firstly building an application inhabited with trustworthy agents and secondly analysing trustworthy interactions.

1. Trust is difficult to establish and easily damaged.

2. Trust is rarely established instantaneously.

3. Trust can be established through familiarity.

4. Trust can be established through displays of competence.

5. Trust is dependent on situational contexts.

With these points in mind, the complexities of understanding and actually establishing trust in interfaces are outlined which indicates a clearer perspective and definition of the

concept and meaning of trust in human computer interaction in order to evaluate it in interfaces.

In an attempt to define the meaning of trust, McKnight (1996) explains the multidimensional facets of trust and concludes, like Corritore et al. (2000), that a cross-disciplinary conceptualisation may in fact be beneficial, initiating more structured explorations into the establishment of trust between agents and users and also enabling "a richer, more well-balanced view of the phenomenon". McKnight's framework has been used in this experiment as a reference model to understand trust within a human-computer interaction situation. Having a richer model of the complex nature of trust allows a more constructive approach to be taken when analysing users' cognitive and behavioural characteristics during interactions, which can be measured through Likert scales and post-experiment interviews.

The six constructs that McKnight defines in order to conceptualise trust are trusting intention; trusting behaviour; trusting beliefs; system trust; dispositional trust; and situational decision to trust. The framework that links these constructs is illustrated in Figure 7.1. Each of the constructs supports one or more of the other constructs in various ways and each has a role to play during the analysis of trust within the experiment described in this chapter. In essence this model states that users' beliefs and expectations lead to the formation of intentions, which are manifested in displayed behaviour. In addition, the level of trust or confidence that the user has in the system can also depend on the context of the system (situational and system trust) and the subjective nature of the user (dispositional trust), which can in turn influence their intentions and thus their behaviour.

Participants in the experiment were invited to converse with embodied conversational agents in two contrasting application environments. At the point when participants are asked to converse with the agent, participants begin to form the belief that the agent can understand their spoken input, which leads them to have the intention of conversing with the agent. The resulting manifested behaviour is displayed when the user actually speaks to the agent. For this interaction to develop into a trusting association the users' beliefs, intentions and behaviour must be manifested as trusting beliefs, trusting intentions and trusting behaviours.

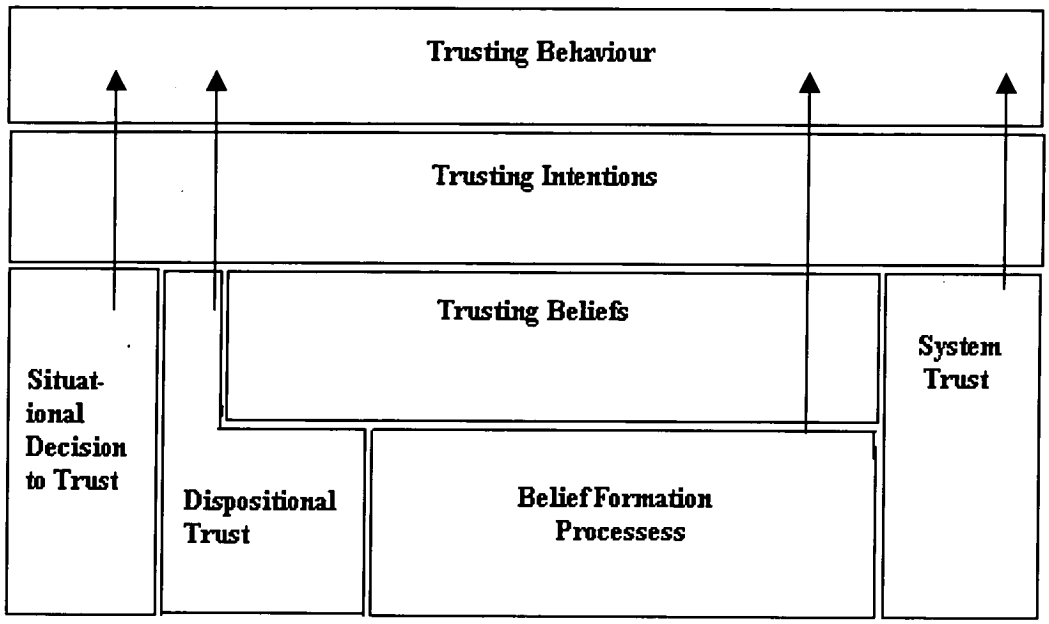


Figure 7.1 Multi-disciplinary framework for trust (McKnight & Chervany, 1996)

A trusting belief can be described as the users' confidence in their intentions, which is actually their willingness to depend and rely on another in a given situation, that is their displayed willingness to depend on the agent to assist them with the task. As mentioned, trusting intentions and behaviour also depend on the level of system trust, and the safeguards in place to reduce any uncertainty with respect to the system. These safeguards include building systems that are responsive to known user expectations, such as inhabiting the environments with ECA that display the appropriate verbal and non-verbal communicative behaviour. In addition to this, system trust can also depend on the functionality of the interface where users are provided with the opportunity to engage in the most comfortable and efficient interaction possible. To be both comfortable and efficient, the user must feel content using the application, they must feel in control and they also must feel that the application serves a purpose. Table 7.1 details each of the six constructs as defined by McKnight and their possible manifestation with respect to the experiment described in this chapter.

In the event of participants in the experiment not assigning a degree of trust to the agents or to the application the McKnight guide can effectively be used to analyse the particular areas of the application that need to be developed further in order to promote a trusting environment in which a user can interact. To summarise, in order to investigate user

perceptions of trust in retail applications an evaluation was designed where interface features, such as providing the user with text input and text output, were introduced. Participants interacted with agents who were dressed casually in the cinema application and formally in the banking application (see Chapter 6 results) and were asked to complete questionnaires after they experienced the various interface conditions in both applications. The experiment predictions, procedure and results are presented, followed by a discussion of the implications of the research findings.

Trust Construct	Description of Construct	Experiment Manifestations
Trusting Intention	Willingness to depend on another in a situation with a feeling of security, despite possibility of negative consequences.	Willingness to depend on an ECA to complete task despite possibilities of speech mis-recognition.
Trusting Behaviour	Actual displays of dependency on another party.	Actually conversing with ECA to complete tasks.
Trusting Beliefs	Confidence in intentions i.e. the extent to which another person or thing is trustworthy in a situation.	Established when there is knowledge that the task will successfully be completed, comfortably and efficiently.
System Trust	Establishing structures to enable a successful endeavour. Trust can be established through dependence due to safeguards or due to the system's capability to reduce uncertainty.	Designing and creating the expected behaviour for the ECA can establish system trust. Also, developing a robust recognition system to prevent mis-recognitions occurring.
Dispositional Trust	Personal individual differences lead to different levels of trusting intention, primarily due to development in expectations through life-experiences.	Gender and age balanced experiment design can detect group difference. One-to-one interview can probe further individual differences.
Situational Decision to Trust	Trusting without regard to the specific persons/objects involved because benefits of trusting outweigh possible negative outcomes of trusting.	

Table 7.1 Manifestations of trust with respect to this experiment

7.2 Experiment Conditions

As mentioned text input and text output facilities were added to the interface. To evaluate these different features four experiment conditions were created, as described in Table 7.2.

Experiment Condition	User Input Modes	Agent Output Modes
Cond(0)	Speech	Speech
Cond(1)	Speech and Text	Speech
Cond(2)	Speech	Speech and Text
Cond(3)	Speech and Text	Speech and Text

Table 7.2 Experiment Condition Descriptions

The experiment platform was identical to that described in Chapter 6, where users could speak to the agents and their speech output was captured and processed using a Nuance™ speech recogniser. Chapter 6 provides a more in-depth description of the spoken language capabilities of the interactive system and the construction of the dialogue manager, which controls the direction of the dialogue while the user completes an application task. Additional interface features were added to the applications in order to evaluate user perceptions to various experiment conditions, which may impact attitudes to trustworthiness and associated attributes. Figure 7.3 illustrates the cinema application with and without the text input and text output features. In should be noted that each of these interface features were expertly evaluated before being included in the experiment design (see Chapter 3).

7.2.1 Text Input (Cond(1) & Cond(3))

Participants were informed that when completing application tasks with the agent they would sometimes have the option to type the required task information, instead of entering the details via speech. To capture the text input a command line appeared in the interface, as is visible in Figure 7.3. A text recognition function, activated by the dialogue manager, recognised keywords relevant to each task. This text function permitted the participants to type entire sentences or just the keywords. It also was capable of recognising numbers in the form of digits.

In addition to entering text via the keyboard, participants also had the opportunity to enter security number details via a virtual number pad, which appeared on the desk in the retail environment. Using the mouse to click on the displayed digits, the text recognition function in the dialogue manager was activated and the entered details were processed. The keyboard could also be used.

7.2.2 Text Output (Cond(2) & Cond(3))

In the virtual world, a computer monitor appeared on the assistant's desk and this was normally facing the agent. For the text output condition, this monitor was rotated so that the user could see the screen and the details of the information they had entered. As the conversation progressed more of the data input appeared on the screen. Identical cinema and banking tasks to those used in Chapter 6 were used in this experiment. Table 7.3 presents example tasks for each application. The monitor screens that appeared in the interfaces of both the cinema and banking applications are provided in Figure 7.2.

Application	Example Task
Cinema	Book 2 tickets for the movie Casablanca on Friday at 3 pm
Bank	Transfer £50 from savings account to current account on December 9th.

Table 7.3 Examples of the Experiment Tasks per Application

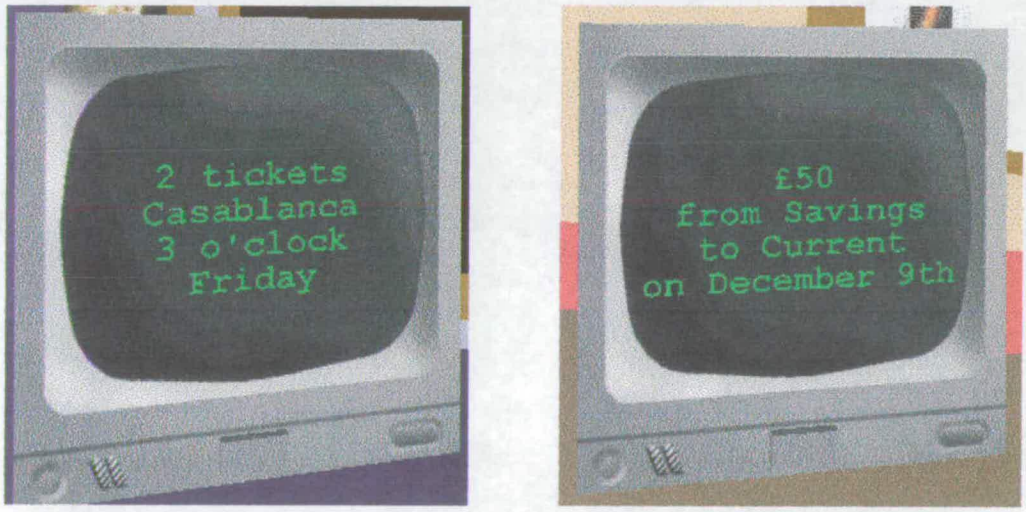


Figure 7.2 Agent's Computer Monitor in Both Applications (Cinema and Bank)

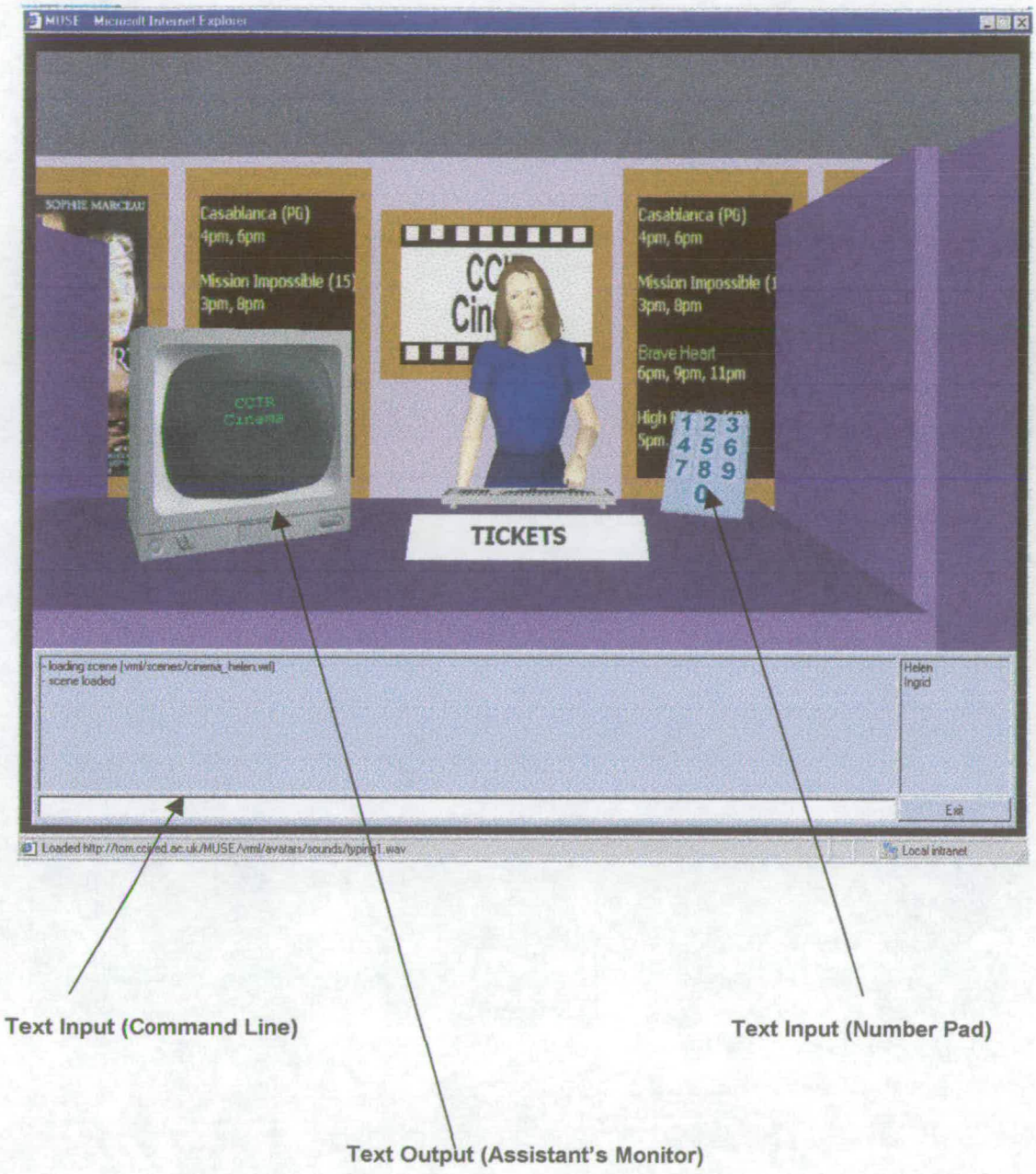


Figure 7.3 Experiment Conditions Illustrated in the Cinema Application

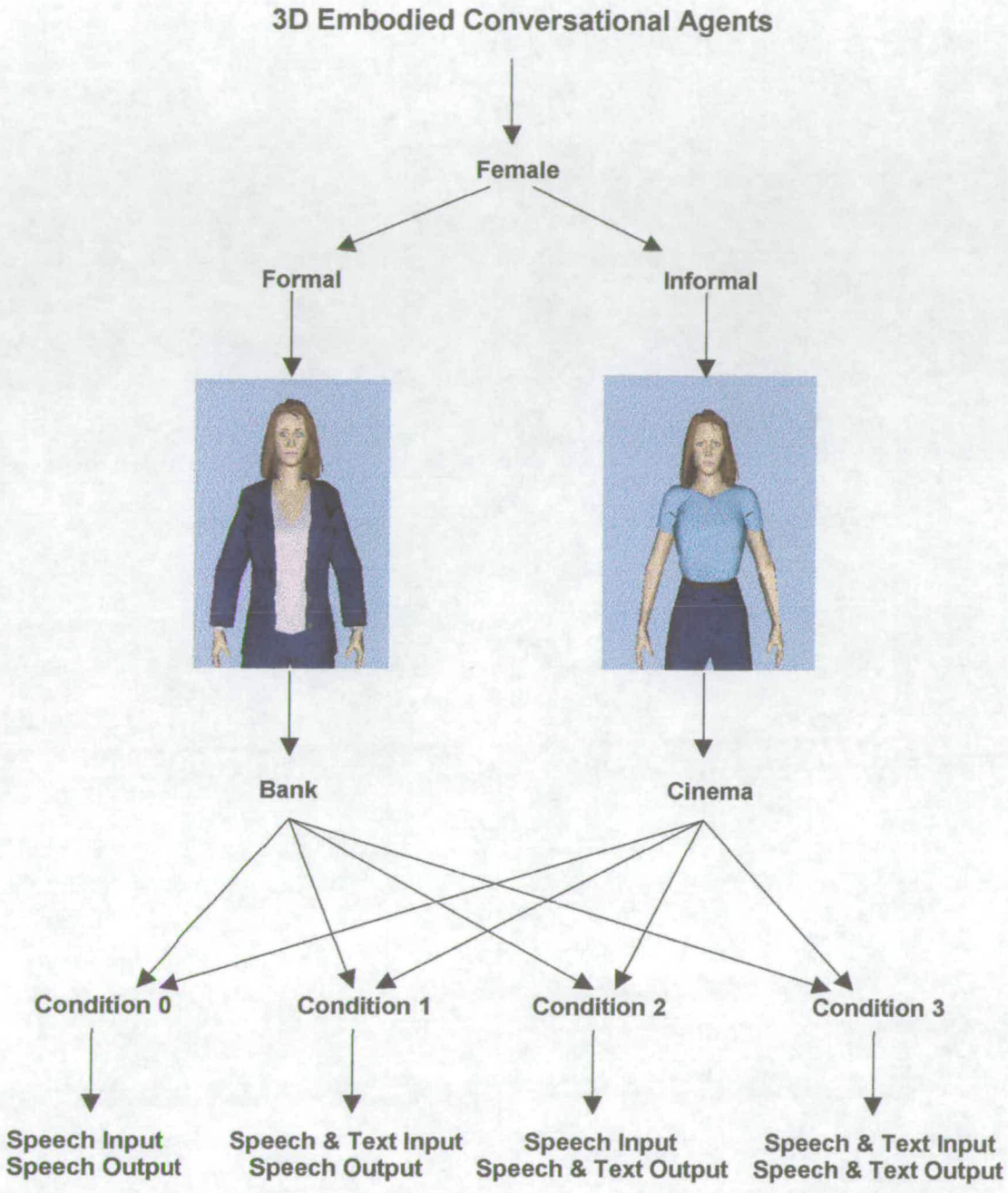


Figure 7.4 Illustration of Interactive Experiment II Implementation

7.3 Experiment Predictions

1. In applications where users must disclose financial information, providing users with the opportunity to type information in addition to speaking will give them greater flexibility than speech input alone and will increase user confidence in the system and user trust in the agent.

2. In applications where users must disclose financial information, providing users with speech and text output feedback in the interface will improve the user's confidence in the system in comparison to a speech output only system and this in turn will increase the user trust in the agent.

3. In applications where users must disclose financial information, combining multi-modal input (speech and text) with multi-modal output (speech and text) will increase user confidence in the system and user trust in the agent in comparison to speech input and output alone.

7.4 Experiment Design

As mentioned the experiment described in this chapter was performed using the experiment platform described in Chapter 6. Two application environments were selected for the evaluation. The virtual cinema box-office and the virtual bank were chosen as significant differences emerged for perceived trustworthiness with respect to these two applications. As agent gender issues were previously investigated in Chapter 6, one agent gender was included in this experiment, the female agent. Results from the previous evaluation also dictated the appearance of the ECA in the applications and so an informally dressed ECA appeared as the assistant in the cinema application and a formally dressed ECA appeared in the banking application. The experiment aimed to assess the effect of introducing text input and text output in combination with speech input and speech output as mechanisms to possibly increase user confidence in the systems and thus infer a greater degree of trustworthiness to the retail applications, especially those where users must disclose financial information. The effect of these interfaces with respect to the perception of the embodied agents was also assessed.

A total of 48 participants (distributed according to gender and age as shown in Table 7.4) took part in the experiment. As stated in Chapter 3, cognitive walkthrough evaluation was completed for the experiment design. From this it was felt that participants in the sample must have computer experience as they were told they would have the opportunity to use the mouse or keyboard at points during the interactions.

	Interactive Evaluation I Evaluation		Total
Participant	**Male**	**Female**	
Age 18-35	8	8	16
Age 36-49	8	8	16
Age 50+	8	8	16
	24	24	48

Table 7.4 Analysis of Participants by Gender and Age Group

The experiment procedure required participants first of all to read an information sheet regarding the application they were about to use (Appendix 3.1). For instance, if the participant was going to use 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 observe the assistant and the application carefully. They were also told that they might be asked for security number information, which was presented to them before the interaction began. Directly before experiencing each experiment condition participants were reminded about the features of the interface, e.g. that they had the option to use the keyboard and number pad or that they would see the assistant's computer screen. In this repeated measures design the presentation of the experiment conditions to the participants was randomised across applications and the presentation of the two applications was balanced amongst the participants. Four similar tasks (identical to those used in Chapter 6) were created for each application, which were randomised amongst the experiment conditions. After the participants experienced each experiment condition they were asked to fill out a questionnaire (7-point Likert attitude questionnaire statements) relating to the agent and the application. The questionnaire statements are shown in Table 7.5. Within the questionnaire, statements were balanced for polarity (equal number of positively and negatively worded stimulus statements). The actual questionnaires used for the experiment can be examined in Appendix 4.1.

As with the previous evaluations described in this thesis, the blueprint method described in Chapter 3 was used to create the questionnaire statements. The content areas were

defined as (1) agent and (2) application. The manifestations were further defined using McKnight as a reference (McKnight & Chervany, 1996) who defines attributes of trust to include the following: reliability, confidence, dependability, competence, goodness and shared understanding. The resulting matrix assisted in producing statements that would provide a true reflection of the concept of trust in the interface. Questionnaire statements 1-13 were answered for all experiment conditions.

Questionnaire Statements	
Applications	1. The service was not reliable.
	2. This service is not a good idea
	3. The service was efficient
	4. I had no confidence in the service
	5. I did not think this service was useful
	6. I felt in control using this service
Agents	7. I felt the assistant was trustworthy.
	8. I did not enjoy speaking to the assistant.
	9. I could depend on the assistant to do the job.
	10. I did not think the assistant was reliable.
	11. I felt confident the assistant understood me.
	12. The assistant was credible.
	13. The assistant was competent.

Table 7.5 Questionnaire Statements

The dependent variables in the experiment were the responses to the individual usability attribute statements in the questionnaire and the responses given during a post experiment interview. The independent variables were the experiment conditions and the VRML retail applications (cinema, bank). These were treated as within-subject variables in a repeated measure design. When participants had experienced all four conditions in both applications they took part in an interview designed to elicit further information about the agents, which also gave participants the opportunity to make suggestions for improvements to the system.

Title	Interactive Evaluation II: Trusting 3D ECA in VRML Retail Applications	
Design		One Independent Sample
Predictions	7.1	Text and speech input would improve user confidence over speech input alone.
	7.2	Text and speech output would improve user confidence over speech output alone.
	7.3	Text and speech input and text and speech output would improve user confidence over speech input and speech output alone.
Dependent Variables		Attitude Questionnaire Responses (1-7 Likert scale)
Other Data		Interview Answers
(Experiment) Independent Variables:	1	Interface Condition (4 levels)
	2	Application (2 levels)
(Participant) Independent Variables	1	Gender (2 levels)
	2	Age Group (3 levels)
Extraneous Variables:	Order	Interface conditions randomised within applications
		Tasks randomised within applications.
		Application presentation balanced between participants.
Location		Edinburgh - CCIR Premises, Central Edinburgh
Cohort		N = 48 50% male, 50% female
Remuneration		£10
Duration:		50 Minutes

Table 7.6 Summary Table of Interactive Multi-Modal Evaluation

7.5 Results

A series of 2 x 4 repeated measures ANOVA were completed for each usability attribute listed in the questionnaires. The results are divided into two sections: attitude to the applications and attitude to the agents. At the beginning of each of these sections, graphs depicting the mean scores with respect to the independent variables of experiment condition and application are given. This is followed by a discussion of the main findings with the ANOVA tables.

7.5.1 Attitude to Applications

Figure 7.5 illustrates the mean scores for usability attributes relating to participants' attitudes toward the applications, specifically focusing on attributes that are associated with perceived trustworthiness. Figure 7.5 shows that for most of the usability attributes in this section, the mean scores for experiment conditions favoured interfaces that included text output (Cond(2) and Cond(3)). The details of the individual attributes are presented next.

Figure 7.5(i) Usability Attributes for Application by Application

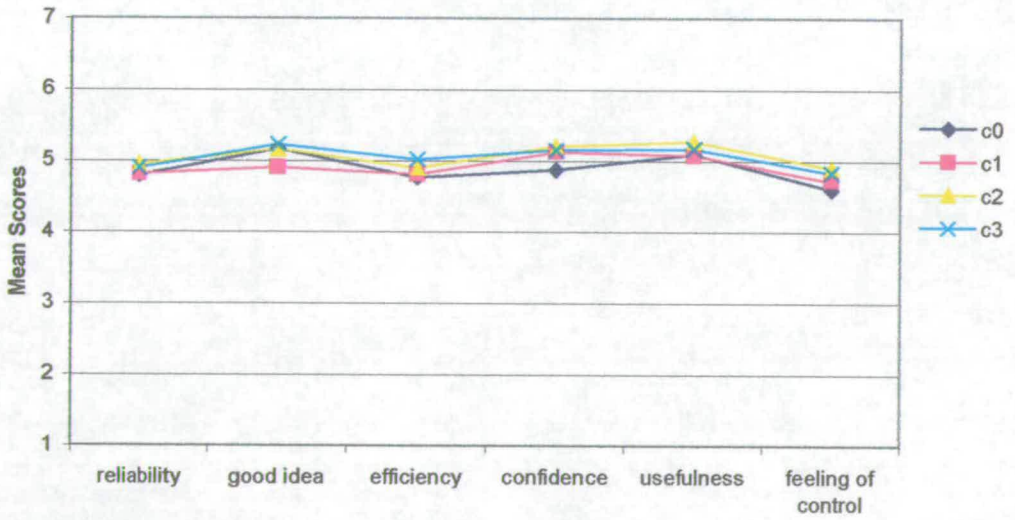


Figure 7.5(ii) Usability Attributes for Application by Experiment Condition

7.5.1.1 Usability Attribute – "Reliability"

The service was reliable	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	12.811	1	12.811	6.965	**.012**
Application * P(Age)	13.277	2	6.639	3.609	**.036**
Application * P(Gender)	1	4.594	2.498	.122	1
Error(Application)	77.252	42	1.839		
Condition	1.552	3	.517	.432	.731
Condition * P(Age)	4.562	6	.760	.634	.703
Condition * P(Gender)	15.116	3	5.039	4.203	**.007**
Error(Condition)	151.047	126	1.199		
Application * Condition	2.175	3	.725	.477	.699
Error(Application * Condition)	191.428	126	1.519		
Between Subject Effects					
P(Age)	40.291	2	20.145	2.814	.071
P(Gender)	27.702	1	27.702	3.869	.056
Error	300.685	42	7.159		

Table 7.7 ANOVA for Usability Attribute "Reliability"

Significant differences emerged (Table 7.7) due to users' attitude toward the reliability of the applications with the cinema application thought to be more reliable than the banking applications (mean cinema = 5.05, mean bank = 4.68). A marginally significant interaction between participant age and application also emerged and a post hoc t-test showed participant in the first age group (18-35) felt that the cinema application was significantly more reliable than the bank (Table 7.8).

Age Group	Mean Rating Score Application	
	Cinema	Bank
Age 18-35	4.95	4.06
Age 36-49	4.83	4.79
Age 50+	5.37	5.19

Table 7.8 Usability Attribute "Reliability" - Mean Score by Participant Age and Application

Although there was no overall effect for experiment condition, there was a highly significant interaction between participant gender and experiment condition. With respect to experiment condition, attitudes from the female participants were significantly

more positive than the male participants with respect to reliability when text output appeared in the interface (Cond(2), Cond(3), see Figure 7.6).

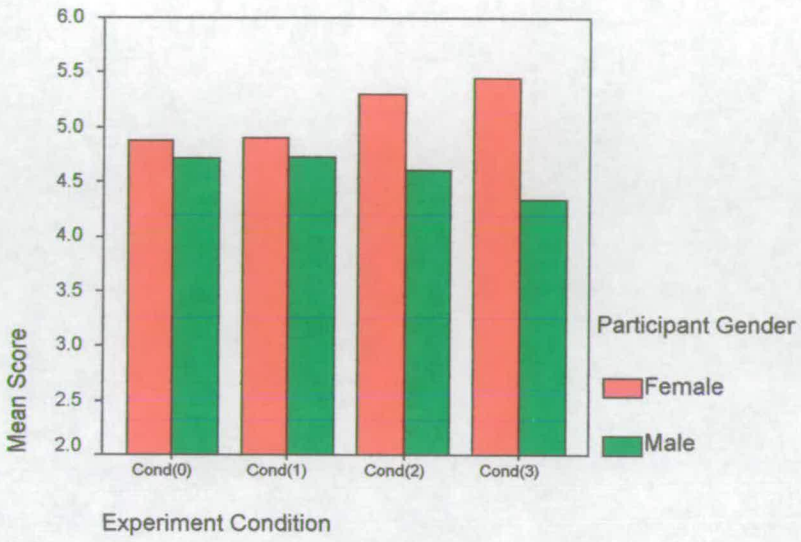


Figure 7.6 Usability Attribute "Reliability" - Mean Score by Participant Gender and Experiment Condition

7.5.1.2 Usability Attribute – "Good idea"

As suggested by Figure 7.5, where the mean scores were graphically illustrated, no significant differences emerged with respect to this usability attribute and in agreement with the results from Chapter 6, participants felt that both applications were good ideas (Table 7.9).

However with respect to experiment condition significant differences did emerge. T-tests showed further that experiment condition, Cond(3), with text input and text output was thought to be a significantly better idea than condition Cond(1) with text input alone, $p < 0.01$ (Table 7.10).

This service is not a good idea	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	8.508E-02	1	8.508E-02	.060	.807
Application * P(Age)	4.384	2	2.192	1.552	.224
Application * P(Gender)	1.128	1	1.128	.799	.377
Error(Application)	59.322	42	1.412		
Condition	5.740	3	1.913	2.384	**.042**
Condition * P(Age)	5.393	6	.899	1.120	.354
Condition * P(Gender)	2.446	3	.815	1.016	.388
Error(Condition)	101.122	126	.803		
Application * Condition	2.657	3	.886	.986	.402
Error(Application * Condition)	113.209	126	.898		
Between Subject Effects					
P(Age)	45.824	2	22.912	1.232	.302
P(Gender)	7.773E-02	1	7.773E-02	.004	.949
Error	781.191	42	18.600		

Table 7.9 ANOVA for Usability Attribute "Good idea"

Application	Mean Rating Score (max 7)
Cond(0)	5.16
Cond(1)	4.90
Cond(2)	5.16
Cond(3)	5.23

Table 7.10 Usability Attribute "Good idea" - Mean Scores for Experiment Condition

7.5.1.3 Usability Attribute – "Efficiency"

Participants felt both the cinema and banking applications were equally efficient. The ANOVA table for this attribute shows that no main effects emerged for the independent variables of application or experiment condition (Table 7.11).

This service was efficient	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.473	1	1.473	.883	.353
Application * P(Age)	1.554	2	.777	.466	.631
Application * P(Gender)	9.277E-02	1	9.277E-02	.056	.815
Error(Application)	70.045	42	1.668		
Condition	3.572	3	1.191	1.091	.355
Condition * P(Age)	8.639	6	1.440	1.320	.253
Condition * P(Gender)	5.093	3	1.698	1.556	.203
Error(Condition)	137.489	126	1.091		
Application * Condition	2.518	3	.839	.692	.558
Error(Application * Condition)	152.739	126	1.212		
Between Subject Effects					
P(Age)	54.404	2	27.202	2.577	.088
P(Gender)	16.953	1	16.953	1.606	.212
Error	443.422	42	10.558		

Table 7.11 ANOVA for Usability Attribute "Efficiency"

7.5.1.4 Usability Attribute – "Confidence"

I had no confidence in this service	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	1.675	1	1.675	1.030	.316
Application * P(Age)	2.517	2	1.259	.774	.468
Application * P(Gender)	1.044	1	1.044	.642	.427
Error(Application)	68.276	42	1.626		
Condition	6.151	3	2.050	1.924	.129
Condition * P(Age)	6.086	1.000	6.086	1.904	.175
Condition * P(Gender)	6.086	3	2.029	1.904	.132
Error(Condition)	134.263	126	1.066		
Application * Condition	2.696	3	.899	.818	.486
Error(Application * Condition)	138.384	126	1.098		
Between Subject Effects					
P(Age)	12.266	2	6.133	.616	.545
P(Gender)	5.750	1	5.750	.578	.451
Error	418.004	42	9.952		

Table 7.12 ANOVA for Usability Attribute "Confidence"

The results for this usability attribute also showed no significant differences between applications or experiment conditions, indicating that participants had similar levels of confidence regardless of the independent variables. The mean scores are graphically illustrated in Figure 7.5.

7.5.1.5 Usability Attribute – "Usefulness"

Participants also felt that both the applications were useful, regardless of the experiment condition that they encountered in each. The statistical evidence is presented in the ANOVA table (Table 7.13).

I did not think this service was useful	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.158	1	.158	.096	.759
Application * P(Age)	2.955	2	1.478	.895	.416
Application * P(Gender)	1.157	1	1.157	.701	.407
Error(Application)	69.370	42	1.652		
Condition	2.143	3	.714	.865	.461
Condition * P(Age)	1.009	6	.168	.204	.975
Condition * P(Gender)	2.288	3	.763	.923	.432
Error(Condition)	104.054	126	.826		
Application * Condition	3.345	3	1.115	1.686	.173
Error(Application * Condition)	83.331	126	.661		
Between Subject Effects					
P(Age)	49.334	2	24.667	1.377	.263
P(Gender)	.448	1	.448	.025	.875
Error	83.331	126	.661		

Table 7.13 ANOVA for Usability Attribute "Usefulness"

7.5.1.6 Usability Attribute – "Feeling of control"

Echoing the results of Chapter 6, this usability attribute did show a preference for the cinema application with respect to the participant's perceived level of control (mean cinema = 4.92, mean bank = 4.60). The qualitative results will show that the cinema application was preferred because many participants were uncomfortable with the notion of divulging financial information to an animated agent, regardless of the fact that in many cases they had the opportunity to enter more crucial information using the keyboard as opposed to speech entry (Table 7.14).

I felt in control using this service	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	9.341	1	9.341	8.358	**.006**
Application * P(Age)	2.017	2	1.009	.903	.413
Application * P(Gender)	1.698	1	1.698	1.519	.225
Error(Application)	46.939	42	1.118		
Condition	4.946	3	1.649	1.733	.164
Condition * P(Age)	8.438	6	1.406	1.478	.191
Condition * P(Gender)	2.689	3	.896	.942	.423
Error(Condition)	119.886	126	.951		
Application * Condition	.296	3	9.856E-02	.142	.934
Error(Application * Condition)	87.237	126	.692		
Between Subject Effects					
P(Age)	53.489	2	26.745	1.894	.163
P(Gender)	12.713	1	12.713	.900	.348
Error	592.980	42	14.119		

Table 7.14 ANOVA for Usability Attribute "Feeling of control"

7.5.2 Attitude to Agents

The second set of quantitative data addressed participants' attitudes toward various aspects of trustworthiness with respect to the embodied agents who appeared as assistants in the application environments. Figure 7.7 below illustrates the pattern of results for the agents with respect to application and experiment condition.

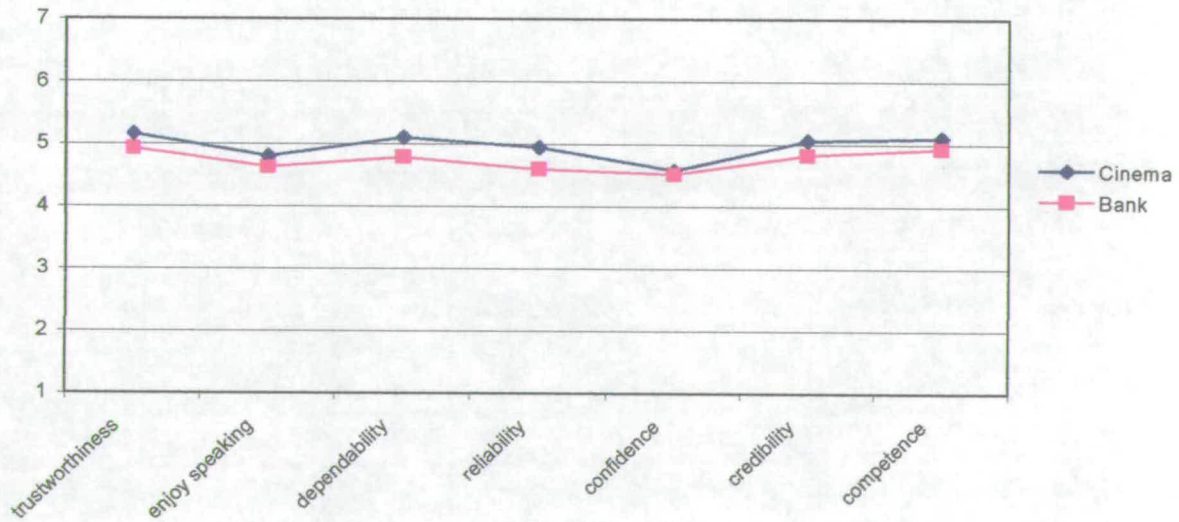


Figure 7.7(i) Usability Attributes for Agents by Application

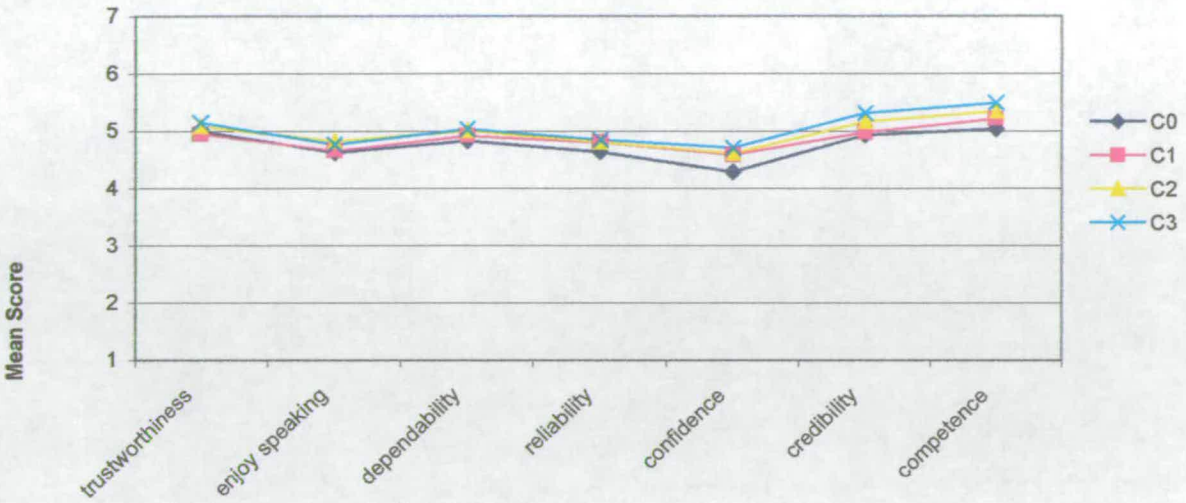


Figure 7.7(ii) Usability Attributes for Agents by Experiment Condition

7.5.2.1 Usability Attribute – "Trustworthiness"

Participants felt less encouraged to assign a degree of trust to the agents in the banking application, regardless of the experiment condition and the ANOVA (Table 7.15) results showed that the agent in the cinema application was more trustworthy than the agent in the banking application (mean cinema score = 5.15, mean bank score = 4.93).

I felt the assistant was trustworthy	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	5.025	1	5.025	7.754	**.008**
Application * P(Age)	.974	2	.487	.751	.478
Application * P(Gender)	.982	1	.982	1.516	.225
Error(Application)	27.219	42	.648		
Condition	2.162	3	.721	1.073	.363
Condition * P(Age)	3.273	6	.545	.812	.562
Condition * P(Gender)	1.846	3	.615	.916	.435
Error(Condition)	84.606	126	.671		
Application * Condition	2.339	3	.780	.986	.402
Error(Application * Condition)	99.650	126	.791		
Between Subject Effects					
P(Age)	12.380	2	6.190	.709	.498
P(Gender)	9.111	1	9.111	1.043	.313
Error	366.822	42	8.734		

Table 7.15 ANOVA for Usability Attribute "Trustworthiness"

7.5.2.2 Usability Attribute – "Enjoy speaking"

I enjoyed speaking to the assistant	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	2.993	1	2.993	9.863	**.003**
Application * P(Age)	.949	2	.474	1.564	.221
Application * P(Gender)	2.270E-02	1	2.270E-02	.075	.786
Error(Application)	12.744	42	.303		
Condition	2.493	3	.831	1.756	.159
Condition * P(Age)	5.372	6	.895	1.892	.087
Condition * P(Gender)	.429	3	.143	.302	.824
Error(Condition)	59.612	126	.473		
Application * Condition	.258	3	8.586E-02	.184	.907
Error(Application * Condition)	58.771	126	.466		
Between Subject Effects					
P(Age)	31.123	2	15.562	1.083	.348
P(Gender)	7.931	1	7.931	.552	.462
Error	603.717	42	14.374		

Table 7.16 ANOVA for Usability Attribute "Enjoy speaking"

Participants also significantly preferred speaking to the agent in the cinema application rather than the agent in the banking application (mean cinema score = 4.80, mean bank score = 4.63).

7.5.2.3 Usability Attribute – "Dependability"

Participants felt they could depend on the assistant more in the cinema application rather than the banking application to help with the task (mean cinema score = 5.09, mean bank score = 4.78).

7.5.2.4 Usability Attribute – "Reliability"

This usability attribute showed that the assistant in the cinema application was significantly more reliable than the agent in the banking application (mean cinema score = 4.95, mean bank score = 4.59).

I could depend on the assistant to do the job	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	9.184	1	9.184	6.823	**.012**
Application * P(Age)	3.336	2	1.668	1.239	.300
Application * P(Gender)	.860	1	.860	.639	.429
Error(Application)	56.530	42	1.346		
Condition	2.534	3	.845	.894	.446
Condition * P(Age)	2.026	6	.338	.358	.904
Condition * P(Gender)	1.235	3	.412	.436	.728
Error(Condition)	119.026	126	.945		
Application * Condition	1.419	3	.473	.462	.710
Error(Application * Condition)	129.100	126	1.025		
Between Subject Effects					
P(Age)	33.096	2	16.548	2.041	.143
P(Gender)	19.876	1	19.876	2.452	.125
Error	340.521	42	8.108		

Table 7.17 ANOVA for Usability Attribute "Dependability"

I did not think the assistant was reliable	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	11.540	1	11.540	7.942	**.007**
Application * P(Age)	3.335	2	1.667	1.147	.327
Application * P(Gender)	3.035	1	3.035	2.088	.156
Error(Application)	61.032	42	1.453		
Condition	2.506	3	.835	.653	.582
Condition * P(Age)	7.055	6	1.176	.920	.483
Condition * P(Gender)	2.674	3	.891	.697	.556
Error(Condition)	161.117	126	1.279		
Application * Condition	7.027	3	2.342	2.011	.116
Error(Application * Condition)	146.748	126	1.165		
Between Subject Effects					
P(Age)	60.461	2	30.230	2.952	.063
P(Gender)	23.827	1	23.827	2.327	.135
Error	430.068	42	10.240		

Table 7.18 ANOVA for Usability Attribute "Reliability"

7.5.2.5 Usability Attribute – "Confidence"

I felt confident the assistant understood me	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	.283	1	.283	.143	.707
Application * P(Age)	1.679	2	.839	.425	.656
Application * P(Gender)	.283	1	.283	.143	.707
Error(Application)	82.879	42	1.973		
Condition	9.326	3	3.109	2.505	**.042**
Condition * P(Age)	11.173	6	1.862	1.500	.183
Condition * P(Gender)	3.478	3	1.159	.934	.426
Error(Condition)	156.380	126	1.241		
Application * Condition	3.882	3	1.294	.786	.504
Error(Application * Condition)	207.547	126	1.647		
Between Subject Effects					
P(Age)	55.843	2	27.921	2.270	.116
P(Gender)	23.171	1	23.171	1.884	.177
Error	516.598	42	12.300		

Table 7.19 ANOVA for Usability Attribute "Confidence"

A significant effect emerged with respect to experiment condition for this usability attribute, and post-hoc t-tests explained showed that participants significantly felt more confident that the agents understood them when text output also appeared in the interface (Cond(2) and Cond(3)). The mean scores for condition are presented in Table 7.20.

Application	Mean Rating Score (max 7)
Cond(0)	4.28
Cond(1)	4.58
Cond(2)	4.69
Cond(3)	4.70

Table 7.20 Usability Attribute "Confidence"
Mean Scores by Experiment Condition

7.5.2.6 Usability Attribute – "Credibility"

Following the trend set by previous usability attributes the participants perceived the agent in the cinema application to be significantly more credible than the agent in the banking application (mean cinema score = 5.05; mean bank score = 4.82).

The assistant was credible	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	5.266	1	5.266	6.693	**.013**
Application * P(Age)	1.953	2	.977	1.241	.299
Application * P(Gender)	9.674E-02	1	9.674E-02	.123	.728
Error(Application)	33.044	42	.787		
Condition	4.634	3	1.545	2.830	**.041**
Condition * P(Age)	1.302	3	.434	.795	.499
Condition * P(Gender)	1.392	6	.232	.425	.861
Error(Condition)	68.769	126	.546		
Application * Condition	1.601	3	.534	.954	.417
Error(Application * Condition)	70.456	126	.559		
Between Subject Effects					
P(Age)	35.717	2	17.859	1.817	.175
P(Gender)	9.111	1	9.111	.927	.341
Error	412.782	42	9.828		

Table 7.21 ANOVA for Usability Attribute "Credibility"

In addition to this there was also a significant effect for experiment condition and post hoc t-tests showed that the agents in the interfaces that displayed text output (Cond(2), Cond(3)) were thought to be significantly more credible than the agents in the interfaces where text output was not visible (Cond(0), Cond(1)). The differences between Cond(2) and the first two interfaces (Cond(0), Cond(1)) were marginally significant (both at $p <$ 0.05), However, the differences between Cond(3) and these two interfaces (Cond(0), Cond(1)) were highly significant ($p < 0$ .01).

Application	Mean Rating Score (max 7)
Cond(0)	4.93
Cond(1)	4.97
Cond(2)	5.20
Cond(3)	5.30

Table 7.22 Usability Attribute "Credibility"
Mean Scores for Experiment Condition

7.5.2.7 Usability Attribute – "Competence"

An effect emerged indicating that agents in the interfaces which displayed text output (Cond(2), Cond(3)) scored significantly higher than those without text output (Cond(0), Cond(1)) and these agents were perceived as being more competent.

The assistant was competent	Sum of Squares	df	Mean Square	F	p
Within Subject Effects					
Application	2.721	1	2.721	2.137	.151
Application * P(Age)	1.910	2	.955	.750	.478
Application * P(Gender)	1.943E-02	1	1.943E-02	.015	.902
Error(Application)	53.468	42	1.273		
Condition	11.225	6	1.871	2.795	**.014**
Condition * P(Age)	8.666E-02	3	2.889E-02	.043	.988
Condition * P(Gender)	11.633	3	3.878	5.793	**.001**
Error(Condition)	84.337	126	.669		
Application * Condition	1.061	3	.354	.419	.740
Error(Application * Condition)	106.397	126	.844		
Between Subject Effects					
P(Age)	47.343	2	23.671	2.867	.068
P(Gender)	14.001	2	7.000	.848	.436
Error	346.745	42	8.256		

Table 7.23 ANOVA for Usability Attribute "Competence"

Pair-wise comparisons showed that there were marginally significant differences between Cond(2) and Cond(1), and between Cond(2) and Cond(0), both at $p < 0.05$. This effect was highly significant between Cond(3) and Cond(0), and Cond(3) and Cond(1), both at $p < 0.01$.

Application	Mean Rating Score (max 7)
Cond(0)	5.03
Cond(1)	5.20
Cond(2)	5.34
Cond(3)	5.49

Table 7.24 Usability Attribute "Competence"
Mean Scores for Experiment Condition

7.5.3 Application Preferences

Participants had the opportunity to express their application preference and were encouraged to give reasons for their choice. The results of participants' preferences are illustrated graphically in Figure 7.8. The results show that the cinema was the preferred application. These qualitative results augment the quantitative results already reported and the results can be neatly summed up by quoting one of the participants, who stated

that "*Although the experiences were similar, I preferred the cinema because it was a more entertaining service and was more enjoyable than the bank, but this is the difference between the nature of the services and I prefer to do more entertaining things*". The results suggest that no matter what additional features are added to the interface, so much that they improve the interaction for the user, more entertaining tasks may always be preferred. Nevertheless, as with the findings in Chapter 6, the results here also indicate a lack of confidence in the capabilities of the banking system, resulting in a reluctance to respond positively to the application and although the addition of interface features such as text output may improve the interaction, this may not alone improve the experience for the user, so much so that they would actually complete banking tasks or display a willingness to exchange financial information with an embodied conversational agent.

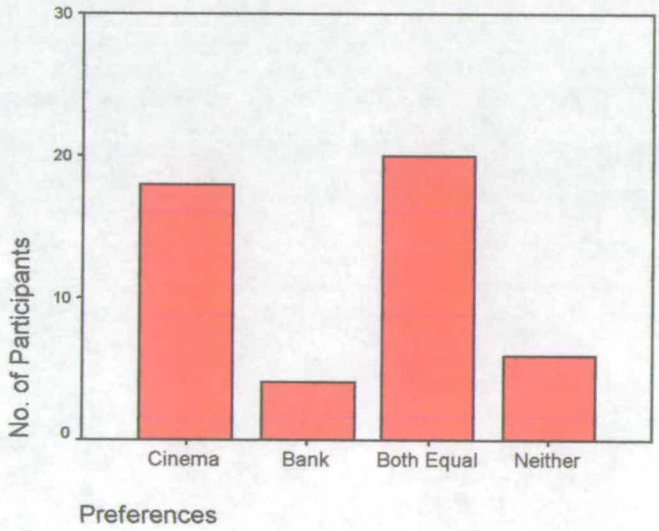


Figure 7.8 Participants' Preferences for Application

7.5.3.1 Cinema Application

Participants were asked to state any general comments about the cinema application. Of the 48 participants in the experiment, 36 responded positively to the cinema application claiming it was easy to use, straightforward and it would be a very useful service and had lots of potential. Participants stated that the application was very realistic giving the feeling that they were in the cinema, providing a real-life situation. The assistant in the application was thought to be polite, pleasant and competent and overall the interaction

was efficient and one participant is quoted as saying that they "*surprisingly felt more engaged with the assistant and was better than telephone interactions because you actually see the person*". Of the 12 participants who responded negatively to the cinema application many were largely concerned with the speed of the interaction, stating it was too slow. One participant stated that the "*delay was confusing and made it unnatural*".

7.5.3.2 Banking Application

Participants expressed concern about this application and stated that they are "*more wary, because it deals with my money*". Another participant is quoted as saying that "*I am concerned with this [banking application], because of the risk factor and would want it to be more reliable*". Although the quantitative results were not significant with respect to reliability participants did state that interface features such as text output were encouraging, for example one participant is quoted as saying that "*it is much more useful seeing the assistant's monitor, because seeing is believing*". However, the delays in the system's responses made the interaction less than authentic. Although the cinema application may have been preferred because of its entertaining nature, these qualitative findings suggest that there is potential for users to engage in banking with ECA, particularly when the interface also provides text output and more so when the speed of the real-time interaction is improved.

7.5.4 Interview Feedback

Participants took part in a structured interview after they had experienced the four experiment conditions in both applications. The questions primarily probed for information about users' attitudes to the different versions of the systems, corresponding to each of the four experiment conditions. Each of the features was mentioned and participants were encouraged to comment on the benefits of such features in the interface.

7.5.4.1 Speech Input and Output

Participants were invited to give comments as to how they felt about speaking to the assistants and were asked how they felt in both the cinema and banking applications. Overall participants enjoyed speaking to the assistants, but did admit that it was

annoying and sometimes frustrating if the assistant did not recognise what was begin said. Participants said that the interactive experience is an improvement to just looking at text on the screen. It was also thought to be an improvement on the telephone adding a more "personal touch" to the interaction. There were comments from some participants about the speed of the systems and it can be deduced that if the delays are eliminated the spoken interaction could be more enjoyable.

7.5.4.2 Text Input

Approximately half of the participant sample felt that having the opportunity to type was a good option, in particular in the event of any mis-recognition. The other half of the sample expressed satisfaction at just speaking to the assistant, as it seems more "*efficient and easier*". One participant stated that she didn't think to use the keyboard when the option was there, because the agent started the interaction by speaking and so it seemed natural to also reply using speech. Another participant commented that she found it easier to stick to the one way of entering data, rather than two modes. Four participants commented that entering data via the keyboard perhaps was more secure, as people would not overhear the information being entered and the number could be entered discretely.

One quarter of the participant sample thought the number pad was a good option in the interface. In total, seven participants actually attempted to use it. Comments from these participants suggested that the number pad was good for privacy and security. The majority of participants (36/48) were not bothered with the number pad, comments being that they didn't notice it or were happy speaking to the assistant. One participant stated that they would rather use the keyboard to enter the security number. Another commented that as long as the assistant could understand there is no need to use it. Another commented they would rather use the keyboard instead of the mouse and four participants commented that it seemed unnatural to use it and would be better to use the keyboard. Another commented that they did not use it because it seemed too complicated to switch between speech and keyboard and mouse. Three participants commented that the number pad may serve a great purpose in the banking application.

7.5.4.3 Text Output

All participants in the experiment responded positively to seeing the information on the assistant's monitor. Participants commented on the helpfulness of seeing the monitor, and by providing this textual information the user was more reassured that the information was received correctly. The monitor acted as a verification tool, providing "*excellent feedback*" and it is "*essential to see it*". Some comments stated that it was a good idea to see this information in combination with the speech combination and that the two methods complemented each other and increased confidence. Participants felt more reassured, and there was less room for doubting that the inputted information was incorrect. Overall the textual output was a helpful and useful tool to have in the interface. Again the majority of participants stated it was important to see this in both the applications.

7.6 Discussion

The primary aim of this experiment was to introduce additional features to retail interfaces to investigate if they could significantly improve user confidence, especially in applications where users must disclose financial information. A virtual cinema box-office application and a banking application were used in the evaluation. The cinema application was assessed in a previous evaluation (Chapter 6) where it was perceived as being significantly more trustworthy than the banking application. Interface features, which included the addition of text input and text output, were introduced in a repeated measures design. Three predictions were made prior to the evaluation, two of which were actually supported. However, a number of interesting findings emerged associated with the design of retail interfaces that could elicit target feelings such as trustworthiness and other associated attributes.

The first prediction stated that combining text input with speech input would improve user confidence in the system, and therefore infer a greater sense of trustworthiness in the conversational agent and the interactive application as a whole. The overall mean trend for the usability attributes showed that there were no significant differences between the experiment conditions. In fact the inclusion of text input in the interface did not alter the reliability, efficiency, usefulness or improve user confidence within the

applications. Participants did feel however that when speech and text input were combined with speech and text output, the application was considered to be a significantly better idea than simply having speech input alone.

It was predicted that text input in the interface would be an improvement to the banking application, offering more security and privacy for the user. Participants thought the banking application was less reliable than the cinema, and they also felt more in control when completing tasks in the cinema application. However, no interactions emerged between the independent variables of application and experiment condition, suggesting that regardless of the inclusion of text input participants thought the applications were good ideas, they felt confident using the applications and they were equally efficient and useful. The qualitative analysis specifically focusing on the text input feature did somewhat augment these findings where the overall perception of textual input was that it was a good idea, that its presence was reassuring, however users were satisfied using one mode of input and since the interaction began via speech, this seemed to be the most natural way to complete the interaction. It is therefore concluded that the first prediction was not supported and that text input in the interface did not improve aspects of user confidence. However the interface was considered to be a significantly better concept with text input, in particular when it was complemented with text output (Cond(3)), and the addition of text input does serve as a convenient function for the user if necessary.

The second prediction stated that text output combined with speech output would improve user confidence in the system over speech output alone. This prediction was supported, and when speech and text output were combined with speech and text input this improvement was significant with respect to some attributes examined in the questionnaires. With respect to the independent variable of application, no differences were evident, and regardless of the inclusion of text output in the interface, participants thought the applications were equally reliable, efficient and useful. They felt more in control using the cinema application and also thought it was more reliable, but the lack of an interaction between applications and experiment conditions showed that this difference was independent of the presence of textual output. As mentioned earlier, participants did think it was significantly better to include some interface features, and when textual output was combined with textual input this was significantly better than providing the user with speech output alone.

The results favoured the agent in the cinema application, who was thought to be significantly more trustworthy, dependable, reliable, and credible than the agent in the banking application. Independent of application, the addition of text output did improve other aspects of the perception of the agent and users in particular felt significantly more confident that the agent understood them better when the interface had text output and the presence of the information in the interface reassured the user that the information was correct. In addition results showed that the agent was also thought to be more credible and competent when the interface had text output. Although there was not detected difference with respect to trustworthiness directly, the associated attributes of confidence, credibility and competence were significantly improved by the presence of text output.

When the interfaces combined text input with text output the difference between the experiment conditions were also significant. In support of the third prediction (Prediction 7.3) the combination of the two modes of communication was an improvement to the system. Participants felt that the applications were more reliable when they combined speech and text input with speech and text output, and this was significantly the attitude of the female participants in the sample. The participant sample also felt the agents was more credible and competent when combining both speech and text input with speech and text output.

Although the results favour the cinema application with respect to attributes of trustworthiness the evaluation has shown that including text output in the interface can infer greater agent credibility, competence and confidence of understanding, however these alone are not sufficient to infer a more trusting and comfortable environment for users to disclose financial information in banking applications. It is known that trust is rarely established instantaneously (Nikander & Karvonen, 2000; Karvonen, 1999) and virtual applications in which financial information is disclosed are not perceived as trusting environments. Despite the fact that some interface features can elicit trusting emotions in users (Kim & Moon, 1997) this experiment showed that trust remains difficult to establish and maintain in certain applications. However, it is also known that trust can be established through displays of competence, and the results of this evaluation show that text output in the interface infers greater competence, making it possible to deduce that text output in these interfaces may be a factor in improving user trust in banking applications.

Referring to McKnight's (1996) framework for trust produces evidence of further improvements that are likely to improve trust in the interface, in particular in applications where users must disclose personal financial information. As mentioned trusting beliefs are formed when users have confidence in their intentions and the results of this application show that users are more confident that they were understood when text output is displayed, thus deducing that they have more confident intentions toward the applications, which can manifest itself as more trusting behaviour.

Convincing potential users that transactions over the interface are secure and safe is a challenging issue, but is likely to be the key to improving the situational and contextual decision to trust. To advance the research that precedes the construction of usable tools the correct system safeguards must be implemented. This evaluation showed that text output in combination with speech output may be one factor to improve user confidence in retail applications and thus trust in agents, but other interface features and techniques still need to be considered.

7.7 Summary

This evaluation completes a series of experiments exploring the use of embodied conversational agents in retail applications. Using the interactive system that was described in Chapter 6, the empirical evidence showed that despite the addition of certain interface features, there was no overall improvement in the perception of the trustworthiness of a banking application inhabited with embodied conversational agents. However, the results showed that the spoken language interfaces augmented with textual output significantly improved attitudes to an agent's credibility and competence and it is displayed competence that is one of six key issues to demonstrate in order to form a trusting environment. The results showed that text output also improved user confidence that the agent understood them during an interaction. Although users' perception of the trustworthiness of agents in banking applications were not improved, this evidence is suggestive that text output together with speech input and output could serve as a component of banking interfaces as a mechanism to initiate a trusting environment through displayed competence.

Chapter 8

Research Contributions and Design Implications with Respect to Embodied Conversational Agents in Virtual Retail Environments

8.1 Main Findings

The research in this thesis was undertaken to advance knowledge of the effectiveness of ECA's in electronic retail applications and to make an empirical contribution to the innovative area of embodied conversational agents. The thesis engineered computer interfaces to allow the evaluation of the effectiveness and efficiency of communication between computers and their human users through the modality of speech with the addition of non-verbal behaviour provided through the creation of embodied conversational agents. In this way, users could use their innate communicative capabilities for the benefit of a more engaging interaction. The research showed that anthropomorphising the interface with lifelike behaviour animated this communicative process and the extension of the persona metaphor to retail applications was successful, highlighting the potential of personalised real-time interactions for users. Throughout the thesis the research strategy employed aimed to provide design guidelines by gathering empirical evidence to support the successful deployment of ECA's in retail applications. The interdisciplinary investigation comprised a series of progressive evaluations which reflected the research themes of assessing attitudes toward agents' physical realisations, agents' gender and perception of agents' behaviour, personality and their functionality in retail applications. Table 8.1 provides a summary of each of the four empirical evaluations and the main findings gathered from quantitative and qualitative results.

At the beginning of the thesis, Figure 1.2 illustrated the fact that humanoid ECA's could be physically realised in many different forms: 2D, 3D, animated or photo-realistic. It was made clear in Chapter 2 that there was little prior research available comparing the benefits of these different variations of the persona metaphor. The first two experiments in the thesis addressed these issues by evaluating a range of humanoid male and female photo-realistic agents. In the second evaluation a range of humanoid male and female

animated agents were also included. To complete these experiments a retail interface template, described in Chapter 4, was constructed to serve as an experiment platform from which to assess the ranges of embodied conversational agents. The evaluations contributed evidence in support of conversational agents in retail interfaces, but showed that certain physical realisations of agents are preferred.

Evaluation Type	Evaluation Topic	Evaluation Findings
Passive Viewing	Photo-realistic Agents	Conversational retail interfaces are liked. Photo-realistic agents must display sophisticated human-like non-verbal behaviour.
	Animated Agents	3D fully-embodied agents are perceived as being more competent, friendly and polite in comparison to 2D counterparts or talking heads.
Interactive Participation	3D Embodied Agents: Appearance and Behaviour	The degree of formality of presentation of the agent should be consistent with the perceived formality of the application. Agents in banking tasks are perceived to be less trustworthy than agents in informal applications.
	Agent Trustworthiness	Spoken language interfaces augmented with textual input are convenient in case of mis-recognitions (see Section 7.5.4.2) Spoken language interfaces augmented with textual output offers reassurance to the user (See Section 7.5.4.3). Spoken language interfaces augmented with textual output in the interface improves perceived agent credibility and competence. Spoken language interfaces augmented with textual output does did not appear to improve agent trustworthiness.

Table 8.1 Main Experiment Findings

8.1.1 Evaluating Humanoid Photo-Realistic Agents

This evaluation included male and female video agents, 3D talking heads, images with facial movement, still images and disembodied voices. Using observation techniques as a method of assessment Chapter 4 reported on the results from the evaluation of humanoid photo-realistic agents, and showed that participants expected these photo-realistic agents to display corresponding human-like non-verbal behaviour, otherwise they preferred to use the auditory mode of communication alone, stating that the visual mode is otherwise

a distraction. Because participants expected the photo-realistic images to display human-like non-verbal behaviour, the video agent types were rated the highest and were also perceived to be more competent, more friendly and more natural with regard to their overall appearance. If the agent did not display sophisticated communication strategies, it was found that the users' attitude was significantly reduced, thus discouraging further interactions.

8.1.2 Evaluating Humanoid Animated Agents

The main findings reported from the second passive viewing experiment (Chapter 5), which assessed humanoid animated agents using the same retail interface template for the first passive viewing evaluation, showed that during a comparison of 2D and 3D animated agents realised as talking heads and fully-embodied agents, 3D fully embodied characters conveyed a greater sense of friendliness, lifelikeness and politeness. In the preceding evaluation of humanoid photo-realistic agents there was a dichotomy of response within the participant sample regarding the presence of a visible agent in the interface. Should the agent not visibly meet participants' expectations, then an agent realised as a disembodied voice would be sufficient to act as a retail interlocutor. In this evaluation of animated agents, more encouraging support for ECA's was found with two-thirds of the participant sample stating that the presence of a visible agent in the form of an animated character enhanced the application. Although the experiment did determine that participants had a preference for the agents to be realized in three-dimensions and for them to be full bodied, a number of unanswered issues emerged with respect to the deployment of such characters in 3D virtual retail applications. For instance, how should the 3D agent appear and behave in different 3D application environments? Is the personality and functionality of the agent perceived differently in interactive situations? As discussed in Chapter 6, graphically representing interfaces in three dimensions is becoming more feasible with the latest developments in 3D descriptive languages (Web Consortium, 2001), making it possible and likely that such environments could be inhabited with 3D animated fully embodied agents. Consequently, the third evaluation in this thesis was designed to investigate underlying issues with respect to the use of 3D animated agents as assistants in 3D retail applications.

8.1.3 Assessing 3D Embodied Agents

An interactive spoken language system was designed and implemented to act as an experiment platform from which to evaluate issues relating to 3D ECA's in retail applications. A technical description of the system was provided in Chapter 6 followed by a description of an interactive experiment where participants were invited to engage in conversation with 3D agents in three retail applications in order to evaluate perceptions of the agents' gender, appearance, personality, behaviour and functionality in contrasting applications. The applications themselves were selected to represent different points on the spectrum of more and less serious tasks to be completed. The first, the cinema box-office application, was selected as an entertainment driven task; the second was a travel agency with a more complicated booking task; the third was a banking task where users completed a personal financial transaction. The results showed that participants enjoyed conversing with the agents in these novel interactive applications. No agent gender differences emerged with respect to the agents in the different applications, but it was discovered that participants expected agents to be informally dressed agents for informal applications, such as a cinema box-office, and agents to be formally dressed agents in more serious financial applications, such as banking. Regardless of their appearance, the banking applications inspired less confidence in users and the agents that appeared in this application were perceived to be less trustworthy than the agents in other applications. As the establishment of a trusting environment is imperative in order for successful interactions and transactions to be completed in retail applications, this became the research motivation for the fourth and final empirical evaluation in this thesis. The evaluation contributed evidence in support of the inclusion of multi-modal features in the interface, such as textual output, to instil a greater sense of user confidence and credibility in the application.

8.1.4 Methods to Improve Trustworthiness

This experiment used the interactive spoken dialogue system described in Chapter 6. Interface features, which included the addition of text input and text output, were added to the cinema box-office application and the banking application. In a repeated measures design, participants interacted with agents and experienced each of the different input and output modalities. Text input in the interface did not improve user confidence within the applications, although participants did feel its presence was reassuring. When

speech and text input were combined with speech and text output, there was a significant improvement over having speech output alone. Participants felt more confident that the agent understood them better when the interface had text output and the presence of the information in the interface served to reassure the user that the information was correct. Participants felt that the applications were more reliable and credible when they combined speech and text input with speech and text output. The participant sample also felt the agents in the applications were more credible and competent when both speech and text input and speech and text output were featured in the interface. Although the results favoured the cinema application with respect to perceived agent trustworthiness the evaluation showed that including text output in the interface can instil greater agent credibility, competence and confidence of understanding in the user. Although these alone are not sufficient to produce a trustworthy agent, the experiment demonstrated that they could serve to improve user confidence in certain applications as a means toward creating a trusting environment.

8.1.5 Attitude to Agents' Voices

Throughout the evaluations many findings emerged with respect to the voices of both the male and female agents. In the first evaluation (Chapter 4), the results showed that the physical realisation of the agent types produced significant cross-modal effects with respect to the quality of human-like voice outputs. Despite the fact that the one male voice was used for all five male photo-realistic agents and all five female agents had an identical female voice output, participants preferred the voices of the video agents and the disembodied voice agents to the other agent types. They also perceived the voices of these two agent types to be clearer and the conversation between the customer and these agents to be more natural. These results suggest that if the agent's visual appearance does not complement the verbal behaviour, attitude to the agents' voices may be significantly lower.

The second experiment showed a general preference for the female voice over the male voice, which was the opposite finding to that reported in Chapter 4, where the male voice was preferred to the female voice. After the poor perception of the voices in the first experiment, new voices were selected for the voice output of the animated agents of Chapter 5. The intention was to give the agents fluent, conversational speech. The results showed in this case that the new female voice was significantly more natural than

the new male voice. This result, combined with similar evidence in Chapter 4, strongly indicates that if and when selecting natural rather than synthetic voices for agents it is essential to pre-evaluate the voices before using them.

In the interactive system that was developed and used for the evaluation of the 3D embodied agents in the 3D retail environments, concatenated speech was used to allow real-time speech interaction between the agents and the users. As described in Chapter 6, pre-recorded natural speech prompts were concatenated to form each of the agents' output sentences. Despite the fact that the agent voices were pre-evaluated prior to building the system, various other issues did arise. For instance, due to the concatenated nature of the output utterances from the agents, in particular with respect to the male voice, it was felt that the concatenated male utterances were not as natural as the female voice output. It is important in the development of real-time applications that caution is exercised to ensure that the transitions between words in agents' output sentences are perceived as being natural.

8.2 Interface Design Implications

As previously stated, this empirical research was carried out with the intention of creating design guidelines and interface development strategies for the creation of effective ECA in virtual retail applications. The first experiments used evaluation by observation techniques; later experiments used real-time interactive spoken language systems to produce empirical information in support of the use of embodied conversational agents in retail applications. In Chapter 2 it was speculated that the appearance of ECA in retail applications could serve as effective marketing tools, increasing company profiles, offering personalised twenty-four hour services to customers and improving customer relationships with companies. From the potential customer's perspective, this thesis showed that the use of embodied conversational agents in retail applications is welcomed and the conversational capabilities of the interfaces were liked by the majority of participants that took part in the experiments. As stated the research themes focused on users' attitudes towards the agents' physical realisation, agents' gender and perception of agents' behaviour, personality and their functionality in retail applications. Many salient design implications emerged from empirical findings detailed throughout the thesis. Firstly, with respect to the

representation of humanoid agent types in the interface it is important to ensure, when using photo-realistic agents, that the agents' non-verbal behaviour is sophisticated enough to complement the sophisticated graphical image. Addressing the alternative representation of agents in the interface, the use of animated agents realised as 3D fully embodied agents can promote the friendliness, lifelikeness and politeness of the agents. These factors are promoted through the agents' gesturing and three-dimensional representation. In order to create successful 3D representations, it is essential that the agents have the correct human movement corresponding to the physical make-up of the human body, otherwise the agent will appear unnatural.

Agents represented as disembodied voices are the preferred alternative for conversational assistants should the visual representation of the agent not meet user expectations. However, the empirical evidence also showed that caution must be exercised when selecting natural voices for the agents. Firstly, fluent conversational voices should be selected and it is recommended that these voices be evaluated prior to incorporating them in interfaces. A second recommendation relates to the use of natural voice in real-time interactive systems. As concatenated speech is often used for speech output, this thesis strongly suggests that transitions between the concatenated words must be made to sound natural and be given the correct inflection.

Despite the fact that some gender differences did emerge during the experiment, these could be accounted for by cross-modal effects of perceptions of voice and the physical representation of the actual agent type. It is therefore recommended to use either male or female agents as assistants, as both offered similar enhancements to real-time retail applications, in particular applications such a virtual cinema box-office, a virtual travel agency and a virtual bank. However, the appearance of these agents whether male or female, must be considered. It is recommended the agents appear formally or smartly dressed in more formal applications such as banking, and agents should be informally or casually dressed in less formal applications, for example a cinema box-office. As 3D retail interfaces mirror real life environments, the appearance of the agents that inhabit these applications should also mirror their real-life counterparts.

With regard to design implications of the applications themselves, more serious applications, in particular those that require users to conduct financial transactions, are thought to be less trustworthy than other more entertaining tasks, and users are unlikely or unwilling to disclose information to animated agents. In order to improve user

confidence in banking applications that are inhabited with ECA it is recommended to provide textual output on the interface detailing the information that is being exchanged. This interface feature supports the interaction providing the user the necessary information through not just one audio channel, but also through visible textual output.

8.3 Future Work

As desktop virtual reality graphics are advancing, more 3D e-commerce sites have begun to appear on the Internet in an attempt to create more intimate interfaces between humans and computer imagery (Brand, 1988). As Chittaro (2000) states these immersive environments promise to make interactions more natural, attractive and fun for users and have the advantage, if correctly implemented,

- emulating the real-world shopping experience,
- supporting buyer's natural shopping actions (walking, browsing),
- satisfying the emotional needs of the buyer, through immersion and interaction,
- satisfying the social needs of the buyer, through interactions with sales people.

Although this thesis did find that interactions with 3D embodied agents in 3D retail environments were welcomed, further study needs to be carried out in order to investigate the actual spatial representation of the environments. Attitudes, perceptions and expectations toward navigation in these environments also need to be analysed to determine how these digital spaces should be represented and also what the content should be. How far does the metaphor need to be extended or realised for interactions in these virtual worlds to be successful?

Future research should also focus on the representation of virtual humanoid agents in environments. This thesis did show significant preferences for humanoid agents to be realised as 3D embodied characters, and subsequently demonstrated that attitudes to the physical appearance of these agents can vary depending on the context of the retail application. The evaluations in this thesis employed agents that interacted with users in constrained environments, and although the agents effectively demonstrated real-time

capabilities for task specific interactions, more research needs to be completed in order to create robust agents that display autonomous natural human-like movement. The importance of non-verbal behaviour emerged in Chapter 6, where it was shown that the politeness, friendliness and lifelikeness of the embodied characters were promoted through gesturing. Churchill et al. (2000) call for more formal evaluations of agents' gesturing, to determine user perceptions of this non-verbal behaviour for the benefit of the interaction. To create real-time autonomous gesturing, rule-based parameterised generation is being explored (Badler, 1999). However to create effective, believable and natural movement, gesture alignment and gesture transition models need to be created.

Although virtual reality interfaces and virtual agents demonstrate that computer interfaces can be more intuitive (Billinghurst, 1996), more research is required specifically focusing on the representation of the application environment, together with developments that govern the physical expression of the embodied agents. However, to be intelligent, it is necessary, but not sufficient for these agents to be physically expressive. To be intelligent these agents need to be also emotionally expressive and it is through such emotional behaviour that the social competence of the agents may be used as means to influence user satisfaction (Ball & Breese, 2000).

Integral to the establishment of a social relationship is the perception of a trustworthy environment. As explained in Chapter 7 trust is difficult to establish and even harder to maintain and such facts outline more challenging problems that need to be addressed in order to convince users to converse with agents to complete retail tasks in virtual environments. To do this it is suggested that systems provide users with evidence of successful interactions and display intelligent back-up upon system failure, leaving the user with sufficient confidence that their transactions are both safe and secure. To study the effects of these various mechanisms it is necessary for longitudinal user studies to be completed. Such studies would gather information over time and monitor users' interactions with agents on repeated interactions. As explained in Chapter 3, the importance of user input should not be underestimated as their input has shown to be essential to the creation and development of novel systems and applications. Longitudinal studies could help to determine long-lasting conclusions about the agents. As this thesis based its research on evidence of users' first encounters with applications and the conversational agents that appeared as assistants in these scenarios, and although it was determined that the majority of participants expressed a desire to interact with the

agents in fully-functional applications, the effects of subsequent interactions need to be addressed, to investigate how interactions develop over time and how expectations from each user may change.

Minsky (Stork, 1997) explains that computational problems are being compounded by the computer's inability to reason and learn and in his landmark book, 'The Society of Mind' (Minsky, 1985) he argues that in order to give the computer intelligent capabilities, the combined functionality of smaller agent programs, when working together, could produce rational reasoning, learning, creativity and understanding in agent based systems. Each agent sub-program would have responsibilities and capabilities for certain operations. Separately, the output of these individual agent programs is not effective, but the combined mechanisms of each, results in 'effective agency' (Massaro, Cohen, Beshow & Cole, 2000). Through this 'effective agency' the machine can display intelligence and understanding and as a result the first steps can be taken towards creating social relationships between agents and users. To create these personalised social interactions, the agents must have the capability to adapt to the individual user and then adapt to that user's temporal state. To do this, the agent must have some level of memory capability, and must then be able to decipher and respond to the user's changing state effectively for the benefit of the user. This affective agent capability has become known as emotional intelligence or adaptive social awareness.

According to Goffman (Kendon, 1988), in order to create effective interaction between agents and their human users there must be a sense of co-presence. The concept of co-presence describes how the members of a conversation can sense each other and can adjust their actions appropriately by gauging the others.

> Persons must sense that they are close enough to be perceived in whatever they are doing, including their experiencing of others, and close enough to be perceived in this sensing of being perceived.

For agents and human users to be perceived as being co-present, the agent must be capable of adjusting its actions by gauging the actions of the human user. Essentially the agent must then be sensitive to the user's goals, intentions and desires during the interaction. This thesis suggests future work to create such socially aware agents, by giving agents intentions, goals, desires and beliefs and allowing them to respond to a

user's intentions, goals, desires and beliefs. In Turkle's opinion (Turkle, 1984) machine limitations are attributed to the lack of intentionality, as it is intentionality that is the means for social competence to be displayed in order to influence user satisfaction.

A believable agent is defined as one that has the ability to display and feel emotions and must be able to communicate and interact following the rules of face-to-face interaction. It must be able to conceive, represent and convey all the possible meanings that natural language and multimodal interaction may convey in humans. In agreement with Pelachaud and Poggi further research is necessary, with particular attention being paid to the rules that can be used to formalise, represent and implement emotional and social agents (Pelachaud & Poggi, 2001).

Baron-Cohen (1997) describes the condition where one is blind to the social and mental state of intentionality as 'Mindblindness'. This is a salient symptom seen in autistic children, and the research into this psychological area can be applied to determine and describe conceptual models that may be used to develop social awareness in autonomous interactive agents. In order to create a conceptual cognitive model of these communicative features, Baron-Cohen presents the 'Theory of Mind' and to date some recent work in robotics has shown that elements of this theory can be effectively applied to embodied robots to provide them with social awareness or emotional intelligence. This thesis proposes further work in developing this theory of mind for software agents, giving the agents both intentions and goals, desires and beliefs in order for the agent to communicate socially for the benefit of the interaction as a whole. There is a need to project the illusion of awareness and intentionality from the agents, not just the illusion of life.

Social interactions depend upon the recognition of other points of view, other mental states, and the recognition of complex non-verbal signals of attention and emotional states. Social dynamics rely upon the ability to correctly attribute beliefs, goals, percepts and intentions to other people. Recognising that others have perceptions, goals and intentions that differ from our own is critical to human development, self recognition and grounding (Kendon, 1988). Baron-Cohen's model has a number of modules, which when combined together aim to recognise perceptions, goals and intentions in others. These modules include an 'intentionality detector' that detects the basic movements of

approach and avoidance, an eye direction detector to interpret gaze direction as a perceptual state, and a shared attention mechanism that detects mutual understanding.

Scassellati (2001) believes that a robot with such modules would allow for social interactions, that were previously not possible. The robot would be capable of expressing its internal state and could recognise the goals and desires of others. Scassellati's research investigates the initial components to develop these detectors, such as the ability to detect and distinguish animate and inanimate motion. When the robot recognises animate motion, it begins to use the other detectors to infer the social dynamics. Through the use of vision detection processes a software agent could be given the capability to detect the human user, the user's gaze and then manipulate this information for increased social awareness. Already, Magnenat-Thalmann (2000) has shown that the user's facial display can be recognised and interpreted, and inferences about an emotional state can be made from observations of emotional expression and behaviour in order to generate synthetic emotional states in agents. There exist systems such as the Affective Reasoner by Elliot (1992), which addresses the problem of representing emotions, by grouping emotions according to cognitive conditions. Elliot's system uses the 'OCC Model' (Ortony, Clore & Collins, 1988) to synthesise cognitive emotions, which demonstrates the capability for systems to reason about emotions. Future work is required to combine these emotion synthesis systems as part of the system requirements for the 'brains' that will inevitably be part of smart autonomous agents, that can effectively respond to users' states to create personalised real-time experiences, which, as stated in Chapter 2, is one of the primary aims to boost electronic retail services inhabited with embodied conversational agents.

> The excitement is not so much in what has been accomplished but in what has yet to be done. And there is satisfaction in working on one small piece of a puzzle whose shape is continually unfolding.
>
> Machinery of the Mind (Johnson, 1987)

Bibliography

Allbeck, J. (2001). Consistent Communication with Control. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents*. 5th International Conference on Autonomous Agents, 21-26.

Allwood, J., Nivre, J. & Ahlsén, E. (1990). Speech Management: On the Non-Written Life of Speech. *Nordic Journal of Linguistics*, 13:1-48.

André, E., Müller, J. & Rist, T. (1999). Employing AI Methods to Control the Behaviour of Animated Interface Agents. Journal of *Applied Artificial Intelligence* 13: 415-48.

André, E. and Rist, T. (1998). 'Personalising the User Interface: Projects on Lifelike Characters at DFKI'. Proceedsing of the *3rd Workshop on Conversational Characters*, Oct,167-170.

Arafa, Y. & Mamdani, A. (2000). Virtual Personal Service Assistants: Towards Real-Time Characters with Artifical Hearts. Proceedings of the *International Conference on Intelligent User Interfaces*, New Orleans, LA.

Arafa, Y., Dionisi, G., Mamdani, A., Martin, S., Pitt, J. & Witkowski, M. (2000). Towards Building Loyalty in ECommerce Applications: Addressing Issues on Personalisation, Persistence & Presentation. Proceedings of the *Workshop on Agents in Industry*. 5th International Conference on Autonomous Agents.

Ardissono, L. & Goy, A. (1999). Tailoring the Interaction with Users in Electronic Shops. Proceedings of the 7th International Conference on User Modeling, 35-44. Canada, Berlin: Springer-Verlag.

Argyle, M & Cook, M. (1976). *Gaze and Mutual Gaze*. England: Cambridge University Press.

Badler. N., Bindiganavale, R., Allbeck, J., Schuler, W., Zhao, L. & Palmer, M. (2000). Parameterized Action Representation for Virtual Human Agents. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents*, Cambridge, MA, London, England: MIT Press. ISBN 0-262-03278-3.

Badler, N., Palmer, M., and Bindiganavale, R. (1999). Animation Control for Real-Time Virtual Humans. In *Communications of the ACM*, Vol. 42, No. 8, 65-73.

Badler, N., Phillips, C., & Webber B. (1993). *Simulating Humans: Computer Graphics Animation and Control.* New York: Oxford University Press.

Ball, G., and Breese, J. (2000). Emotion and Personality in a Conversation Agent. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents*, Cambridge, MA, London, England: MIT Press. ISBN 0-262-03278-3.

Ball, G., Ling, D., Kurlander, D., Miller, J., Pugh, D., Skelly, T., Stanosky, S., Thiel, D., and Wax, T. (1997). Lifelike Computer Characters: The Persona Project at Microsoft Research'. In J. M. Bradshaw (ed.), *Software Agents.* Menlo Park, California: AAAI/MIT Press, 1997. ISBN 0-262-51090-1.

Baron-Cohen, S. (1997). *Mindblindness.* Cambridge: MA; London, England: MIT Press. ISBN 0-26252225-X.

Bates, J. (1994). The Role of Emotion in Believable Agents. In *Communications of the ACM* Vol. 37, 122-25.

Bates, J. (1991). Virtual Reality, Art and Entertainment. In *Presence: The Journal of Teleoperators and Virtual Environments.* Cambridge, MA, London, England: MIT Press.

Bernsen, N., Dybkjær, H. & Dybkjær, L. (1998). *Designing Interactive Speech Systems: From First Ideas to User Testing.* Berlin: Springer-Verlag. ISBN 3-540-76048-2.

Bernsen, N. (1996). Towards a Tool for Predicting Speech Functionality. In *Speech Communication*, Vol. 23.

Bickmore, T. and Cassell, T. (2000). How about this weather? Social Dialogue with Embodied Conversational Agents. Proceedings of *Socially Intelligent Agents: The Human in the Loop.* AAAI Fall Symposium, 4 –9. ISBN 1-57735-127-4.

Biermann, A. et al. (1997). *More Than Screen Deep: Toward Every-Citizen Interfaces to the Nation's Information Infrastructure.* National Academy Press.

Billinghurst, M. & Savage, J. (1996). Adding Intelligence to the Interface. Proceedings of the *IEEE Virtual Reality Annual International Symposium*, 168-73.

Bohlin, P., Bos, J., Larsson, S., Lewin, I., Matheson, C., and Milward, D (1999). *Survey of Existing Interactive Systems.* Trindikit 1.0 Manuel, Deliverable D1.3, TRINDI.

Bolt, R. (1985). Conversing With Computers. *Technology Review*, March. Cambridge, MA: MIT Press.

Bradshaw, J (1997). An Introduction to Software Agents. In J. Bradshaw (ed) *Software Agents.* Menlo Park, CA: AAAI Press. Cambridge, MA, London, England: MIT Press.

Brand, S. (1988). *The Media Lab: Inventing the Future at MIT.* New York: Viking Penguin. ISBN 0-14009701-5.

Bricken, M. (XXXX). *Virtual Worlds: No Interface to Design.* Technical Report R-90-2, Washington Technology Center, WA.

Bristow, G. (1986). *Electronic Speech Recognition: Techniques, Technology and Applications.* London: Collins.

Burgoon, J., Bonito, J., Bengtsson, B., Cederberg, C., Lundeberg M., and Allspach,.L. (2000). Interactivity in Human-Computer Interaction: a Study of Credibility, Understanding, and Influence. In *Computers in Human Behavior*, Vol. 16, 553-574.

Cañamero, L. & Fredslund, J. (2001). How Does it Feel? Emotional Interaction with a Humanoid LEGO Robot. Proceedings of *Socially Intelligent Agents: The Human in the Loop*, AAAI Fall Symposium, Nov 2-5, 23-8. Menlo Park, CA: AAAI Press.

Carr, K & England, R. (1995). *Simulated and Virtual Realities: Elements of Perception.* London: Taylor & Francis. ISBN 074-84012-96.

Cassell, J. & Bickmore, T. (2001). *External Manifestations of Trustworthiness in the Interface.* In Communications of the ACM. Vol. 43(12).

Cassell, J., Bickmore, T., Vilhjalmsson, H., and Yan, H. (2001). More Than Just Another Pretty Face: Affordances of Embodiment. *Knowledge Based Systems.*

Cassell, J. (2000). Nudge, Nudge, Wink, Wink. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents.* Cambridge, Mass., London, MIT Press, 2000. ISBN 0-262-03278-3.

Cassell, J., Thorisson, K. (1999). The Power of a Nod and a Glance: Envelope v's Emotional Feedback in Animated Conversational Agents. In Journal of *Applied Intelligence*, Vol. 13 (3), 519-538.

Churchill, E., Cook, L., Hodgson, P., Prevost, S., and Sullivan, J. (2000). May I Help You?: Desiging Embodied Conversational Agent Allies. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents.* Cambridge, Mass., London, MIT Press, 2000. ISBN 0-262-03278-3.

Chittaro, L. & Ranaonm R. (2000). Virtual Reality Stores for One-to-One ECommerce. Proceedings of *SIGCHI Workshop on Designing Interactive Systems for One-to-One Ecommerce.*

Clark, H. (1992). *Arenas of Natural Language Use.* Chicago, Illinois: University of Chicago Press.

Clark, H. & Brennan, S. (1990). Grounding in Communication. In L. Resnick, J. Levine & S. Bahrend (eds.). *Perspectives on Socially Shared Cognition*, 127-149. American Psychological Association.

Cohen, P. & Oviatt, S. (1995). The Role of Voice Input for Human-Machine Communication. Proceedings of the *National Academy of Sciences*. Vol. 92(22), 9921-27.

Cole, R. (1999). Tools for Research and Education in Speech Science. In Proceedings of the *International Conference of Phonetic Sciences*, San Francisco, CA. http://cslu.cse.ogi.edu/toolkit

Coolican, H. (1994). *Research Methods and Statistics in Psychology.* London: Hodder & Stoughton. ISBN 0-340600-829.

Corritore, C., Kracher, B. & Wiedenbeck, S. (2000). Working Paper. Creighton University, Omaha, NE.

Coyne, R. (1999). *Technoromanticism.* Cambridge, MA: MIT Press. ISBN 0-26203260-0.

Cuddihy, E. & Walters, D. (2000). Embodied Interaction in Social Virtual Environments. Proceedings of *Collaborative Virtual Environments*, San Francisco, CA. ACM Press.

Dautenhahn, K., (2001) Socially Intelligent Agents: The Human in the Loop. In *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.*

Dehn, D., and Van Mulken, S. (2000). The Impact of Animated Interface Research:A Review of Empirical Research. In *International Journal of Human-Computer Studies*, Vol. 52(1),1-22.

Deepmatrix, (2001). http://www.deepmatrix.com

Dix, A. (1998). *Human Computer Interaction.* London: Prentice Hall Europe. ISBN 0-132398-648.

ECommerce Trust Study. (1999). *Cheskin Research and Studio Arhetype/Sapient.* http://www.studioarchetype.com/cheskin

Ekman, P. & Friesen, W. (1978). *Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. & Friesen, W. (1975). *Unmasking the Face.* New Jersey: Prentice-Hall.

Ekman, P. & Friesen, W. (1969). The Repertoire of Non-Verbal Behaviour: Categories, Origins, Usage, and Coding. In *Semiotica*, 1, 49-98.

Elliot, C. (1992). *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System.* Ph.D. Thesis. Institute for the Learning Sciences, Northwestern University.

Ezzat, T., and Poggio, T. (1998). MikeTalk: A Talking Facial Display Based on Morphing Visemes. Proceedings of *Computer Animation Conference*. June, 96-102.

Flanagan, J. (1995). Multimodality. In G. Varile & A. Zampolli (eds.) *Survey of the State of the Art in Human Language Technology*. National Science Foundation.

Fogg. B. & Tseng. H. (1999). The Elements of Computer Credibility. Proceedings of *SIGCHI*, Pittburgh, PA, May ,15-20.

Gould, J. (1988). How to Design Usable Systems. In M. Helander (ed.) *Handbook of Human-Computer Interaction*. Amsterdam: Elsevier Science Publishers.

Granström, B. (1999). Multi-modal Speech Synthesis with Applications. In G. Chollet, M. Di Benedetto, A. Esposito & M. Marinaro (eds.) *Speech Processing, Recognition and Artificial Neural Networks*. Berlin: Springer.

Gratch, J. & Marsella, S. (2001). Tears and Fears: Modeling Emotions and Emotional Behaviors in Synthetic Agents. Proceedings of the *5th International Conference on Autonomous Agents*, Montreal, 278-85. ACM Press. ISBN 1-58113-326-X.

Guye-Vuillieme, A., Capin, T., Pandzic, I., Magennet-Thalmann, N., and Thalmann, D. (1999). Non-verbal Communication Interface for Collaborative Virtual Environments. In.*The Virtual Reality Journal*, Vol. 4, 49-59.

Hagen, E. (2000). A Flexible Spoken Dialogue Manager. Proceedings of the *3rd International Workshop on Human-Computer Conversation*, July, 68-73.

Hauptmann, A. (1989). Speech and Gestures for Graphic Image Manipulation. Proceedings of *SIGCHI*, May, 241-5.

Hayes-Roth, B. (2001). Characters Everywhere. Seminar on *People, Computers and Design*, Stanford University, March, 2001.

Humanoid Animation Specification, (2001). http://www.h-anim.org

Isbister, K. (1998). *Reading Personality in Onscreen Interactive Characters: An Examination of Social Psychological Principles of Consistency, Personality Match, and Situational Attributes Applied to Interaction Characters*. Ph.D. Thesis. Communication Department, Stanford University, CA.

Johnson, G. (1987). *Machinery of the Mind*. Redmond, WN: Tempus. ISBN 1-55615010-5.

Johnson, L., and Rickel, J. (2000). Animated Pedagogical Agents: Face-to-face Interaction in Interactive Learning Environments. In *International Journal of Artificial Intelligence in Education*.

Johnson, L. (2000). Socially Intelligent Agent Research at CARTE. Proceedings of *Socially Intelligent Agents: The Human in the Loop*, AAAI Fall Symposium, Nov 2-5, 77-82. Menlo Park, CA: AAAI Press.

Jording, T. & Michel, S. (1999). Personalised Shopping in the Web by Monitoring the Customer. Proceedings of the Active Web Conference. Stafford, UK.

Karvonen, K. (1999). Creating Trust. Proceedings of the *4th Nordic Workshop on Secure IT Systems*, Nov 1-2, Sweden, 21-36.

Kendon, A. (1988). Goffman's Approach to Face-to-Face Interaction. In P. Drew & A. Wootton (eds.) *Erving Goffman: Exploring the Interaction Order.* Cambridge: Polity.

Kim, J. & Moon, J. (1997). Emotional Usability of Customer Interfaces: Focusing on Cyber Banking System Interfaces. Proceedings of *SIGCHI.*

King, J., and Ohya, J. (1996). The Representation of Agents: Anthropomorphism, Agency and Intelligence. Proceedings of *SIGCHI Conference Companion on Human Factors in Computing Systems: Common Ground.* April, 289-90, ACM Press.

Kipp, M. (2001). From Human Gesture to Synthetic Action. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents*, 5th International Conference on Autonomous Agents, 9-14.

Koda, T. (1996). *A Study on the Effects of Personification of Software Agents.* M.Sc. diss., Massachusetts Institute of Technology, Cambridge, Massachusetts,1996.

Krueger, R. (1994). *Focus Groups: A Practical Guide for Applied Research.* California: Sage. ISBN 0-803955-669.

Lanier, J. (1995). *Agents of Alienation.* ACM Interactions, July, 66-72.

Laurel, B. (1993). *Computers as Theatre.* Reading, MA: Addison-Wesley. ISBN 0-201-55060-1.

Laurel, B. (1990). Interface Agents: Metaphors with Character. In B. Laurel (ed.) *The Art of Human-Computer Interface Design*, 335-365. Reading, MA: Addison-Wesley Publishing Co.

Lester, J., and Stone, B. (1997). Increasing Believability in Animated Pedagogical Agents. Proceedings of the *1st International Conference on Autonomous Agents*, Feb. 6-21, ACM Press.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., and Bhogal, R. (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. Proceedings of *Human Factors in Computing Systems, CHI '97*, March, 359-366, ACM Press.

Likert, R. (1932). A Technique for the Measurement of Attitudes. In *Archives of Psychology 140*, 55, 1932.

Lindsay, D. (1997). Talking Head. In *Invention and Technology*, 57-63.

Lombard, M. & Ditton, T. (1997). At the Heart of It All: The Concept of Presence. *Journal of Computer Mediated Communication* 3(2). http://www.ascusc.org/jcmc/vol3/issue2/

Maes, P., Darrell, T., Blumberg, B. & Pentland A. (1995). The ALIVE System: Full-body Interaction with Autonomous Agents. In *IEEE Computer, Special Issue on Virtual Environments*, 11-18.

Maes, P., Darrell, T., Blumberg, B. & Pentland A. (1994). Interacting with Animated Autonomous Agents. *AAAI Spring Symposium on Believable Agents Working Notes*, Stanford University, California, March 19-20, 50-52.

Magnenat-Thalmann, N. & Kshiragar, S. (2000). Communication with Autonomous Agents. Proceedings of *Cele-Twente Workshop on Language Technology: Learning to Behave*, 1-8.

Magnenat-Thalmann, N., Kalra, P., and Escher, M. (1998). Face to Virtual Face. *Proceedings of the IEEE.* Vol. 86, No.5, May 1998.

Marsella, S., Gratch J. & Rickel, J. (2001). The Effect of Affect: Modeling the Impact of Emotional State on the Behaviour of Interactive Virtual Humans. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents,* 5th International Conference on Autonomous Agents, 47-52.

Massaro, D., Cohen, M., Beshow, J. & Cole, R. (2000). Designing and Evaluating Conversational Interfaces with Animated Adams. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (eds.) *Embodied Conversational Agents.* Cambridge, Mass., London, MIT Press, 2000. ISBN 0-26203278-3.

Massaro, D. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle.* Cambridge, Mass., London, MIT Press, 1998. ISBN 0-262-13337-7.

Maybury, M., (1998). Toward Co-operative Multimedia Interaction. In *Lecture Notes in Artificial Intelligence 1374, Multi-modal Human-Computer Communication*, 1998.

McBreen, H., and Jack, M. (2001). Evaluating Humanoid Synthetic Agents in E-Retail Applications. To appear in *IEEE Transactions on Systems, Man and Cybernetics - Special Issue on Socially Intelligent Agents: The Human in the Loop*, ed. K. Dautenhahn.

McBreen, H. (2001). Embodied Conversational Agents in ECommerce Applications. In K.Dautenhahn, A. Bond, D. Canamero & B. Edmonds (eds.) *Socially Intelligent Agents - Creating Relationships with Computers and Robots.* Kluwer Publications.

McBreen, H., Anderson, J. & Jack, M. (2001). Evaluating 3D Embodied Conversational Agents in Contrasting VRML Retail Applications. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents*, 5th International Conference on Autonomous Agents, 83-8.

McBreen, H., and Jack, M. (2000). Empirical Evaluation of Animated Agents In a Multi-Modal Retail Application. Proceedings of the *AAAI Fall Symposium: Socially Intelligent Agents – The Human in the Loop*, Nov. 122-126. ISBN 1-57735-127-4.

McBreen, H., and Jack, M. (2000). Animated Conversational Agents in ECommerce Applications'. Proceedings of the *3rd Workshop on Human-Computer Conversation*, 112-117.

McBreen, H., Shade, P., Jack, M., and Wyard, P. (2000). Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications. Proceedings

of the *4th International Conference on Autonomous Agents*, June, 39-45, ACM Press. ISBN 1-581-13230-1.

McGurk, H (1999). Hearing Lips and Seeing Voices. In *Talking Heads*. Haskins Laboratories. URL: www.haskins.yale.edu/haskins/heads

McGurk, H (1981). Listening with Eye and Ear. In T. Myers, J. Laver, J. Anderson (eds.), *The Cognitive Representation of Speech*. Amsterdam: North Holland.

McInnes, F., Attwater, D., Edington, M., Schmidt, M. & Jack M. (1999). User Attitudes to Natural Speech and Text-to-Speech Synthesis in an Automated Information Service. Proceedings of the *6th European Conference on Speech Communication and Technology (Eurospeech)*, 831-34.

McKnight, H. & Chervany, N. (1996). *The Meaning of Trust*. Working Paper 96-04. Carlson School of Management, University of Minnesota, MN.

McNeill, D. (1997). Growth Points Cross-Linguistically. In Nuyts & Pederson (eds.) *Language and Conceptualisation*. Cambridge: Cambridge University Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Meech, J. & Marsh, S. (2000). Social Factors in ECommerce Personalisation. Proceedings of *SIGCHI Workshop on Designing Interactive Systems for One-to-One ECommerce*.

Minsky, M. (1985). *The Society of Mind*. New York: Simon and Schuster.

Molich, R. & Nielsen. (1990). Heuristic Evaluation of User Interfaces. Proceedings of *SIGCHI*, Seattle, Washington, 249-256. ACM Press.

Monk, A. ed. (1985). *Fundamentals of Human Computer Interaction*. London: Academic Press.

Nagao, K, & Takeuchi, A. (1994). Social Interaction: Multi-modal Conversation with Social Agents. Proceedings of the *12th National Conference on Artificial Intelligence*, Vol. 1, 22-28.

Nass, C., Isbister, K., and Lee, E. (2000). Truth is Beauty: Researching Embodied Conversational Agents. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents*. Cambridge, Mass., London, MIT Press, ISBN 0-26203278-3.

Nass, C., Moon, Y., Morkes, J., Kim, E., and Fogg, B. (1997). Computers as Social Actors. In B. Friedman (ed.) *Human Values and the Design of Computer Technology*, CSLI Publications, ISBN 1-57586-081-3.

Nass, C., Steuer, J. & Tauber, E. (1994). Computers are Social Actors. Proceedings of *SIGCHI '94*, Boston, MA, April, 72-28.

Negroponte, N. (1990). Hospital Corners. In B. Laurel (ed.) *The Art of Human-Computer Interface Design*, 347-353. Reading, MA: Addison-Wesley Publishing Co.

Nielsen, J. (1993). *Usability Engineering*. San Diego, CA: Academic Press.

Nikander, P. & Karvonen, K. (2000). Users and Trust in Cyberspace. Proceedings of *Cambridge Security Protocols Workshop*. Springer.

Norman, D. (1998). *The Invisible Computer*. Cambridge, Mass., London, England, MIT Press. ISBN 0-262-14065-9.

Norman, D. (1997). How People Might Interact with Agents. In J. Bradshaw (ed.) *Software Agents*. Menlo Park, CA: AAAI Press. Cambridge, MA, London, England: MIT Press.

Ostermann, J., and Millen, D. (2000). Talking Heads and Synthetic Speech: An Architecture for Supporting Electronic Commerce. Proceedings of *IEEE International Conference On Multimedia and Expo* (ICME), 2000.

Ortony, A., Clore, G. & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge, MA: Cambridge University Press.

Pagano, R. (1990). *Understanding Statistics in the Behavioural Sciences*. St. Paul: West. ISBN 0-314667-920.

Parke, F., and Waters, K. (1996). *Computer Facial Animation*. A.K. Peters, Wellesley, 1996. ISBN 1-568-81014-8.

Pelachaud, C. & Poggi, I. (2001). Multi-modal Believable Agents. Proceedings of the *Workshop on Multi-modal Communication and Context in Embodied Agents,* 5th International Conference on Autonomous Agents, 95-9.

Pelachaud, C., Badler, N., & Steedman, M. (!996). Generating Facial Expression for Speech. In *Cognitive Science*, 20(1), 1-46.

Person, N. (2000). AutoTutor's Conversational Behaviors. Proceedings of *3rd International Workshop on Human-Computer Conversation*, July 3-5, 130-35.

Picard, R. (1997). *Affective Computing*. Cambridge, MA, London, England: MIT Press. ISBN 0-262-16170-2.

Prendinger, H. (2001). Smart Brains in Simplistic Bodies. Proceedings of the *Workshop on Multimodal Communication and Context in Embodied Agents,* 5th International Conference on Autonomous Agents, 41-6.

Ratner, P. (1998). *3D Human Modeling and Animation*. New York: John Wiley & Sons. ISBN 0-471-29229-X.

Reeves, B., and Nass, C. (1996). *The Media Equation*. Stanford University, California. CSLI Publications, 1996. ISBN 1-575-86053-8.

Reilly, S.N. (1996). *Believable Social and Emotional Agents.* Ph.D Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Rickel, J., and Johnson, L. (2000). Task Oriented Dialogs with Animated Agents in Virtual Reality. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents.* Cambridge, MA, London, England, MIT Press. ISBN 0-262-03278-3.

Rist, T. (2001). Boosting E-Services Through Animated Interface Agents. Tutorial Notes of *5th International Conference on Autonomous Agents*, May 28-June 01.

Ruben, P. & Vatikiotis-Bateson, E. (1998). *Talking Heads.* Haskins Laboratories. www.haskins.yale.edu/haskins/heads

Rust, J., and Golomok, S. (1989). *Modern Psychometrics: The Science of Modern Psychological Assessment. London*: Routledge. ISBN 0-415030-595.

Ruttkay, Z. & Hoot, H. (2001). FESINC: Facial Expression Sculpturing with Interval Constraints. Proceedings of the *Workshop on Multimodal Communication and Context in Embodied Agents*, 5th International Conference on Autonomous Agents, 71-6.

Sacks, H. (1992). *Lectures on Conversation*, Vol. I & II. Cambridge, MA: Blackwell.

Sacks, H. (1974). A Simplest Systematics For the Organisation of Turn Taking for Conversation. In *Language*, Vol. 50, 696.

Sanders, G., and Scholtz, J. (2000). Measurement and Evaluation of Embodied Conversational Agents. In J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) *Embodied Conversational Agents.* Cambridge, MA, London, England, MIT Press. ISBN 0-262-03278-3.

Scassellati, B. (2000). Theory of Mind for a Robot. Proceedings *AAAI Fall Symposium: Socially Intelligent Agents – The Human in the Loop*, Nov 2-5, 164-68. ISBN 1-57735-127-4.

Scherer, K., and Ekman, P. (1982). *Handbook of Methods in Nonverbal Behavior Research.* Cambridge University Press, 1982.

Shadbolt, N. (1989). Planning and Discourse. In M. Taylor, F. Néel, & D. Bouwhuis (eds.) *The Structure of Multimodal Dialogue.* New York: North Holland Press.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Reading, MA: Addison-Wesley.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the Interface is a Face. In Journal of *Human Computer Interaction*, 11, 97-124.

Stork, D. (1997). Scientist on the Set: An Interview with Marvin Minsky. In D. Stork (ed.) *HAL's Legacy: 2001's Computer as Dream & Reality.* Cambridge, MA; London: England: MIT Press. ISBN 0-262-19378-7.

Takeuchi, A., and Nagao, K. (1993). Communicative Facial Displays as a New Conversational Modality. Proceedings of *InterCHI: Human Factors in Computing Systems.* April, 187-193.

Thalmann, D. (2000). Autonomous Virtual Humans in Virtual Environments. Tutorial 6 of the *4th International Conference on Autonomous Agents*, June, ACM Press.

Thomas, F. & Johnston, O. (1981). *Disney Animation – The Illusion of Life.* New York, NY: Abbeville Press.

Thorisson, K. (1996). *Communicative Humanoids: Model of Psychosocial Dialogue Skills.* Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996.

Trappl, R. & Petta, P. (1997). *Creating Personalities for Synthetic Actors.* Berlin: Springer-Verlag. ISBN 3-540-62735-9.

Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet.* New York: Simon and Schuster.

Turkle, S. (1984). *The Second Self: Computers and the Human Spirit.* London: Granada.

Van Mulken, S., André, E., and Müller, J. (1999). An Empirical Study on the Trustworthiness of Lifelike Interface Agents. In *Human Computer Interaction: Communication, Cooperation and Application Design*, 152-156.

Van Mulken, S., André, E., and Muller, J. (1998). The Persona Effect: How Substantial Is It? Proceedings of the *Human Computer Interaction Conference*, 53-66, Berlin, Springer.

Walker, H., Sproull, L., and Subramani, R. (1994). Using a Human Face in an Interface. Proceedings of *Human Factors in Computing Systems: CHI*, April, 85-91.

Walker, M. (1989). Natural Language in a Desktop Environment. Proceedings of *SIGCHI '98.* ACM Press.

Waters, K., and Levergood, L. (1995). DECface: A System for Synthetic Face Applications. In *Multimedia Tools and Applications*, 1, 349-366, Kluwer Academic Publishers, 1995.

Web Consortium, (2001). http://www.w3c.org

Whittaker, S. & Walker, M, (1991). Towards a Theory of Multi-modal Interaction. In *AAAI '91 Workshop Notes*, 78-85.

Wyard, P., and Churcher, G. (1999). The MUeSLI Multimodal 3D Retail System. Proceedings *ESCA Workshop on Interactive Dialogue Systems.*

Yngve, V. (1970). On Getting a Word in Edgewise. Papers from the Sixth Regional Meeting. Chicago Linguistics Society, 567-78.

Zue, V. & Cole R. (1995). Spoken Language Input. In G. Varile & A. Zampolli (eds.) *Survey of the State of the Art in Human Language Technology*. National Science Foundation.

Appendix 1

Appendix 1.1

This section details the information sheets that were presented to each of the participants before beginning the evaluation described in Chapter 4.

- Task Sheet 1: Used in Part I of Evaluation (Home Furnishings Service)
- Task Sheet 2: Used in Part II of Evaluation (Personalised CD Service)

Task Sheet 1

Introduction

You are going to hear a customer using an interactive Home Furnishings Service which features an automated shop assistant. The customer is trying to decorate a living room, and the assistant is offering advice to the customer.

You will see the display that the customer sees, and hear the conversation between the customer and the assistant, as if you were looking over the customer's shoulder. You will not see the customer. Some versions of the service have the assistant visible on the screen, and others do not.

Instructions

Your task is to listen to and observe the assistant carefully. Also observe the changes that will be made to the living room. You will be asked to fill out a short questionnaire about your views of the assistant and the service.

Task Sheet 2

Introduction

You are going to hear a customer using an interactive CD service which features an automated shop assistant. The customer is trying to create a CD, and the assistant is offering advice to the customer.

You will see the display that the customer sees, and hear the conversation between the customer and the assistant, as if you were looking over the customer's shoulder. You will not see the customer. Some versions of the service have the assistant visible on the screen, and others do not.

Instructions

Your task is to listen to and observe the assistant carefully. Also observe the changes that will be made to the CD. You will be asked to fill out a short questionnaire about your views of the assistant and the service.

Appendix 1.2

The actual questionnaires used in the evaluation described in Chapter 4 are presented here.

- Questionnaire 1: Corresponds to Agent Types H1, H2 and H3
- Questionnaire 2: Corresponds to Agent Type H4
- Questionnaire 3: Corresponds to Agent Type H5

Questionnaire 1

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I liked the appearance of the assistant.	☐	☐	☐	☐	☐	☐	☐
2. The assistant's voice was not clear	☐	☐	☐	☐	☐	☐	☐
3. I think this service is a good idea.	☐	☐	☐	☐	☐	☐	☐
4. I think this service would be difficult for me to use.	☐	☐	☐	☐	☐	☐	☐
5. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
6. I felt the assistant seemed unfriendly.	☐	☐	☐	☐	☐	☐	☐
7. I felt the assistant was competent.	☐	☐	☐	☐	☐	☐	☐
8. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐
9. I noticed the lips moving.	☐	☐	☐	☐	☐	☐	☐
10. I would like to use this service myself.	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
11. I felt the speech sometimes didn't match the lips.	☐	☐	☐	☐	☐	☐	☐
12. I thought the assistant looked natural	☐	☐	☐	☐	☐	☐	☐
13. I though being able to see the assistant was helpful.	☐	☐	☐	☐	☐	☐	☐
14. The appearance of the assistant was unsuitable for the service.	☐	☐	☐	☐	☐	☐	☐

Comments

Questionnaire 2

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I liked the appearance of the assistant.	☐	☐	☐	☐	☐	☐	☐
2. The assistant's voice was not clear.	☐	☐	☐	☐	☐	☐	☐
3. I think this service is a good idea.	☐	☐	☐	☐	☐	☐	☐
4. I think this service would be difficult for me to use.	☐	☐	☐	☐	☐	☐	☐
5. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
6. I felt the assistant seemed unfriendly.	☐	☐	☐	☐	☐	☐	☐
7. I felt the assistant was competent.	☐	☐	☐	☐	☐	☐	☐
8. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐
9. I would like to use this service myself.	☐	☐	☐	☐	☐	☐	☐
10. I thought the assistant looked natural	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
11. I thought being able to see the assistant was helpful.	☐	☐	☐	☐	☐	☐	☐
12. The appearance of the assistant was unsuitable for the service.	☐	☐	☐	☐	☐	☐	☐

Comments

Questionnaire 3

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐
2. The assistant's voice was not clear.	☐	☐	☐	☐	☐	☐	☐
3. I think this service is a good idea.	☐	☐	☐	☐	☐	☐	☐
4. I think this service would be difficult for me to use.	☐	☐	☐	☐	☐	☐	☐
5. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
6. I felt the assistant seemed unfriendly.	☐	☐	☐	☐	☐	☐	☐
7. I felt the assistant was competent.	☐	☐	☐	☐	☐	☐	☐
8. I would like to use this service myself.	☐	☐	☐	☐	☐	☐	☐

Comments

Appendix 2

Appendix 2.1

This section details the information sheet that was presented to each of the participants before beginning the evaluation described in Chapter 5.

Task Sheet 1

Introduction

You are going to hear a customer using an interactive Home Furnishings Service that features an automated shop assistant. The customer is trying to decorate a living room, and the assistant is offering advice to the customer.

You will see the display that the customer sees, and hear the conversation between the customer and the assistant, as if you were looking over the customer's shoulder. You will not see the customer. Some versions of the service have the assistant visible on the screen, and others do not.

Instructions

Your task is to listen to and observe the assistant carefully. Also observe the changes that will be made to the living room. You will be asked to fill out a short questionnaire about your views of the assistant and the service.

Appendix 2.2

The actual questionnaires used in the evaluation described in Chapter 5 are presented here.

- Questionnaire 1: Corresponds to Agent Types C4, C5, C6,
- Questionnaire 2: Corresponds to Agent Type C2, C3
- Questionnaire 3: Corresponds to Agent Type C1

Questionnaire 1

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. Being able to see the assistant was helpful.	☐	☐	☐	☐	☐	☐	☐
2. I liked the gestures the assistant made.	☐	☐	☐	☐	☐	☐	☐
3. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
4. I liked the appearance of the assistant.	☐	☐	☐	☐	☐	☐	☐
5. The appearance of the assistant was unsuitable for the Home Furnishings application.	☐	☐	☐	☐	☐	☐	☐
6. The assistant was polite.	☐	☐	☐	☐	☐	☐	☐
7. The facial expressions made the assistant appear lifelike.	☐	☐	☐	☐	☐	☐	☐
8. The gestures made the assistant appear unfriendly.	☐	☐	☐	☐	☐	☐	☐
9. The facial expressions made the assistant appear unhelpful.	☐	☐	☐	☐	☐	☐	☐
10. The gestures made the assistant appear unhelpful.	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
11. The assistant was competent.	☐	☐	☐	☐	☐	☐	☐
12. The facial expressions made the assistant appear unfriendly.	☐	☐	☐	☐	☐	☐	☐
13. The assistant's voice was natural.	☐	☐	☐	☐	☐	☐	☐
14. The gestures made the assistant appear lifelike.	☐	☐	☐	☐	☐	☐	☐
15. The assistant's voice was annoying.	☐	☐	☐	☐	☐	☐	☐
16. The lip movement was distracting.	☐	☐	☐	☐	☐	☐	☐
17. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐
18. The facial expressions made the assistant appear lifelike.	☐	☐	☐	☐	☐	☐	☐

Comments

Questionnaire 2

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. Being able to see the assistant was helpful.	☐	☐	☐	☐	☐	☐	☐
2. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
3. I liked the appearance of the assistant.	☐	☐	☐	☐	☐	☐	☐
4. The appearance of the assistant was unsuitable for the Home Furnishings application.	☐	☐	☐	☐	☐	☐	☐
5. The assistant was polite.	☐	☐	☐	☐	☐	☐	☐
6. The facial expressions made the assistant appear helpful.	☐	☐	☐	☐	☐	☐	☐
7. The facial expressions made the assistant appear unhelpful.	☐	☐	☐	☐	☐	☐	☐
8. The assistant was competent.	☐	☐	☐	☐	☐	☐	☐
9. The facial expressions made the assistant appear unfriendly.	☐	☐	☐	☐	☐	☐	☐
10. The assistant's voice was natural.	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
11. The assistant's voice was annoying.	☐	☐	☐	☐	☐	☐	☐
12. The lip movement was distracting.	☐	☐	☐	☐	☐	☐	☐
13. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐
14. The facial expressions made the assistant appear lifelike.	☐	☐	☐	☐	☐	☐	☐

Comments

Questionnaire 3

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
2. The assistant was polite.	☐	☐	☐	☐	☐	☐	☐
3. I felt the assistant seemed unfriendly.	☐	☐	☐	☐	☐	☐	☐
4. I felt the assistant was competent.	☐	☐	☐	☐	☐	☐	☐
5. The assistant's voice was unnatural.	☐	☐	☐	☐	☐	☐	☐
6. The assistant's voice was annoying.	☐	☐	☐	☐	☐	☐	☐
7. I felt the conversation was unnatural.	☐	☐	☐	☐	☐	☐	☐

Comments

Appendix 3

Appendix 3.1

This section details the information sheet that was presented to each of the participants before beginning the evaluation described in Chapters 6 and 7.

Task Sheet 1

Introduction

You are at the CCIR Cinema Box-Office, which can be accessed over the Internet. You will be asked to read your task first and then you will be asked to converse with an automated assistant to book cinema tickets.

Instructions

Please observe the assistant and the service carefully. During the conversation you may be asked for security number information, which is on the desk in front of you. After the conversation you will be asked to complete a questionnaire about your views of the assistant. You will also be asked to complete a short questionnaire about the service.

Task Sheet 2

Introduction

You are at the CCIR Travel Agency, which can be accessed over the Internet. You will be asked to read your task first and then you will be asked to converse with an automated assistant to book flights.

Instructions

Please observe the assistant and the service carefully. During the conversation you may be asked for security number information, which is on the desk in front of you. After the conversation you will be asked to complete a questionnaire about your views of the assistant. You will also be asked to complete a short questionnaire about the service.

Task Sheet 3

Introduction

You are at the CCIR Bank, which can be accessed over the Internet. You will be asked to read your task first and then you will be asked to converse with an automated assistant to transfer money.

Instructions

Please observe the assistant and the service carefully. During the conversation you may be asked for security number information, which is on the desk in front of you. After the conversation you will be asked to complete a questionnaire about your views of the assistant. You will also be asked to complete a short questionnaire about the service.

Appendix 3.2

The actual questionnaires used in the evaluation described in Chapter 6 are presented here.

- Questionnaire 1: Corresponds to Applications
- Questionnaire 2: Application Comparisons
- Questionnaire 3: Corresponds to Agents

Questionnaire 1

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I would use this service myself	☐	☐	☐	☐	☐	☐	☐
2. I felt the service was difficult to use.	☐	☐	☐	☐	☐	☐	☐
3. I do not think this service is a good idea.	☐	☐	☐	☐	☐	☐	☐
4. I think this service is convenient.	☐	☐	☐	☐	☐	☐	☐

Please rate your experience of this service on the following scale (1 = low, 10 = high)

1 2 3 4 5 6 7 8 9 10

Please write down any comments about the service:

Questionnaire 2

Please complete this questionnaire. For the question below, tick one box that best expresses your opinion.

Overall, which service did you prefer?

a) I preferred the cinema service ☐

b) I preferred the travel agency service ☐

c) I preferred the banking service ☐

d) I did not like any of the services ☐

e) I liked all the services equally ☐

Please give reasons for your choice...

Questionnaire 3

Please complete this questionnaire. For the question below, tick one box that best expresses your opinion.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. The assistant was polite.	☐	☐	☐	☐	☐	☐	☐
2. I did not like the appearance of the assistant.	☐	☐	☐	☐	☐	☐	☐
3. I liked the assistant's voice.	☐	☐	☐	☐	☐	☐	☐
4. The assistant was unfriendly.	☐	☐	☐	☐	☐	☐	☐
5. The assistant was competent.	☐	☐	☐	☐	☐	☐	☐
6. The assistant's voice was annoying.	☐	☐	☐	☐	☐	☐	☐
7. I felt the assistant understood me.	☐	☐	☐	☐	☐	☐	☐
8. I did not like speaking to the assistant .	☐	☐	☐	☐	☐	☐	☐
9. The assistant spoke naturally.	☐	☐	☐	☐	☐	☐	☐
10. The assistant was trustworthy.	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
11. The assistant was dressed inappropriately for this service.	☐	☐	☐	☐	☐	☐	☐
12. The assistant was cheerful.	☐	☐	☐	☐	☐	☐	☐
13. The assistant was unsociable.	☐	☐	☐	☐	☐	☐	☐
14. The assistant appeared lifelike.	☐	☐	☐	☐	☐	☐	☐
15. The assistant was agreeable.	☐	☐	☐	☐	☐	☐	☐

Please rate the assistant on the following scale (1 = low, 10 = high)

1 2 3 4 5 6 7 8 9 10

Please write down any comments about the assistant:

Appendix 4

Appendix 4.1

This section details the information sheet that was presented to each of the participants before beginning the evaluation described in Chapters 6 and 7.

Questionnaire 1

Please complete this questionnaire. For each statement below, tick the box that best expresses your opinion of that statement.

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
1. I did not enjoy speaking to the assistant.	☐	☐	☐	☐	☐	☐	☐
2. I felt the assistant was trustworthy.	☐	☐	☐	☐	☐	☐	☐
3. The service was not reliable.	☐	☐	☐	☐	☐	☐	☐
4. I could depend on the assistant to the job.	☐	☐	☐	☐	☐	☐	☐
5. This service is not a good idea.	☐	☐	☐	☐	☐	☐	☐
6. The service was efficient.	☐	☐	☐	☐	☐	☐	☐
7. I felt confident the assistant understood me.	☐	☐	☐	☐	☐	☐	☐

	strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree
8. The assistant was credible.	☐	☐	☐	☐	☐	☐	☐
9. I had no confidence in the service.	☐	☐	☐	☐	☐	☐	☐
10. The assistant was competent.	☐	☐	☐	☐	☐	☐	☐
11. I did not think this service was useful.	☐	☐	☐	☐	☐	☐	☐
12. I felt in control using this service.	☐	☐	☐	☐	☐	☐	☐
13. I did not think the assistant was reliable.	☐	☐	☐	☐	☐	☐	☐

Comments

Publications

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).

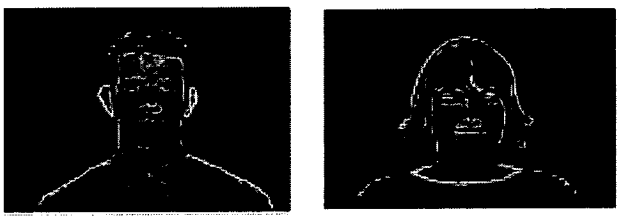


Fig. 1. Photorealistic images of male and female human-like agents.

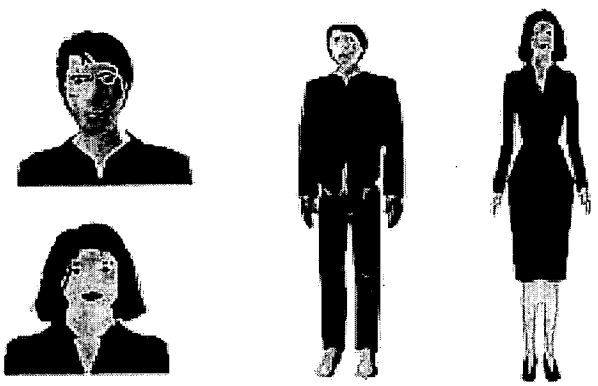


Fig. 2. Two-dimensional heads and embodied cartoon-like agents.

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.

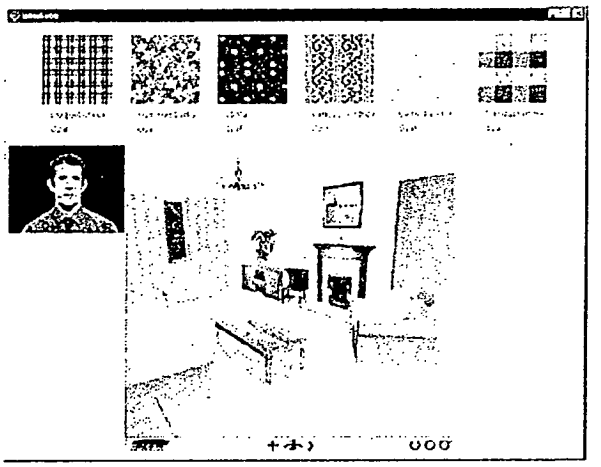


Fig. 3. Interface for application 1: Home furnishings.

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]

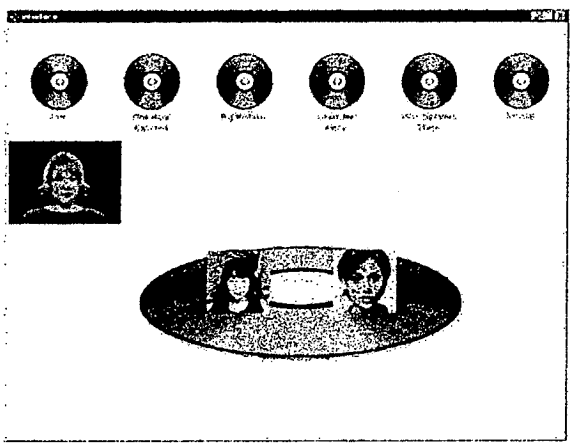


Fig. 4. Interface for application 2: CD service.

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 2×5×2 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 × 5 × 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 × 5 × 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 × 5 × 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 × 5 × 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 × 5 × 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 × 5 × 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 × 5 × 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 × 5 × 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 × 4 × 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 $2 \times 4 \times 2$ 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 $2 \times 3 \times 2$ 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.

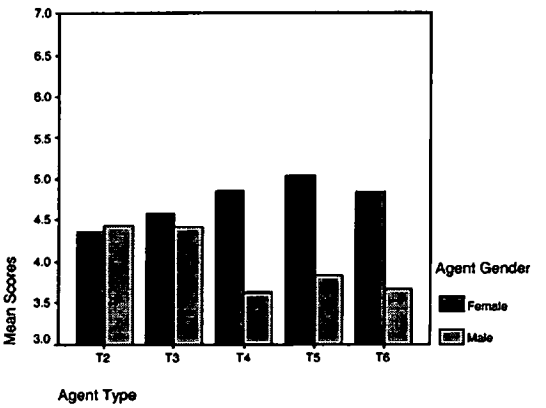


Fig. 5. Mean scores for cartoon-like agent type and gender for questionnaire item: "I liked the appearance of the assistant."

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*

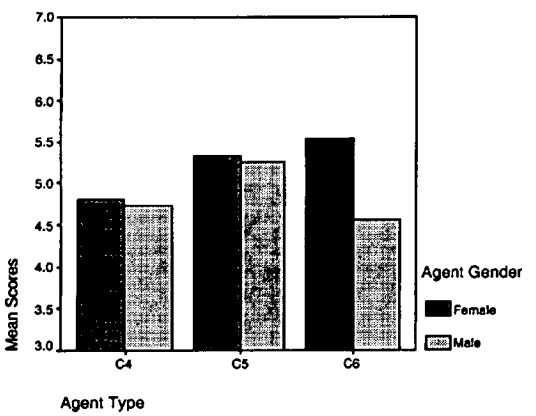


Fig. 6. Mean scores for cartoon-like agent type and gender for questionnaire item: "The gestures made the assistant appear friendly."

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.

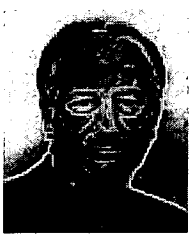
[7] T. Ezzat and T. Poggio, "MikeTalk: A talking facial display based on morphing visemes," in *Proc. Comput. Animation Conf.*, June 1998, pp. 96–102.

[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.

Chapter 1

EMBODIED CONVERSATIONAL AGENTS IN E-COMMERCE APPLICATIONS

Helen McBreen
Centre for Communication Interface Research, The University of Edinburgh

Abstract This section discusses an empirical evaluation of 3D embodied conversational agents, in three interactive VRML e-commerce environments: a cinema box-office, a travel agency and a bank. Results showed participants enjoyed speaking to the agents in the applications and expressed a desire for agents in the cinema application to be informally dressed but those in the banking application to be formally dressed. Qualitative results suggested that participants found it difficult to assign a degree of trust to the agents in the banking application.

1. INTRODUCTION

The emerging interest in embodied conversational agents (ECA's) coupled with the growing evidence [1, 3, 4, 6, 9, 14] that embodiment can enhance user interface design has fuelled a challenging research agenda and developing embodied agents that behave socially in an interaction has become the principal goal for many interdisciplinary researchers involved with the development of intelligent communicative systems. Virtual Reality Modelling Language (VRML) is an effective tool to describe 3D environments increasing the information density for the user and can adding additional layers of perception and meaning to the experience [5]. Inhabiting 3D environments with 3D embodied agents and endowing these agents with conversational capabilities can promote an effective social interaction. Cassell et al [6] have explored the affordances of embodiment and showed that an ECA can improve the interaction and the experience for the user because the agent "enables the use of certain communication protocols in face-to-face conversation which provide for

a more rich and robust channel of communication than is afforded by any other medium available today".

Hayes-Roth [7] has proposed that the Internet should be inhabited with smart interactive characters that can engage users with social communication skills as in the real world, enhancing mundane transactions and encouraging a sense of presence for the user, resulting in more effective and efficient interaction. Developing further this proposal, Ball [3] demonstrated that endowing animated agents with personality and emotion creates a sense of social presence, leading to more useful conversational interfaces. The existence of this social presence is important in order to begin to understand the development of the interaction between the agent and the user. It follows from this that understanding the creation and development of social relationships between the agents and the users is a crucial first step to creating socially intelligent embodied conversational agents.

There is little empirical evidence yet available to demonstrate the effectiveness of ECA's, particularly in e-commerce applications and there is a growing need for the establishment of objective and subjective measures of usability. Ostermann [12] developed an architecture designed to support e-commerce "by providing a more friendly, helpful and intuitive user interface compared to a regular browser". Results from experiments using this architecture showed that facial animation was favoured over text only interfaces. These results are encouraging, but it is also necessary to investigate the range of applications that can be significantly enhanced by the presence of an ECA and what are users' attitudes toward their appearance, personality, trustworthiness and behaviour during the interaction.

The study presented here addresses issues relating to the functionality, behaviour and perception of personality and appearance of ECA's in contrasting e-commerce applications. The goal is to present empirical evidence in support of the use of the agents within e-commerce domains, in addition to documenting qualitative and quantitative data regarding users' subjective experience of successive interactions with the agents. A detailed discussion of the experimental findings is obviously beyond the scope of this section, however the experimental procedure, key findings and challenge problems are presented.

2. EXPERIMENTAL RESEARCH

This experiment aimed to assess 3D male and female embodied agents, appearing as assistants in VRML e-commerce applications (cinema, travel agency and bank). Two types of male and female agents were assessed:

the first was smartly dressed (formal), the second casually dressed (informal). In order to evaluate the agents, a real-time experimental platform system, capable of face-to-face conversation between the user and the agent was used.

The first prediction was that participants would believe ECA's have a role to play as assistants. This prediction was made based on the results of previous experiments, where customers passively viewed conversational agents in retail spaces [9] and indicated a desire to actually converse with them. 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 informal) 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. Finally, as the verbal and non-verbal behaviour for all the agents (male, female; formal, informal) was identical it was predicted that attitudes to the agents' functionality, aspects of personality and trustworthiness would be similar within and between the applications.

2.1. Experimental Platform Design

The system architecture is based on a client-server system. Using a NuanceTM speech recogniser, the users speech input is captured on the client PC. A Java-based dialogue manager controls the direction of the dialogue as the user completes a task in each application. The VRML code is stored on the server PC. VRML is an excellent software language for the creation of interactive simulations that incorporate animation, and real-time user participation. The 3D applications (Figure 1.1) were created using VRML97, the international standard file format for describing interactive 3D multimedia on the Internet.

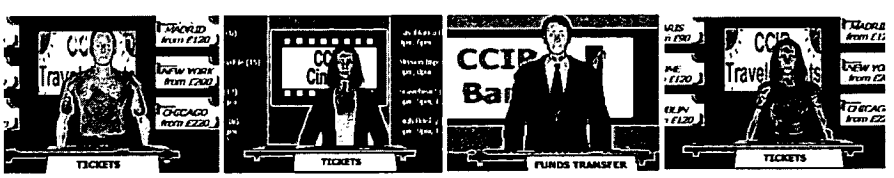


Figure 1.1. Images of ECA's in Applications

The embodied agents were created using MetaCreations Poser 4.0, a character animation software tool. The agents were exported to VRML97, where the code was fitted to the H-Anim specification template [13]. This specification is a standard way of representing humanoids in VRML97. Using this specification it was possible to obtain access to the joints of the agent to create gestures and mouth movements. Four gestures were created for the embodied agents: nodding, waving, shrugging and typing. One male and one female voice recorded the necessary output prompts for the male and female agents respectively. All four agents had the same verbal output.

2.2. Experimental Procedure

Participants were randomly assigned all conditions in a 2 x 2 x 3 repeated measures design: agent gender (male, female), agent type (formal, informal), application (cinema, travel, bank). 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 equally according to gender and age group (age 18-35, 36-49, 50+).

Participants were told they would be asked to speak to assistants to complete a series of tasks in the applications. In all cases the participants were asked to carefully observe the assistant and the application. After the conversation participants completed a 7-point Likert [8] attitude questionnaire relating to the assistant. When participants had seen all four agents in an application they filled out a short attitude questionnaire relating to the application. After the participants had interacted with all fouragents in all three applications they completed a questionnaire stating their application preference. A structured interview followed.

2.3. Experimental Findings

2.3.1 E-Commerce Applications. The mean rating scores from the 10-point (low-high) application rating scale show a largely positive response to the applications. No effects for between-subject variables of age and gender were found. A 3 x 1 repeated measures ANOVA taking experimental application as the independent variable showed no significant effects for applications ($F = 0.76, df = 2.0, p = 0.47$). The cinema was rated the highest, followed by the travel agency and thirdly the bank (mean score: cinema = 6.56; travel = 6.46; bank = 6.12). The 7-point Likert questionnaire used to retrieve information about the par-

ticipants' attitudes toward the applications showed participants felt the applications were easy to use, they were considered to be good ideas and they were equally convenient.

Participants were asked to make a selection as to which application they preferred overall. The cinema application was the preferred choice in comparison to the other applications and a chi-square test showed that it was significantly preferred, ($p < 0.05$). In fact, 40% of participants preferred the cinema application, 14% of participants preferred the travel agency and 14% preferred the banking application. A further 8% did not like any of the applications and 25% of the participant sample liked all applications equally.

Qualitative comments explain the preference for the cinema application. It was preferred because is "seemed easier to use". One participant commented the experience was an improvement because of the feeling of "dealing with someone face to face". Participants commented that more visual content in the form of text output would be an improvement. Participants experienced delayed responses from the system as it was processing information and the general thought was that if the delayed responses in the system could be eliminated, this application would be more successful.

Although the travel agency application was considered to be a good idea and was thought to be easy to use, participants were again concerned with delays in the system's responses, resulting in reduced used confidence in the system.

In addition theere was uncertainty about security, confidentiality and reliability when completing transactions in the banking application. Participants described the information for the banking and travel interactions as being more "critical", with users likely to become anxious if something went wrong. It was also stated that having the opportunity to use the keyboard to enter security numbers may be of benefit. Similar comments about the delays in the system were also made with regard to this application.

2.3.2 Embodied Conversational Agents. A series of 2 x 2 x 3 repeated measure ANOVAs taking agent gender, agent type and application as the within-subject independent variables were conducted to analyse participants' attitudes to the questionnaire items relating to the embodied agents as assistants. The questionnaire addressed key issues relating to the agents' personality, trustworthiness and appearance.

All the agents were perceived as being equally friendly and 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 ($F = 2.97, df = 2.0, p < 0.06$), the mean results did show that the assistants in the bank scored less than the assistants in the other applications (mean score: cinema = 5.15; travel = 5.23; bank = 4.93).

Results showed (Figure 1.2) significant preference for the formal agents in the banking application, ($p < 0.01$). Significant results (Figure 1.3) also showed participants felt it would be more appropriate for agents in the cinema application to be dressed informally and agents to

be dressed formally in the banking application, ($F = 15.65, df = 2.0, p < 0.01$).

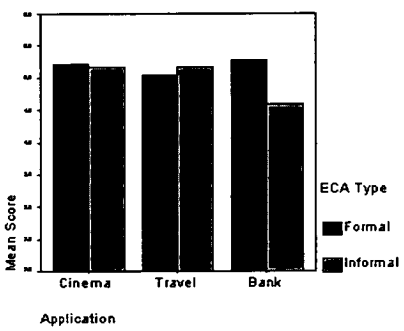


Figure 1.2. Attitude to Appearance

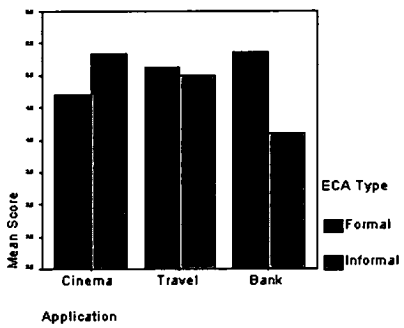


Figure 1.3. Attitude to Appropriateness of Assistants Dress

All participants in the experiment took part in a structured interview. Many comments were ways to improve aspects of the system. Participants felt that the agents' gesturing was at times "a bit awkward". This highlights one of the challenge problems of creating autonomous animated embodied agents with fluid movements. Badler [2] is working to model human movement for the creation of

realistic animated agents. Using parallel transition networks to simulate human-like actions the work aims to "convince the viewer of the a character's skill and intelligence in negotiating its environment, interacting with its spatial situation, and engaging other agents".

Due to real-time technological restraints, some of the output responses were delayed and participants found these delays off-putting and annoying and they gave the impression that the assistant seemed unsure. This again highlights another challenging problem within the area of ECA research. With technological improvements and faster computer speeds, this issue may be resolved. In turn user confidence with respect to the security, confidentiality and reliability of the systems may be improved.

Two thirds of the participants (24 of 36) thought the assistants enhanced the services and they enjoyed speaking to them. One participant is quoted as saying: "Yes, I enjoyed talking to the assistants, I was even polite to them". Participants felt the assistants should be polite and cheerful and should demonstrate competence during the interaction, adding to a more personal interaction. To do this it was suggested that they should smile and provide appropriate verbal and non-verbal feedback.

3. DISCUSSION

It was hypothesised that participants would respond positively to the embodied agents. The results support this prediction suggesting that 3D ECA's have a role to play as assistants in VRML e-commerce applications. It is important to know that ECA's would be welcomed in retail domains especially given the number of commercial websites that are exploring the use ECA's as marketing tools (e.g. Extempo Inc, VirtualFriends). The results supported also a further claim: that casually dressed agents are more suitable in virtual cinemas, and formally dressed agents are more suitable in virtual banking applications.

Participants felt the cinema was more entertaining than the travel agency or banking applications. Although ECA's were welcomed in all three retail applications, reults suggest it is important to consider carefully the seriousness and entertaining nature of the application task and be aware that ECA's might be more effective in less serious applications, where the consequences of failure are less serious. 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. Suggested methods to achieve this included better and faster response times from the agents, having the opportunity to enter data using the keyboard and also seeing additional textual feedback on the interface.

All four agents were perceived to be polite, friendly, competent, cheerful, sociable and agreeable; all traits important for assistants in retail and e-commerce spaces. The trustworthiness of the agents was the only aspect were differences between the applications emerged. The qualitative results showed that participants were less likely to trust agents to complete tasks correctly particularly in the banking application. During the interviews, participants stated that they would be more likely to use the applicaitons if the ECA was more convincing that the inputted information was being processed correctly.

4. CONCLUSIONS & FUTURE RESEARCH

Establishing trust between the agent and the user is of great importance, and on-going research [4] is exploring the construction of a social relationship to assist with establishing trust. Unless users are confident that the agent can understand and process information correctly they may be less likely to trust it, resulting in a less effective interaction. In the study by van Mulken et al [14] results showed personification of interfaces does not appear to be sufficient for raising trustworthiness. If this is the case what other methods could be used for establishing trust in e-commerce applications?

The use of text in the interface could be used to provide feedback to the user as to how much information the agents have received and processed and may improve user confidence. Also, allowing the use of keyboard entry in conjunction with speech input, especially when entering security details is may be an additional improvement. Using the same experimental platform described for this experiment, text-input and text-output will be added to the system and it is intended to further research aspects of user confidence with regard to the use of ECA's in e-commerce applications. Research is suggesting that the development of ECA's in all domains will be dictated not only by technological advances but much more by advances in the understanding and creation of the social interaction between the agent and user, in particular the establishment of trust.

Acknowledgments

Special thanks to colleagues at CCIR for helpful comments, in particular Prof. M.A. Jack and Dr.J.C. Foster. Sincere gratitude is also expressed to Dr. J.A. Anderson for developing the dialogue manager software.

References

[1] E. André and T. Rist. Personalising the user interface: Projects on life-like characters at dfki. In *Proceedings From 3rd Workshop on Conversational Characters*, pages 167–170, October 1998.

[2] N. Badler, R. Bindiganavale, J. Allbeck, W. Schuler, L. Zhao, and M. Palmer. Parameterized action representation for virtual human agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, Mass. London, England, 2000. ISBN 0-262-03278-3.

[3] G. Ball and J. Breese. Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents.* MIT Press, Cambridge, Mass. London, England, 2000. ISBN 0-262-03278-3.

[4] T. Bickmore and J. Cassell. How about this weather? social dialogue with embodied conversational agents. In *Proceedings AAAI Fall Symposium: Socially Intelligent Agents The Human in the Loop*, pages 4–8, November 2000.

[5] M. Bricken. Virtual worlds: No interface to design. Technical Report R-90-2, 1990.

[6] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents.* MIT Press, Cambridge, Mass. London, England, 2000. ISBN 0-262-03278-3.

[7] B. Hayes-Roth. Characters everywhere. Seminar on People, Computers and Design, March 2001.

[8] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:55–62, 1932.

[9] H. McBreen and M. Jack. Empirical evaluation of animated agents in a multi-modal retail application. In *Proceedings AAAI Fall Symposium: Socially Intelligent Agents The Human in the Loop*, pages 122–126, November 2000.

[10] H. McBreen and M. Jack. Special issue on socially intelligent agents. To appear in IEEE Transaction on Systems, Man and Cybernetics, 2001.

[11] C. Nass, Y. Moon, J. Morkes, E. Kim, and B. Fogg. Human values and the design of computer technology. In B. Friedmann, editor, *Computers Are Social Actors: A Review of Current Research*, pages 137–162. CSLI Publications, Cambridge University Press, 1997. ISBN 1-57586-080-5.

[12] J. Ostermann and D. Millen. Talking heads and synthetic speech: An architecture for supporting electronic commerce. In *Proceedings IEEE International Conference On Multimedia and Expo (ICME)*, 2000.

[13] S. Ressler, C. Ballreich, and M. Beitler. Humanoid Animation Working Group, 2001. http://www.h-anim.org.

[14] S. van Mulken, E. André, and J. Muller. An empirical study on the trustworthiness of life-like interface agents. In *Proceedings Human Computer Interaction: Communication, Cooperation and Application Design*, pages 152–156, 1999.

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, trustworthiness 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.

Conference '00, Month 1-2, 2000, City, State.

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 and trustworthiness would be identical within applications. These predictions were made based on the fact that the male and female agents of both agent type (formal and casual) were built on the same platform and they had the same verbal and non-verbal behaviour, with no intentional inclusion of utterances or behaviours that are regarded as being more masculine than feminine or vice versa.

2. SYSTEM DESIGN

The architecture is based on a client-server system [11]. 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.

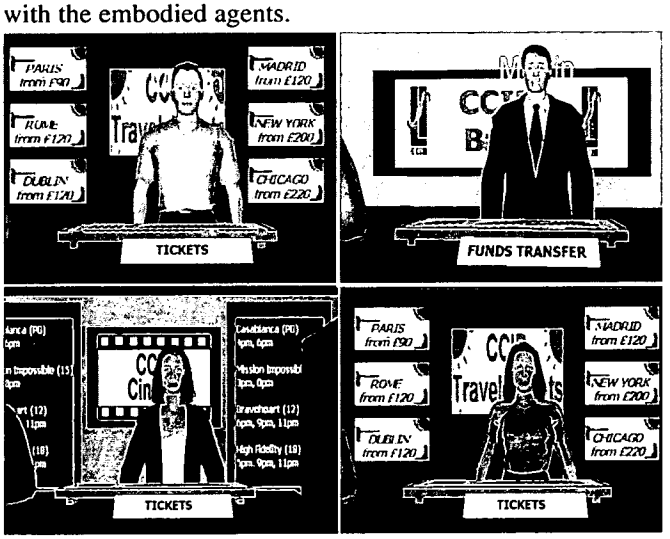


Figure 2: Images of Agents in Applications
(top: left: Casual Male (Travel); right: Formal Male (Bank); bottom right: Casual Female (Travel); left: Formal Female (Cinema))

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. All agent types had the same speech output content.

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
☐	☐	☐	☐	☐	☐	☐

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 Cinema	Mean Travel	Mean 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.

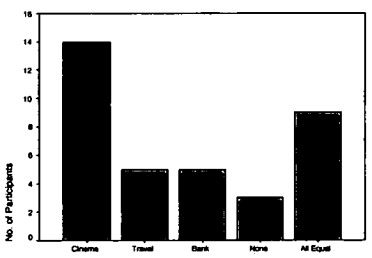


Figure 3: Application Preferences

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 is "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 analysis participants' attitudes to the questionnaire items relating to the embodied agents as assistants. The questionnaire addressed the following key issues: agents' voice, personality, trustworthiness 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 and Trustworthiness

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 also asked if they thought 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, ($p < 0.01$).

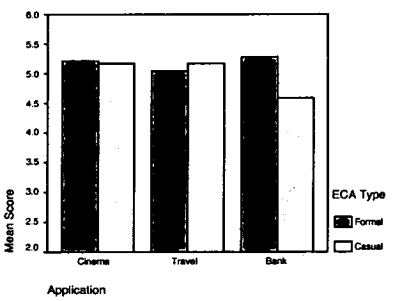


Figure 4: Attitude to Appearance

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

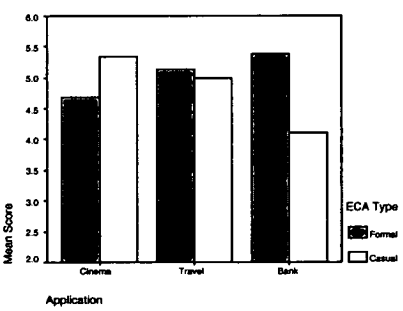


Figure 5: Attitude to Appropriateness of Assistants Dress

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 thought to be polite, friendly, competent, cheerful, sociable and agreeable; all these are 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 the agents would be equally trustworthy. 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

AAAI Fall Symposium: Socially Intelligent Agents – The Human in the Loop (2000)

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 (Wyard 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.

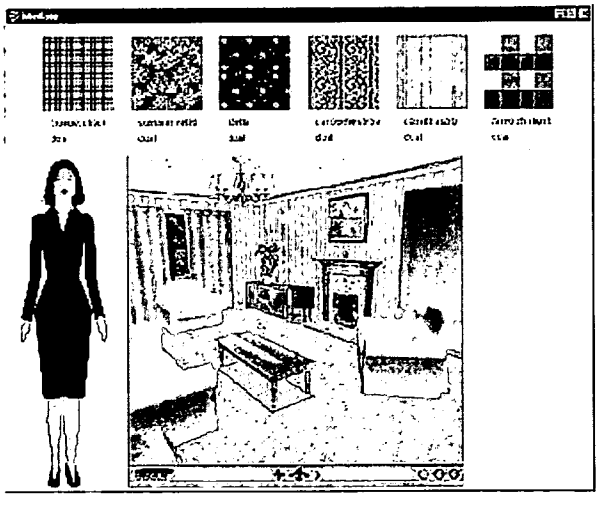


Figure 1: The E-Retail Application Interface with Animated Agent

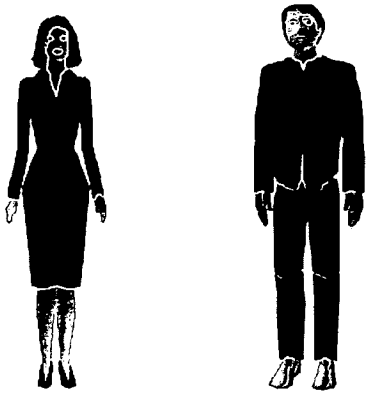


Figure 2: Female and Male Embodied Agents (used for T4, T5 and T6)

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

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 that 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.

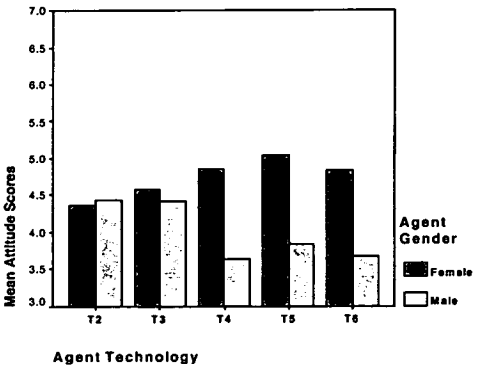


Figure 4: Attitude to Agent Appearance

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 the 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 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.; Billinghurst, 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.

Animated Conversational Agents in E-Commerce Applications

Helen McBreen, Mervyn Jack
Centre for Communication Interface Research,
University of Edinburgh,
80 South Bridge, Edinburgh, EH1 1HN.
+44 131 650 2779

1. Introduction

This research investigates design guidelines to assist in choosing the appropriate conversational agents to act as assistants in interactive e-commerce applications. The first experiment evaluated the effectiveness and user acceptability of human-like agents in an electronic retail application [5,7]. The second evaluated the same cast of agents in another application to explore application dependency effects on the users' attitudes toward the cast of agents. The third experiment evaluated a cast of cartoon-like humanoid agents in an e-retail application. Participants 'eavesdropped' on spoken dialogues between a 'customer' and each of the agents. They completed attitude questionnaires and took part in interviews designed to obtain information relating to their perception of the agents.

Experiments 1 and 2 showed that participants expected a high level of human-like communicative behaviour from the agents. There was a dichotomy in the participant sample between those who wanted to see the agent in the interface and those who preferred just to hear a voice. The second experiment results showed no application dependency: two different groups of participants rated the various human-like agents in similar ways, even though each group witnessed them in a different application. Results from the third experiment showed that with respect to animated humanoid agents, participants prefer 3D to 2D humanoid agents and have a desire to interact with fully embodied agents as opposed to heads alone. In this experiment two-thirds of the participant sample preferred to see an agent in the interface.

The success of these applications is dependent on advances in speech recognition and speech generation. It is for this reason that it has become necessary to investigate users' perceptions of the agents in order to create functional and acceptable e-commerce conversational agents for future applications. It has been pointed out that well-designed embodied interfaces will address needs that are not met in current interfaces. Such interfaces will help discover ways "to make dialogue systems robust in the face of imperfect speech recognition, to increase bandwidth at low cost, and to support efficient collaboration between humans and machines", [1].

2. Experiment Procedure

Different groups of 32 participants took part in each experiment, distributed according to gender and age. They read a brief explanation of the purpose of the experiment and were

primed verbally by the experiment supervisor. They viewed 2-minute videos showing the dialogue between the 'customer' and one of the agents. After listening to each dialogue, they completed a seven point agree-disagree continuum questionnaire [4]. The passive methodology used to assess the agents was practical, it avoided complex technological issues involved in creating fully interactive applications with a large range of agents, but it still allowed an adequate evaluation of each. Ideally an interactive application would have provided more informative results, however to evaluate such a substantial cast of agents this compromise had to be made.

2.1 Experiment 1

The dialogue in this application [7] illustrated the 'customer' conversing with the agent in order to choose room furnishings. The room was visible in the interface as a 3D living room. Male and female human-like agents were all created from photo-realistic images of humans. The cast of agents was based on an incremental technology scale chosen to represent a range of software techniques that may be used in the development of such on-line assistants.

2.2 Experiment 2

The same cast of agents that appeared in Experiment 1 were evaluated here. The agent appeared to the left of the main interface graphic of a CD. The dialogue illustrated the 'customer' conversing with the agent in order to compile a personalised CD. The agents were created to represent a stereotypical assistant who seemed knowledgeable about popular music.

Agent	Description of Technology
H1	Videos of human assistants. Male/Female voice soundtracks used for the other male/female agents respectively
H2	3D talking heads (modelled on T1) with lip-synchronisation to the original male/female voice soundtracks.
H3	Still frames taken from videos (T1) with added graphic lip movement to match the original voice soundtracks.
H4	Still frames taken from the videos of the sales assistants *without* graphic lip movement.
H5	Disembodied voices (male and female)

Table 1: Human-Like Agents

2.3 Experiment 3

Using the home furnishings application twelve cartoon-like agents appeared in the interface.

Agent	Description of Technology
C1	Disembodied voices (male and female)
C2	2D animated head with synchronised lip movement and blinking and smiling during the course of the conversation with the customer.
C3	3D animated head as in C2
C4	2D animated full-bodied agent using the heads of C2, outside the 3D room.
C5	3D animated full-bodied agent using the heads of C3, outside the 3D room.
C6	3D graphically animated full-bodied agent identical to C5, inside the 3D room.

Table 2: Cartoon-Like Agents

3. Human-Like Results

The results of a ten point rating scale completed during the post-experiment interviews were analysed. A 2x5x2 ANOVA, taking agent gender, agent technology and application as the independent variables and the mean rating scores as the dependent variables showed a highly significant main effect for agent technology (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. Table 3 shows the mean ratings for each agent technology (pooled for application and gender) with the results of pair-wise comparisons. The video (H1) was rated best and the 3D talking head (H2) rated worst.

Agent	Mean Rating Score	Technology rated better than... (all at $p < 0.01$)
H1	7.25	H2, H3, H4, H5
H5	5.85	H2, H3, H4
H4	4.70	H2
H3	4.27	H2
H2	3.05	-

Table 3: Ratings for Human-Like Agents

Participants felt the service would be easy to use and was a good idea, depending on the agent that appeared in the interface. Attitudes did follow the overall trend with videos being more popular than H2 ($p<0.05$), H3 ($p<0.01$) and H4 ($p<0.05$). Male agents had more positive scores than their female counterparts.

Although the voices for each male and female agent were identical, participants had varying attitudes toward the clarity of the different technologies. The disembodied voices (H5) were clearer than H2, H3 and H4, $p<0.01$. Also, the videos (H1) were clearer than H2 and H4, $p<0.01$. In addition, the 3D talking heads (H2) were liked the least, significantly less than the H1 and H5, $p<0.01$. As regards gender differences, the male voice was liked more than female. Participants explained that the female voice had a more distinctive accent, which did not appeal to them.

Even though the dialogue between the customer and each agent was identical, the conversation with the videos (H1) was felt to be most natural and was significantly more natural than the conversation with H2, H3, H4 and H5, ($p<0.01$). As indicated previously, the conversation with the male agents was preferred to the female agents. Interestingly, the conversation in the Home Furnishings application was more natural than that in the CD Service application. This could be due to participants not knowing or liking some of the tracks that were chosen, or indeed because of the more relaxed and informal tone of the agents.

As regards the personality of the agents all were thought to be competent in both applications. Participants felt the videos (H1) and disembodied voices (H5) were friendlier than the other agents. Again a gender difference existed with male agents being significantly more friendly than the female. Participants felt that seeing the videos (H1) was more helpful than the other technologies.

The questionnaires also probed participants for information as regards the agents' appearance. Following the

overall trend, pair-wise comparisons showed that the videos (H1) had the most popular appearance and the 3D talking heads (H2) the worst. The appearance of the male agents was significantly preferred to the females. The negative response to the female voice could also have impacted on the attitudes towards the appearance. Participants felt the videos (H1) were significantly more suitable for the applications than H2 ($p<0.01$), H3 ($p<0.05$) and H4 ($p<0.05$) in both experiments.

More concentration was given to the videos (H1) and the still with graphic lips (H4). With respect to gender, more attention was given to the male agents, in all cases except H4. However participants did not attend to this technology significantly, probably due to the fact that it did not move or show much emotion. It may be the case that the more information on the interface, the harder it might be for the participant to concentrate on the agent.

4. Cartoon-Like Results

New male and female speakers who had more experience with recorded speech output utterances for interactive applications were used in this experiment. On completion of the analysis of the results from Experiment 2 no application dependency was found: the human-like agents were rated similarly in both applications. For this reason Experiment 3 used only one application, the Home Furnishings Service. The attitude questionnaires addressed five main issues: voice, personality, appearance, facial expressions and gesturing.

The female voice was significantly preferred to the male voice. Qualitative comments indicated that both voices were clear. However the female voice was more friendly and natural. The male was perceived to be slightly monotonous. T-Tests indicated that the male voice of C4 and C6 ($p < 0.01$) were significantly less preferred to their female counterparts, suggesting that the male embodied characters may be less popular.

Participants felt the disembodied voices and the 3D embodied characters were more polite, suggesting 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. Results showed that the 3D embodied agents (C5, C6) were most friendly and significantly more friendly than the 2D and 3D talking heads.

C4 was significantly more forceful than all other technologies, ($p<0$ .01). Many participants said that agents who made suggestions were more helpful. Although the scripts were slightly different, the scripts of C4 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.

The appearance of the female agents was preferred to the male agents, this

was significantly the case for the full-embodied agents. Many participants thought that the male embodied agents used hand gestures that dominated the interface. The female agents were also thought to be significantly more helpful. A combination of poor attitude responses to the male voice and the exaggerated gestures of the male embodied agents were the probable cause of the gender bias.

The facial expressions of all the agents were friendly. Significant the facial expressions of 3D agents appeared to be the most life-like. In the interviews, many participants said that that the lip movements of the talking heads were distracting because they looked dubbed and artificial. The 2D talking head (C2) had the most distracting lip movement, significantly less so than the embodied characters, $p<0.01$. The facial expressions of C2 were noticeable and it was found that they were unhelpful in comparison to other agents, namely C3 and C5, $p < 0.05$. It may be concluded from this that less obvious and more natural expressions could promote a sense of helpfulness in animated agents.

C4 only had pointing gestures, which were thought to be less exaggerated. Deictic gestures may therefore be more functional in these e-retail interfaces. Many participants felt that the males' gestures were exaggerated which explains why participants may not have liked the male appearance. Gesturing contributed to the perceived friendliness of the agents. The 3D embodied agents 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.

5. Conclusions

The objective of these experiments was to evaluate user acceptability toward a variety of human-like agents and cartoon-like agents in two e-commerce applications. Many key findings, outlined below were discovered during the experimental phases.

Of a cast of human-like agents videos and disembodied voices are preferred. This is due to the strong preference to interact with agents that exhibited human-like facial gestures and emotion. If the human-like agent does not exhibit this non-verbal behaviour many participants prefer to just hear the voice. Results showed that the voice of the agent should be neutral and accent free.

A number of findings were discovered as regards cartoon-like humanoids. Firstly, 3D animated humanoid heads are preferred to their 2D counterparts but 3D embodied characters were thought to be more friendly, helpful and polite. It was shown that gesturing promotes friendliness and helpfulness. Of a cast of animated humanoid agents, two-thirds of the participant sample preferred to see an agent in the interface.

It was seen from the results that competence and forcefulness are 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 all three experiments the voice of the agent strongly affected users' perceptions. Until speech synthesis techniques have advanced it seems more appropriate to use recorded speech, but the selection of the voice is critical. In the first and second experiments the female voice had a distinctive accent that did not appeal to many participants. In the third experiment where new voices were used, the male was deemed monotonous. The results suggest that neutral, accent free voices should be used. A finding that clearly emerged from these experiments 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.

Passive viewing experiments have made it possible to retrieve user feedback and evaluate two e-commerce applications. Continuing with this progression of experiments and reflecting on the feedback retrieved, it is intended to research verbal and non-verbal behaviour of embodied agents in e-retail enterprises in greater depth. The information provided is now being used to create 3D embodied animated agents for a selection of interactive e-commerce applications that will be used as research platforms to address further aspects of human computer communication.

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 University of Edinburgh, in particular Dr. John Foster.

References

[1] Cassell;, J.: 'Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems'. Gesture and Narrative Language Group. MIT, 1998.

[2] King, J., Ohya, J.: 'The Representation of Agents: Anthropomorphism, Agency and Intelligence'. In Proc. CHI 1996.

[3] Koda, T.: 'Agents with Faces: A Study of the Effects of Personification of Software Agents'. MSc Thesis, MIT Media Laboratory, 1996.

[4] Likert, R: 'Some Applications of Behavioural Research'. Paris, 1932.

[5] McBreen, H., Shade, P., Jack, M., Wyard, P.: 'Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications'. In Proc. Autonomous Agents 2000, Barcelona, Spain, June 3-7, 2000.

[6] Thorisson, K: 1996, 'Communicative Humanoids: Model of Psychosocial Dialogue Skills'. Thesis MIT Media Laboratory, 1996.

[7] Wyard, P., Churcher, G.: 'The MUeSLI Multimodal 3D Retail System', ESCA Workshop on Interactive Dialogue Systems, 1999.

Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications

Helen McBreen† Paul Shade†
†Centre for Communication Interface Research
University of Edinburgh
80 South Bridge, EH1 1HN, Scotland, UK
+44 131 650 2779
helen@ccir.ed.ac.uk paul@ccir.ed.ac.uk

Mervyn Jack† Peter Wyard‡
‡BT Laboratories
Adastral Park
Ipswich, England, UK
+44 1473 640 192
maj@ccir.ed.ac.uk peter.wyard@bt.com

ABSTRACT

This paper details results of an experiment to empirically evaluate the effectiveness and user acceptability of human-like synthetic agents in a multi-modal electronic retail scenario. The synthetic personae played the roles of interactive conversational sales assistants. The range of life-like personae differed with respect to gender and technology. Participants took part in the controlled experiment, which involved them eavesdropping on spoken dialogues between a customer and each of the synthetic personae. They also completed questionnaires and took part in a debriefing interview designed to elicit information relating to the effectiveness, believability and perceived quality of each of the personae.

Results show that participants expected a high level of realistic and human-like verbal and non-verbal communicative behaviour in the synthetic personae. This was demonstrated in the strong preference for personae that exhibited natural facial expressions, gestures and emotions. It was also found that disembodied voices were significantly preferred to many of the personae. In addition, results show participants had significantly different attitudes to the voices of the personae.

Keywords

Synthetic personae, anthropomorphic, verbal and non-verbal communication, virtual conversation.

1. INTRODUCTION

Synthetic personae are now a feature of many graphical user interfaces including enhanced multimedia presentations [1] and educational spoken dialogue interfaces featuring conversational characters [8]. Although such personae have been technically developed for applications little is known about user attitudes towards them [12,13]. As a result, this paper describes the first in a series of experiments to investigate user attitudes to different types of synthetic personae in electronic retail environments. The long-term aim of this work is to create a body of experimental data to provide a set of design guidelines for the creation of effective synthetic personae to be used in intelligent multimedia and multi-modal applications involving automatic speech recognition and automatic speech generation technologies.

In such graphical interfaces, the user must be convinced that the synthetic persona can support a face-to-virtual-face conversation. Given the current limited knowledge about the naturalness of such conversations there is a risk associated with the introduction of anthropomorphic characters into GUI's. If the synthetic persona exhibits visually sophisticated communicative behaviour but cannot support the actual dialogue, the interaction between the user and the synthetic persona may collapse. Interface designers need to be aware that user expectations should be determined by the capability of the spoken language interface and this is the reason why fundamental research into understanding the relationship between user attitudes, user expectations and system behaviour is necessary.

It is important firstly to establish the types of synthetic personae with which users would like to interact, if any, and then provide the characters with the necessary verbal and non-verbal communicative behaviour to enhance face-to-virtual-face conversation. In summary, although enhancing interfaces with synthetic personae is no longer a novel idea, theoretical research into the design of such personae is in its infancy. The focus of such interface design should now be "research that can contribute to advancing innovative concepts and that can promote better understanding of what technology works well to make interfaces more usable, useful and accessible", [2].

2. EXPERIMENT PROCEDURE

This experiment aimed to assess a wide range of usability attributes of ten synthetic personae in the context of a multi-modal electronic retail application by having participants 'eavesdrop' on brief dialogues between a customer (represented by a disembodied voice) and a synthetic persona. The passive methodology used to assess the personae was extremely practical, as it avoided the complex technological issues involved in creating a fully functional interactive application with a range of personae, but it still allowed a full evaluation of each of them.

The GUI for the retail application (Figure 1) was created in the style of MUeSLI [14]. 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 synthetic persona was displayed in its own window (top left). The dialogue illustrated the 'customer' conversing with the persona to select colours and patterns in order to decorate the room.

Note: The primary author of this paper is a PhD student

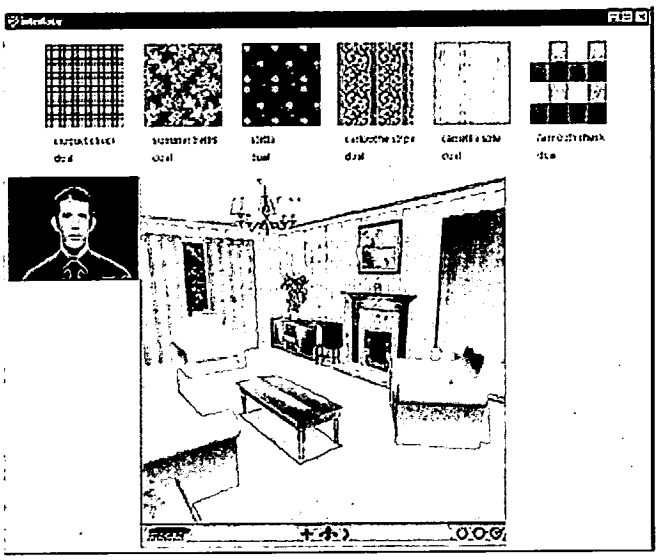


Figure 1: The E-Retail Application Interface

The ten personae differed with respect to the technology (T) used to implement them and gender. Male and female versions of five contrasting technologies were used as follows:

T1: Videos of human (male and female) sales assistants. The female voice soundtrack was used for the other female personae; the male voice soundtrack was used for the other male personae.

T2: 3D talking heads (modelled on T1, T3 and T4) with lip-synchronisation to the original male and female voice soundtracks.

T3: Still frames taken from the videos of the sales assistants with added graphic lip movement to match the original voice soundtracks.

T4: Still frames taken from the videos of the sales assistants (as in T3) *without* graphic lip movement.

T5: Disembodied male and female voices (original voice soundtracks).

The dependent variables in the experiment were the responses to the individual items in the questionnaires, listed in Table 2 and responses given during a semi-structured interview, including an overall rating of each persona. The independent variables were persona gender and technology. Effects of between-subject variables of age, gender and experiment supervisor were also investigated. A total of 32 participants took part in the experiment, distributed according to gender and age as shown in Table 1.

Participant Age Group	No. of Males	No. of Females	Total
18-35	8	5	13
36-49	1	5	6
50+	7	6	13
	16	16	32

Table 1: Analysis of Participants by Gender and Age Group.

The experimental procedure required participants first of all to read a brief explanation of the purpose of the experiment after which they were also primed verbally by the experiment supervisor. They then viewed ten 2-minute videos (created using Macromedia Director 6.5 and presented in randomised order on a Pentium II PC), showing the dialogue between the 'customer' and one of the synthetic personae.

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: *There are more than forty green fabrics, here is a selection.*

Customer: *Try the Cartouche Stripe please.*

Assistant: *Would you like to see it on the chairs as well?*

Customer: *Yes, ok.*

Figure 2: Section of the Dialogue

The same dialogue with the same male and female voice recordings was used throughout. After listening to each dialogue, participants completed a Likert questionnaire [5,7], formatted as shown in Figure 3. 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
☐	☐	☐	☐	☐	☐	☐

Figure 3: Example of a Likert Questionnaire Item

Not all dimensions of interest were relevant to all the synthetic personae, therefore three different questionnaires were used. A total of eight statements were relevant to all the technologies and were therefore included in all questionnaires. Other statements relating to appearance (not relevant to the disembodied voice condition T5) and lip-movement were included as appropriate. The stimulus statements are given in Table 2.

Each session ended with a debriefing interview to investigate:

- participant's views on the use of synthetic personae in e-retail applications
- their effective deployment on screen
- the characteristics required by such personae
- the conversational possibilities with personae in future applications.

Following the interview participants (1) ranked in order of merit their preferred top three sales assistants and (2) rated each sales assistant on a scale of 1 to 10 (ten being the best). At a later time seven participants attended a focus group which examined in greater depth participants' perceptions and opinions about the synthetic personae, their effectiveness and how they might be deployed in applications.

Questionnaire Items	T1	T2	T3	T4	T5
1. I think this service is a good idea	*	*	*	*	*
2. I think this service would be difficult for me to use.	*	*	*	*	*
3. I would like to use this service myself	*	*	*	*	*
4. I felt the assistant was friendly	*	*	*	*	*
5. I felt the assistant seemed competent	*	*	*	*	*
6. The assistant's voice was not clear enough.	*	*	*	*	*
7. I liked the assistant's voice.	*	*	*	*	*
8. I felt the conversation was unnatural.	*	*	*	*	*
9. I liked the appearance of the assistant.	*	*	*	*	
10. I thought the assistant looked natural.	*	*	*	*	
11. I thought being able to see the assistant was helpful.	*	*	*	*	
12. The appearance of the assistant was unsuitable for the home furnishings scenario.	*	*	*	*	
13. I looked at the assistant more than the living room.	*	*	*	*	
14. I felt the speech sometimes didn't match the lips.	*	*	*		
15. I noticed the lips moving.	*	*	*		

Table 2: Questionnaire Items for Each Technology

3. RESULTS

3.1 Overall Ratings

In order to obtain an overall rating of the ten synthetic personae, the results of the ratings provided during the post-experiment interview were analysed. The mean scores for each persona are shown in Figure 4. The videos were rated the best; the 3D talking heads were rated the worst.

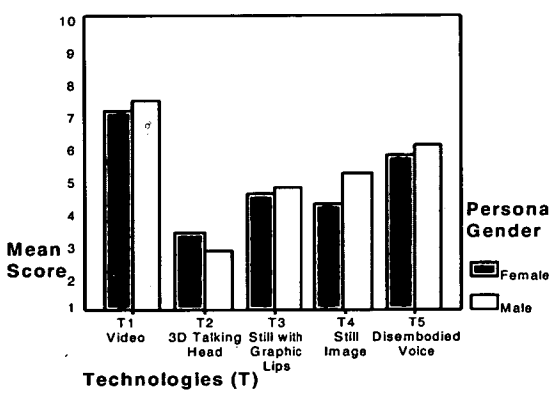


Figure 4: Ratings by Persona Technology and Gender

A 2 x 5 repeated measures ANOVA taking persona gender and persona technology as the independent factors was carried out. Standard multivariate tests such as Pillai's Trace showed highly significant differences due to persona technology ($F = 21.88$, $df = 4.0$, $p < 0.01$) and significant differences for persona gender ($F = 5.10$, $df = 1.0$, $p < 0.05$). There were no significant interactions between persona gender and technology.

Between-technology pair-wise comparisons showed similar though not identical patterns for male and female personae. With respect to the male personae, the video (T1) and the disembodied voice (T5) were rated similarly and both were rated higher than the other three technology types ($p < 0.05$ in all cases). The still image (T3) and still image with lip animation (T4) were rated similarly, and both were rated significantly higher than the 3D talking head (T2), both at $p < 0.01$. Participants' responses to the 3D talking head were poor. This technology was rated the worst of the persona technologies used in this experiment, ($p < 0.01$). The results for the female personae were similar except that the 3D talking head (T2) was rated statistically the same as the still image (T4, $p = 0.062$).

Due to the number of pair-wise tests carried out for this analysis, significance values close to 0.05 need be treated with caution. Consequently, the differences reported above between male T4 and T5 ($p = 0.044$) and between female T3 and T5 ($p = 0.025$) may not in fact be significant. However, the overall pattern remains unaffected with technologies T1 and T5 being the most preferred and T2 being the least preferred.

Analysis of the effects of between-subject variables showed there was no gender bias within the participant sample with respect to persona gender (i.e. male and female participants showed the same preferences for male and female personae). Similarly there was no participant gender bias towards persona technology. No effects were found for participant age group or for experiment supervisor.

3.2 Analysis of Attitude Questionnaire Items

During the experiment an attitude questionnaire was completed, the individual items of which are discussed and analysed below. The results re-enforce the overall ratings presented above, but differ in interesting ways. It should be noted that Likert scales and ratings scales differ in the type of data they provide. The ten-point rating scale used was possibly more sensitive than the 7-point Likert questionnaire because participants were explicitly drawing comparisons between personae, rather than focusing their attention on individual personae when completing the Likert questionnaires. This may be the reason why certain items (i.e. Item's 4 and 5) produced results that differed from the trend suggested by the ten-point rating scale.

Three items in the questionnaire related to the users' attitudes to the *service as a whole*. The service was considered to be a good idea (mean = 5.42), and easy to use (mean = 5.52). Participants also agreed that they would use the service if it were available (mean = 5.05). Persona technology or gender did not influence these positive attitudes.

More surprisingly, the perceived friendliness of the assistant (mean = 4.53) and the perception of the assistant's competence (mean = 5.35), both of which were very positive, were also uninfluenced by either persona gender or technology. A tentative suggestion to the lack of variation between technologies here could have been due to the passive viewing nature of the experiment. Perhaps if the participants had interacted with the sales assistant, they may have had more informed opinions about friendliness and competence.

3.2.1 Attitude to the Voice

With respect to the voice used in the service (one male and one female voice used throughout), the mean for item 6 indicated that participants found the voices of all personae were clear (mean = 5.72). There were no significant differences between persona types. Item 7 produced a mixed response as shown in Figure 5. A

2 x 5 ANOVA did not show significant differences due to persona gender (F = 2.82, df = 1.0, p = 0.104) but did show differences for technology (F = 5.62, df = 4.0, $p < 0.01$).

There was one significant result between the female technologies, that the voice of T1 was significantly preferred to T4, $p < 0.01$. The voice of the male video (T1) was significantly preferred to the male T2 ($p < 0.01$) and T3, ($p < 0.05$). Moreover, it was preferred to all five female personae, all at $p < 0.01$. Attitudes to the voice of male technologies T1, T3, T4 and T5 were significantly better than the voice of the male 3D talking head (T2), all comparisons at $p < 0.01$.

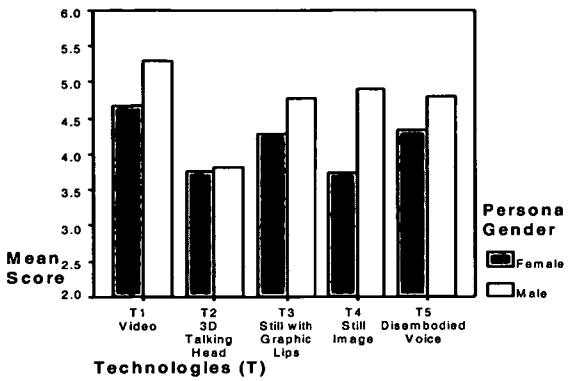


Figure 5: Attitude to Persona Voice (Item 7)

3.2.2 Naturalness of the Conversation

A 2 x 5 ANOVA for item 8 showed significant differences between technologies (F = 2.99, df = 4.0, $p < 0.05$), but no significance differences for persona gender (F = 1.92, df = 1.0, p = 0.179). Figure 6 shows that participants felt the conversation with the female 3D head (T2) was more unnatural than the conversation with the T1, T3 and T5, $p < 0.05$. This persona was also significantly lower than male technologies, T1, T4 and T5 (all at $p < 0.05$). With respect to the conversation with the male personae, the still with graphic lip movement (T3) was considered to be the least natural with a significantly lower score than either the video (T1) or the disembodied voice (T5), both at $p < 0.05$.

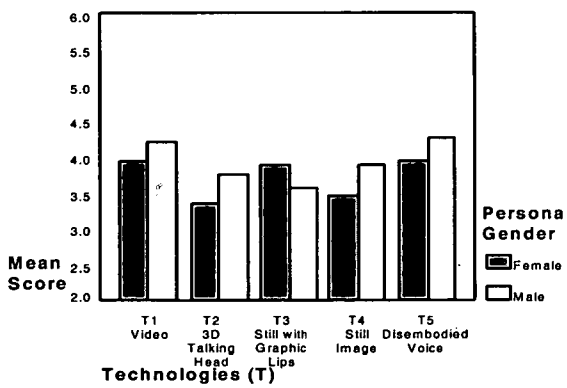


Figure 6: Attitude to the Naturalness of the Conversation (Item 8)

3.2.3 Appearance of the Personae

Five questionnaire items were included for use with technologies T1-T4). A 2 x 4 ANOVA showed that, with respect to the appearance of the personae (questionnaire item 9) there were significant differences due to technology (F = 9.474, df = 3.0, $p < 0.01$) and gender (F = 5.49, df = 1.0, $p < 0.05$).

T-Tests showed the appearance of the male 3D talking head (T2) was significantly worse than T1, T3 and T4, all at $p < 0.01$. Similarly the female 3D talking head (T2) was significantly worse that T1 and T3, both at $p < 0.05$. The male video (T1) was significantly better than all other male technologies and in fact this persona was significantly better than all female technologies, (all at $p < 0.01$). Both T1 genders were significantly better than both T2 genders, all four comparisons at $p < 0.01$.

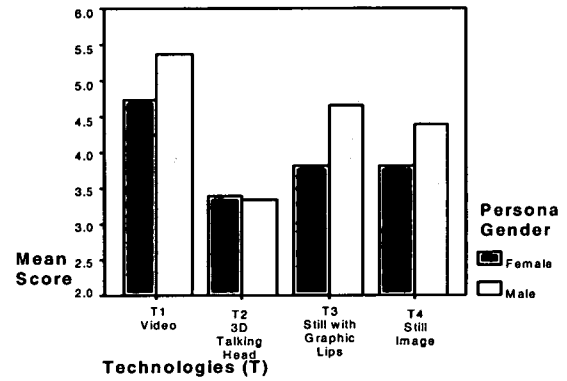


Figure 7: Attitude to the Appearance of the Assistant (Item 9)

Results from a 2 x 4 ANOVA indicated that for item 10 ('I thought the assistant looked natural') there were no significant results for personae gender (F = 0.029, df = 1.0, p = 0.867), but there were significant results for persona technologies (F = 10.574, df = 3.0, $p < 0.01$). On completion of paired sample T-Tests the following significant results were obtained. The male video (T1) was significantly favoured over T2 and T3, both comparisons at $p < 0.01$. The female video (T1) was significantly favoured over T2, T3 and T4, all comparisons at $p < 0.01$. In addition for both male and female personae, T3 was favoured over T2 (both at $p < 0.01$). And finally the male still image (T4) was favoured over the male talking head, T2 ($p < 0.01$).

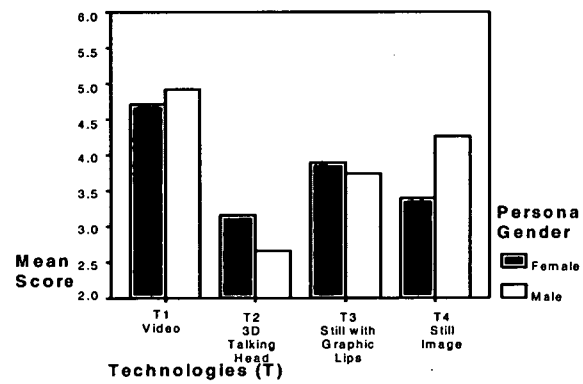


Figure 8: Attitude to the Visual Naturalness of the Assistant (Item 10)

Item 12 ("The appearance of the assistant was unsuitable for the home furnishings scenario") produced significant results with respect to persona gender (F = 3.04, df = 1.0, $p < 0.05$), but no significant differences between persona technologies (F = 1.6, df

= 3.0, p = 0.2). In particular male technologies T2, T3 and T4 were significantly more suitable for the home furnishings scenario than the equivalent female technologies.

3.2.4 Helpfulness of the Personae

A 2 x 4 ANOVA of item 11 showed significant differences between technology (F = 2.55, df = 3.0, $p < 0.05$). There were no significant differences due to persona gender (F = 1.26, df = 1.0, $p < 0.272$). Participants felt that being able to see either the male or female videos (T1) was more helpful than being able to see either the male or female 3D talking heads (T2), all four comparisons at $p < 0.01$. Moreover for the male personae, participants felt that seeing T3 and T4 was more helpful than the 3D talking head, T2 ($p < 0.01$). This is again comparable with the results obtained from the participant ratings in the post-experiment interviews, where the 3D talking heads (T2) were least liked.

3.2.5 Attention given to the Personae

Item 13 ('I looked at the assistant more than the living room') was asked with regard to the personae that were visible in the interface (T1, T2, T3, and T4). A 2 x 4 ANOVA showed that there was no overall significance between persona gender (F = 1.46, df = 1.0, p = 0.237). There was significance between persona technologies (F = 3.14, df = 3.0, $p < 0.05$), that is participants looked at certain persona more than the living room.

For the female personae, when the video (T1) appeared as the assistant, participants looked at the persona more than the living room and significantly more so than the 3D female talking head (T2, $p < 0.05$) and the still image (T4, $p < 0.01$). For the male personae significant results showed that participants looked less at the still image (T4) than the living room compared with other male technologies T1, T2 and T3 ($p < 0.05$).

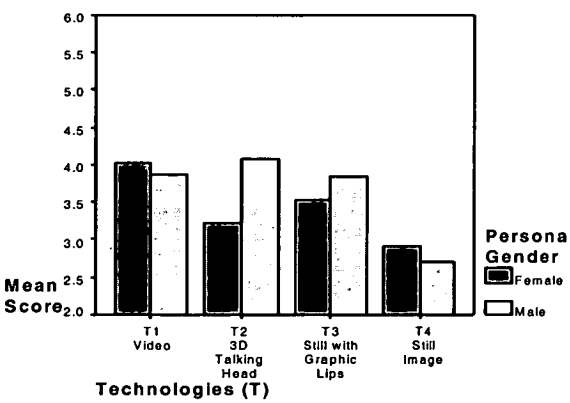


Figure 9: Visual Attention Given to each Personae (Item 13)

3.2.6 Lip Synchronisation

Two questionnaire items were related to the lip movement of the synthetic personae. Only three personae technologies had lip movement and hence a 2 x 3 ANOVA was used to analyse these items.

The ANOVA results for item 14 ('I felt the speech sometimes didn't match the lips') showed a significant difference for technology (F = 5.32, df = 2.0, $p < 0.05$), but not for persona gender (F = 0.023, df = 1.0, p = 0.729). Pair-wise comparisons showed that participants felt both male and female videos (T1) had significantly better synchronised lip movement than the 3D talking heads, T2, and still image with graphic lip movement, T3 (both $p < 0.01$). Moreover, the means for T2 and T3 were below neutral and therefore participants did not think the lip movement matched that well.

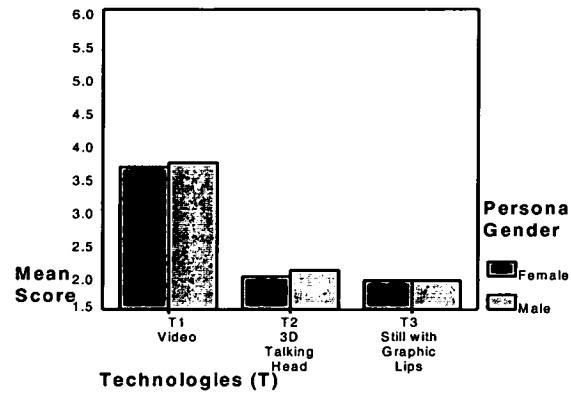


Figure 10: Lip Synchronisation (Item 14)

The results from item 15 ('I noticed the lips moving') showed that participants did notice the lip movement of the personae. A 2 x 3 ANOVA indicated that there were no significant differences within technologies (F = 0.804, df = 2.0, p = 0.46) or between technologies (F = 0.94, df = 1.0, $p < 0.34$).

3.3 Ranking and Rating

In the semi-structured interviews participants were asked to rank the best three assistants. The results are shown here in Table 3.

The male video received the majority of the votes from the 32 participants with the disembodied voices receiving over one third of the votes. There is an interesting relationship between the first and second preferences for the male and female videos (T1) and male and female disembodied voices (T5).

Rank	No. of Votes T1		No. of Votes T2		No. of Votes T3		No. of Votes T4		No. of Votes T5	
	F	M	F	M	F	M	F	M	F	M
1	6	13	0	1	0	1	0	0	6	5
2	11	5	0	0	1	2	1	3	3	6
3	5	5	1	0	4	5	2	6	0	3

Table 3: Preferences for Each Synthetic Persona

13 participants voted for the male video as a first preference. From this group of participants the majority of second preferences votes (8 votes in total) were for the female video. 6 participants voted for the female video first. 4 participants in this group voted for the male video second.

Similarly, of the 6 participants who voted for the female disembodied voice, a total of 4 participants went on to vote the male disembodied voice as their second preference. Finally three of the 5 participants who voted the male disembodied voice as their first preference, voted for the female disembodied voice second. In all cases over half of the second preferences votes went to the opposite gender of the same technology. Other first preference votes went to the male talking head T2 (1 vote) and the male still image with graphic lip movement T3 (1 vote). The highest amount of third preference votes (6 votes in total) were for

the male still image with graphic lip-synchronisation. The rest of the votes were distributed between the other synthetic personae. One participant failed to make a third preference vote.

3.4 Interview Feedback

Feedback from the semi-structured interviews indicated that there was a negative attitude towards the lack of facial expression. They thought the assistants should smile more to exercise more '*civility*'. One participant wrote, "*the assistant's appearance looked very unnatural because of minimal facial expressions*". This suggests that usable applications need to achieve an appropriate level of facial expression.

People noticed that the lip synchronisation of T2 and T3 did not match the speech. Many thought this was distracting and annoying; these participants favoured either the disembodied voices, where they did not have to look at the head, or they preferred the video where the lip movement was more accurate.

There were a substantial number of negative comments about the female voice, but none for the male voice. For example the female voice was described as: "disinterested", "horrible", "tone was over confident", "her voice was harsh and irritating", "tinny". These comments raise interesting questions for the design of synthetic personae. A neutral, accent free voice may be more acceptable than the female voice used in this experiment.

All the participants who expressed preference for the disembodied voices (T5) commented that the picture of the sales assistant was distracting: having to look at the picture distracted them from the visual changes that were happening in the virtual environment. It is important to note that no matter how sophisticated the technology of the visual appearance of the assistant, if the interface demands too much attention from the user, synthetic personae may not actually enhance the service at all.

During the interview participants were asked to indicate the worst assistant and give reasons for their choice. 87.5% of votes were for the 3D talking heads; with 11 votes for the female talking head, 11 votes for the male talking head and 6 votes for both male and female talking heads. Obviously, the technological sophistication of these synthetic personae is far from acceptable with users indicated a marked preference to interact with more natural looking characters.

In summary, the most popular assistants were those which were most realistic and human-like viz. the male and female videos and the disembodied voices.

3.5 Focus Group Feedback

The first issue addressed in the focus group concerned opinions about the sales assistants. The focus group members made comments that the dialogue between the customer and the assistant didn't 'gel together' and seemed 'contrived'. Some participants felt that the assistant was impolite and 'pressurised the buyer'. To improve this aspect of the service, it was suggested that the assistants should have smiled more. In general focus group members felt that the assistant's appearance and voice, lacked emotion.

The focus group members were then shown short excerpts of the ten recordings and were asked to comment on each one. Comments made were that the female video, T1, had an unusual voice and it was apparent that she was reading from a script. There were no comments about the male video. Participants commented that the talking heads, T2, were 'awful', the female head looked distorted and 'dummy' like. They said the male talking head had annoying head movements and seemed 'suspicious'. The female talking head 'distracted attention' from the home furnishings application.

Thirdly, comments about the still frame with hand crafted lip synchronisation, T3, suggested that the male assistant was not interested and was even seen as hostile. The female version was 'comical' and 'too unnatural'. The still image, T4, for both the male and female heads was distracting. Participants said that there didn't seem to be any point in having a still image. They expected the face to move.

Finally the disembodied voices T5 prompted comments that the female assistant appeared to be 'more confident' than the other female assistants. This is interesting considering the dialogue was identical in each video recording. Some members of the group would have preferred to see a face to match the voice. People who favoured the disembodied voice over other technologies suggested that the user should have the facility to remove the picture of the assistant after a certain time.

The participants were asked to suggest ways to improve the assistants. The issue of adding emotion to the face arose. In addition it was pointed out that hand gestures may enhance the performance of the synthetic personae as sales assistants. There was general agreement when participants were asked if they would like to see the sales assistant in the electronic retail environment as a 3D character. Interacting with cartoon characters was suggested.

4. DISCUSSION

The key findings of the experiment showed that:

- There was significant preference for the videos and disembodied voices
- People had a preference to interact with synthetic personae that exhibited facial gestures and emotion
- The 3D talking heads were rated the worst
- The male voice was preferred to the female voice

The aim of the experiment was to evaluate user acceptability toward a variety of male and female synthetic personae. The results showed that participants favoured the male and female videos and disembodied voices to all other technologies used in the experiment. It must be pointed out that it was difficult to gather more detailed information from the participants about the personae, as they did not actually interact with the sales assistant themselves; they merely overheard a recorded conversation.

The popularity of the disembodied voice raises interesting issues about the need for synthetic characters in interfaces. It seems that if the task is visually demanding the user may find a picture of the assistant distracting. It is therefore necessary for interface designers to assess services and applications carefully to establish if a synthetic persona is an actual enhancement.

Questionnaire item 7 retrieved information about user attitudes toward the synthetic personae voices. The same male and female voices were used for each set of five technologies however significant results showed preferences for the voices of the videos (T1) and disembodied voices (T5) suggesting the visual appearance of the persona has an impact on user attitudes toward the voice. This highlights an important cross-modal effect. The voices of T1 were most preferred, suggesting that the more natural

the facial movements, the more acceptable the voice. T5 was also significant here, raising issues also mentioned in the interviews, that if human-like personae do not exhibit natural facial expressions, participants may find them visually distracting.

Once it has been established that there is a need for a synthetic persona in an interface the next question is which type of persona? It is concluded from this experiment that people prefer to interact with more natural looking personae that exhibit human-like facial expressions, facial gestures and emotions. Essentially, the user demands that the system support natural face-to-face dialogue for it to be successful. The low ratings of the 3D talking heads suggest that the technology used to create them is underdeveloped.

During the focus group and semi-structured interviews participants were invited to make suggestions for improving the appearance of the sales assistants. Signals of friendliness (such as smiling) were given high priority for enhancing retail services. In general, facial expressions play a crucial role in attitude and perception of services. Lip synchronisation was highlighted: bad lip-synchronisation was found to be distracting and synthetic personae in which this was the case were undesirable. The choice of a suitable voice is also an issue that needs to be carefully considered.

The dialogues used for e-retail applications need to be examined more closely. The responses of the sales assistant impact significantly on the attitude responses of the participants. The output dialogue will contain a multitude of information about the personality of the persona. These are issues that will be even more important when creating fully functional systems. Virtual sales assistants must have traits that are conducive to selling and advising; stable, relaxed, sociable, conscientious, cheerful, patient and diplomatic.

This experiment was the first in a series. The next stage is to evaluate the same ten synthetic personae in a different e-retail environment. In this way the effectiveness of these synthetic personae will be investigated in a contrasting environment and it will be established if participants evaluate the same personae differently in this new environment. Secondly, the cast of synthetic personae will be extended to include five additional characters, which will include 2D and 3D humanoid cartoon-like characters. Based on issues that arose while evaluating the results of this experiment, a variety of voices or the voice of a professional actor may be used.

5. 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 University of Edinburgh, in particular Dr. John Foster.

6. REFERENCES

[1] Andre, E et al:, 'Personalising the User Interface: Projects on Life-like Characters at DFKI'. In Proc. Workshop on Conversational Characters, Tahoe City, California, 1998.

[2] Biermann, A. et al. (National Research Council): 'More Than Screen Deep: Toward Every-Citizen Interfaces to the Nation's Information Infrastructure'. National Academy Press. 1997.

[3] Bernsen, N. O.: 'Towards a Tool for Predicting Speech Functionality'. In Speech Communication, vol. 23, 1996.

[4] Cassell, et al. J.: 'An Architecture for Embodied Conversational Characters'. In Proc. Workshop on Conversational Characters, Tahoe City, California, 21-30, 1998.

[5] Foster, et al: 1998, 'An experimental evaluation of preferences for data entry method in automated telephone services'. In Behaviour and Information Technology, vol. 17, no.2, 82-92.

[6] King, J. and Ohya, J.: 'The Representation of Agents: Anthropomorphism, Agency and Intelligence'. In Proc. CHI 1996.

[7] Likert, R: 'Some Applications of Behavioural Research'. Paris, 1932.

[8] Massaro, D: 'Perceiving Talking Faces: From Speech Perception to a Behavioural Principle'. MIT Press, 1998.

[9] Maybury, M. T.: 'Toward Co-operative Multimedia Interaction'. In Lecture Notes in Artificial Intelligence 1374, Multi-modal Human-Computer Communication, 1998.

[10] Parke, F.I. and Waters, K.: 'Computer Facial Animation'. A.K. Peters, Wellesley, 1996.

[11] Shneiderman, B.: 'Looking for the Bright Side of User Interface Agents'. University of Maryland, 1994.

[12] Thalmann, et al: 'Digital Actors for Interactive Television'. MIRALab, University of Geneva, 1998.

[13] Thorisson, K: 1996, 'Communicative Humanoids: Model of Psychosocial Dialogue Skills'. Thesis MIT Media Laboratory, 1996.

[14] Wyard, P et al: 'The MUeSLI Multimodal 3D Retail System', ESCA Workshop on Interactive Dialogue Systems, 1999.