Persuasive Interactive Non-Verbal Behaviour in Embodied Conversational Agents

A thesis submitted for the degree of Doctor of Philosophy (Ph.D)

at Newcastle University

by

John Shearer

BEng(Hons)(York)

School of Computing Science

Claremont Tower

Newcastle University

NE1 7RU

United Kingdom

October 2008

Abstract

Realism for embodied conversational agents (ECAs) requires both visual and behavioural fidelity. One significant area of ECA behaviour, that has to date received little attention, is non-verbal behaviour. Non-verbal behaviour occurs continually in all human-human interactions, and has been shown to be highly important in those interactions. Previous research has demonstrated that people treat media (and therefore ECAs) as real people, and so non-verbal behaviour is also important in the development of ECAs. ECAs that use non-verbal behaviour when interacting with humans or other ECAs will be more realistic, more engaging, and have higher social influence.

This thesis gives an in-depth view of non-verbal behaviour in humans followed by an exploration of the potential social influence of ECAs using a novel Wizard of Oz style approach of synthetic ECAs. It is shown that ECAs have the potential to have no less social influence (as measured using a direct measure of behaviour change) than real people and also that it is important that ECAs have visual feedback on their interactants for this social influence to maximised. Throughout this thesis there is a focus on empirical evaluation of ECAs, both as a validation tool and also to provide directions for future research and development.

Present ECAs frequently incorporate some form of non-verbal behaviour, but this is quite limited and more importantly not connected strongly to the behaviour of a human interactant. This interactional aspect of non-verbal behaviour is important in human-human interactions and results from the study of the persuasive potential of ECAs support this fact mapping onto human-ECA interactions. The challenges in creating non-verbally interactive ECAs are introduced and by drawing corollaries with robotics control systems development behaviour-based architectures are presented as a solution towards these challenges, and implemented in a prototypical ECA. Evaluation of this ECA using the methodology used previously in this thesis demonstrates that an ECA with non-verbal behaviour that responds to its interactant is rated more positively than an ECA that does not, indicating that directly measurable social influences will be possible with further development.

Acknowledgements

I would like to acknowledge the following people who have been involved with my PhD. I am grateful to you all. Pressies and actual spoken thanks on its way.

I also wish to express my deepest sadness that my good friend Rumesh is no longer here to have come to this point together, but had a wonderfully energetic and positive impact on me and so many others.

Patrick Olivier	for reading, re-reading, and then saying "whatever". seriously thanks so much.
Dan Jackson	code god #1
Chris Malkin	the rock of Devonshire fun
James Shearer	for about a million hours proof-reading and being great
Annabel Bixby	from day 1
Jennifer Shearer	for politely remembering not to mention the certain 3 letters
Phil Heslop	bringing it right back
Wayne Smith	just cause I want to
Louise Pease	cause she know the right answer to: pub?
Cas Ladha	so that's what 15,000 volts feels like
Fiona Shearer	got out of proofreading the easy way
Peg Shearer	for another million hours proof-reading and being lovely
Tania Brodie-Clarke	for letting me win
Bob Hurling	for being a vital part of the Unilever crew
Ali McGowan	handstands rule ok
Robyn Taylor	definitely on the list of best visitors
Dave Mawer	officially holds the prize of longest housemate ever
Paul Dunphy	for that peaceful easy feeling
Jurgen Wagner	for being the legend of the Jurg
Stephen Lindsay	didn't want to miss a thing
Jon Hook	VJ, smeeJ
Marco De Boni	Unilever made it happen
Jayne Wallace	gromit?
Qaz Chaudry	dude, where's my car?
Pauline Addis	what are all these bees about?
Krish Korvi	for doctoring so well
Christine Niedermeier	you thought I'd forget!

Everyone else in space 2 who read through stuff or just was generally good fun
Everyone who took part in my studies – couldn't have done it without you all

Publications

Shearer, J., Olivier, P., Heslop, P., & Boni, M. D. (2006). *Requirements of non-verbal communication in believable synthetic agents.* Paper presented at the AISB'06: Adaptation in Artificial and Biological Systems conference, Bristol, UK.

Heslop, P., Shearer, J., & Olivier, P. (2006). *Synthetic character fidelity through non-verbal behaviour in computer games.* Paper presented at the iDiG International Digital Games Conference conference, Portalegre

Shearer, J., Olivier, P., & Boni, M. D. (2007). *On the simulation of interactive non-verbal behaviour in virtual humans.* Paper presented at the AISB 07 conference, Newcastle

Shearer, J., Olivier, P., Boni, M. D., & Hurling, R. (2007). *Exploring persuasive potential of embodied conversational agents utilizing synthetic embodied conversational agents.* Paper presented at the Persuasive 2007 conference, Stanford, CA

Shearer, J., Heslop, P., Olivier, P., & De Boni, M. (2008). Non-verbal behaviour for believable synthetic agents. *AISB Journal*(to appear).

Contents

Figures and Tables

1. Introduction

In interactions between people, non-verbal behaviour is highly important and natural, and occurs in all interactions. It serves to communicate a large variety of information, both intentionally and otherwise and also assists with the dynamics of the interaction, providing cues such as who wants to talk. People are highly sensitive to and read a lot into the non-verbal behaviour of other peoples.

This thesis investigates non-verbal behaviour in humans and how it can be applied to Embodied conversational agents (ECAs). More specifically, the focus is on how non-verbal behaviour can influence the persuasiveness of ECAs and how ECAs could be developed to be more persuasive.

Cassell (2000) defines ECAs as "*computer-generated cartoon-like characters that demonstrate many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication*". Presently, ECAs occur very frequently in computer games, but for the most part, at this point, the characters in computer games do not engage in two-way conversation and so aren't generally termed ECAs. A few games, such as Half-Life 2 (Valve Corporation, 2004), have begun to add limited conversational abilities to their characters, and the expectation is that characters in games in the future will be further developed in this regard.

ECAs have received significant attention from the research community, usually with a view to creating service agents – agents that assist with some task such as giving directions or providing information. These characters build on many decades of research in the fields of natural language processing (NLP) and artificial intelligence (AI) which since the development of Eliza (Weizenbaum, 1966) in 1966 and SHRDLU (Winograd, 1968) in 1968 have begun to be able to hold text-based conversations. ECA research develops this to provide embodiment for these conversational agents, along with speech synthesis and sometimes speech recognition capabilities. Present-day ECAs can understand natural language (though usually only through a text-based interface) and can generate appropriate responses, including looking for appropriate information in

knowledge bases. Only recently has attention been given to providing non-verbal behaviour for these agents, such as gaze behaviours and gesture. This non-verbal behaviour has usually been highly task-oriented, such as providing gestures to go with directions (Kopp et al., 2004).

Along with verbal behaviour, non-verbal behaviour can influence the beliefs, attitudes and actions of others. This influence occurs both within the conscious awareness of interactants and also outside conscious awareness. For example, Alice may smile at Bob while talking with him and even though he may not notice he will subconsciously be guided to like her more, find her happier, and be more amenable to her. The influence over others that non-verbal behaviour provides gives advantages to individual humans, and presumably given its frequency it offers advantages to the human (and other) species as a whole. This is in part, by allowing societies to function by providing control without resorting to physical influence – "*society is a massive group of people influencing, persuading, requesting, demanding, cajoling, exhorting, inveigling, and other manipulating each other to further their ends. We call it society because we persuade instead of physically coerce"* (Rhoads, 1997). Non-verbal behaviour is an aspect of social behaviour. Darwin raised the questions of "*why do wrinkle our nose when we are disgusted, bare our teeth and narrow our eyes when enraged, and stare wide-eyed when we are transfixed by fear?*" (referenced in Zanna, 1996) and proposed that "*they are vestiges of serviceable associated habits – behaviours that earlier in our evolutionary history had specific and direct functions. For a species that attacked by biting, baring the teeth was a prelude to an assault*". Behavioural ethologists before Darwin (Hinde, 1972; Tinbergen, 1952) suggested that "*humans to these things because over the course of their evolutionary history such behaviours have acquired communicative value: they provide others with external evidence of an individual's internal state. The utility of such information generated evolutionary pressure to select sign behaviour , thereby schematizing them and, in Tinbergen's phrase, 'emancipating them' from their original biological function"*.

In addition to humans having non-verbal influence over each other, people also have non-verbal influence over animals and vice versa. This is clear from natural interactions with social animals, especially pets, and is supported by scientific studies (Allen, 2003). People readily ascribe intention and human social attributes to non-human entities, such as pets, and this suggests that people would be likely to anthropomorphise entities such as various forms of media, including television, video, film and computers. This has been demonstrated most clearly by Reeves and Nass in *The Media Equation* (Reeves & Nass, 1996), in which it is shown that people treat computer interfaces, even with human or animal form, as social actors – people or things that perform social actions within an interpretive sociological perspective (Weber, 1978). Treating these entities as social actors also introduces the possibility that such entities may even have social influence, and this has in fact been shown by Bailenson and Yee (Bailenson & Yee, 2005).

It would be expected that ECAs with their strong realism, conversational abilities and sense of intention could have a strong social influence. In other words, ECAs could be used to affect the beliefs, attitudes and actions of real people, for positive or negative ends. There could be significant value in ECAs of this kind. They could be used to affect better eating or exercise habits, to stimulate people to give more money to a charity, or to persuade people to buy a certain product. Effective use of non-verbal behaviour in ECAs could enable ECAs to have greater social influence. In addition to providing more effective service or assistive agents, and more effective advertising agents, these could also provide enhanced engagement and realism for game characters, and a variety of other applications.

While ECAs may have many advantages, being natural, emotionally expressive (if so desired), engaging, and familiar, they are not suitable for all situations. For example, an ECA as a component of an in-car interface would create a significant hazard by drawing the visual attention of the driver away from the road. That said, in many circumstances, ECAs do have some strong positive points. These include the fact that humans are comfortably with faces as a form of interaction; that mouth and head movement help people understand speech (Massaro & Stork, 1998; Munhall et al., 2004); that eyes assist

in determining whose turn it is to speak; and that faces help people understand the underlying 'mind' (of a computer) – for example, by reflecting confusion, distraction or busyness. One of the most significant challenges in developing ECAs is that people expect them to behave like real people. When an ECA does not get its behaviour quite right, there is the risk that it may create a strong negative response similar to the effect seen where highly realistic but imperfect humanoid robots cause a sense of revulsion among human observers named the 'uncanny valley' (Mori, 1970). Furthermore, people expect ECAs to have a set of abilities matching that of real people – to hold full conversations, to think for themselves, to remember facts they're told, etc. – and when this is not the case there is risk of confusion and disappointment.

This thesis focuses on the extent to which much non-verbal behaviour in ECAs can affect the actions or behaviour of real people, which aspects of non-verbal behaviour may be important in creating a persuasive effect, and how these aspects may used to aid the development of ECAs. Throughout the thesis attention is given to how ECAs can be evaluated in objective empirical studies, for social influence effects or otherwise. The thesis does not give attention to the ethical issues and possible impacts of ECAs that can, possibly strongly, influence people, nor to their possible presence on the web or in computer games where their behaviour may be less managed than in many environments, and where they may be interacting with vulnerable groups, such as children. Furthermore, the thesis focuses on non-verbal behaviours and avoids significant discussion of the role or impacts of facial and emotional expressions. For the most part, emotional content appears to be an overlay on behaviours, and is itself reflecting a set of values of various emotional attributes such as anger, happiness or fear.

2. Non-verbal behaviour in people

The non-verbal behaviour of ECAs must be based upon the non-verbal behaviour of real humans. Therefore, in order to develop ECAs with non-verbal behaviour a deep understanding of non-verbal behaviour in humans is required. An in-depth view of non-verbal behaviour in humans and how and when it occurs in interactions between people, especially in duologues – interactions between only two people – is presented this chapter. Non-verbal behaviour is defined, with a description of how it differs from verbal behaviour; an indication is given of the various roles non-verbal behaviour serves in interactions between people and which body parts may perform various non-verbal behaviours. An overview of previous investigations and studies into non-verbal behaviour is presented, along with some of the techniques for elucidating various aspects of non-verbal behaviour. The question of the importance of non-verbal behaviour is addressed, with specific attention given to the frequent issue of how much of communication is non-verbal. Discourse convention is discussed along with the role of non-verbal behaviour in managing conversation.

Classifications of non-verbal behaviour are discussed, such as kinesics – body movements including self-adaptors, object-adaptors, gesture – and gesture is taken as an example to illustrate the complexity of non-verbal behaviour, and its additionally complex relationship to speech. The various types of gesture – emblematic, iconic, metaphoric, deictic, emphatic, and cohesive – are described.

The role which eyes play in communication is described, based on five functions of gaze behaviour beyond information gathering with conversation – namely, regulating the flow of communication; monitoring feedback; reflecting cognitive activity; expressing emotions; and communicating the nature of an interpersonal relationship. Attention is given to how eye behaviours relate to the underlying speech stream and to the internal state of the underlying system.

How the body is used and arranged in the physical world is described (proxemics) with special attention to how these distances and arrangement change depending on context and the significant variations in proxemic behaviour across cultures. Touching (haptic)

behaviours are also discussed along with the power of haptic behaviours to portray basic states such as love, affection, or hostility. Also discussed is how the levels of allowed haptic behaviour relate to the social bond contexts – functional bonds, social bonds, friendship, love, and sexual bonds – along with consideration of how these haptic behaviours vary between cultures.

Finally, this chapter discusses vocalisations which are typically not included in the phonological description of language (paralanguage), such as intensity, prosody, laughing, and short utterances such as 'uh-huh', along with the effects of smells and passive communication (for example, clothing play in interactions).

Non-verbal behaviour covers all behaviour other than the spoken word and is highly important in interactions between human beings. Non-verbal behaviour is displayed both with and without intention during all interactions with other people, and perceived non-verbal behaviour affects a viewer both consciously and sub-consciously. For example, if Alice points to a teacup while speaking to Bob, this affects Bob such that he understands that Alice is referring to the teacup.

Non-verbal behaviour generally serves to communicate and it is therefore important to discuss what is meant by communication. In common usage, human communication means the transference of information with some intention, where intention means some mentally formed high level outcome of meaning or significance. The aspect of intention is important, but also confusing. Without intention, communication would be simple information transference, which is too loose a definition. By that definition when, for example, Alice sees a teacup, its colour, shape, etc. would be *communicated* to her. This is not what is usually meant by communication. Communication requires some form of intention, either on the sender's side, or on the side of the receiver perceiving intention from the sender. So, if Alice intended to transfer information to Bob then that is communication, or if Bob thinks that Alice intended to transfer some information, that is also communication. This dual-intention definition of communication is used from here on in this thesis. Studies have shown that "*people can differentiate, even without speech,*

between gestures that are intended to convey meaning and gestures that only seem to emphasize what a speaker is saying. This conclusion may be extended to suggest that most body gestures, facial expressions and so on, are often specifically produced to be understood as part of a person's overall communicative intentions and must be recognised as such for successful interpersonal interactions to occur" (Gibbs, 1999).

Non-verbal behaviour occurs even when communication is not occurring. For example, when Alice gestures while talking with Bob on the telephone there is no information transference of the gesture so there is no communication of gesture-based content, but there clearly is non-verbal behaviour. For this reason, within the thesis the term non-verbal behaviour is used rather than non-verbal communication, though in practice non-verbal behaviour displayed during any form of interaction is usually communicative, and much of the more complex non-verbal displays appear to exist mainly for communication, and are frequently intentional. In contrast, verbal behaviour is almost entirely used as explicit communication.

2.1 Purposes of non-verbal behaviour

The important aspect of both verbal and non-verbal behaviour is how the interaction and the parties involved are influenced by those behaviours. In other words, the purpose or purposes served by an individual behaviour during an interaction must be ascertained. As such, non-verbal behaviour can be categorised by the purpose(s) served, though it is more common to categorise non-verbal behaviour by the section of the anatomy used.

The following are typical anatomical categories:

Kinesics Movement of the body and visible behaviours such as gesture are termed kinesic behaviour. Generally these movements involve the hands and the head, though other body parts may be used (Birdwhistell, 1971; Kendon, 1972).

Oculesics	Eye behaviours, termed oculesic behaviours, such as gazing and eye contact perform functions such as regulating the flow of conversation, monitoring feedback, reflecting cognitive activity, expressing emotion, and indicating the nature of an interpersonal relationship.
Haptics	Touch and touching behaviours are called haptic behaviours and predominantly use the hands, although other body parts can be involved, especially in non-conversation scenarios.
Proxemics	The use and arrangement of the self in the environment, especially in relation to other people, are proxemic behaviours. This is the use of personal space.
Paralanguage	Aspects of vocalisations other than the actual words, such as emphasis, are termed paralanguage, literally meaning 'alongside language'.
Olfactory	Unlike other mammals, humans do not significantly generate aromas as an active behaviour, and are also not as sensitive to them, but smells are certainly important.
Appearance	The way an interactant looks, from hair colour to the clothes worn, can communicate a variety of things to an interactant, though in a passive way. There is no specific action at the time of an interaction.
Symbolism	Symbols hold strong meanings within cultures and their presence or absence indicate certain things, such as allegiance to some association or group. These symbols

may be individual items or patterns, or components of other items.

Artefacts The placement and movement of artefacts within an interaction are another important form of passive communication.

Observed behaviour The way that another person behaves can be observed from a distance and this can provide important cues about that person. For example, if Eve sees Alice act in an aggressive way towards Bob, she can conclude that Alice may pose a threat to other people as well.

Chronemics How fast an act occurs and when it occurs indicates certain information about that act or about things more generally. For example, if Alice is always late when meeting Bob, then Bob may conclude that Alice isn't enthusiastic about meeting him. Or if Alice performs an action slowly, Bob may conclude that she is unenthusiastic about that action or, alternatively, unenthusiastic or tired more generally.

Verbal and non-verbal behaviour in interactions have long been studied by psychologists (Beattie, 2003; DePaulo & Friedman, 1998), linguists (McCafferty, 1998), psycholinguists (Fischer, M. & Zwaan, 2008), anthropologists (Hall, Edward Twitchell, 1973), sociologists (Key, 1980), health care professionals (Derlega, 1995), and business consultants (Greatbatch & Clark, 2005). Riggio and Feldman (2005) provide an overview of various applications areas of non-verbal behaviour. Common sense experience provides strong experience of 'body language' and what non-verbal cues are important. While non-verbal behaviour is highly important in human interactions, it is not vital. For example, the written word is purely verbal, lacking even paralanguage, while telephones and radios remove all non-verbal cues other than paralanguage.

The importance of non-verbal behaviour is unquestionable in human interactions. One of the most frequent discussions of non-verbal behaviour involves how much of our communication is non-verbal as opposed to verbal. This is a very difficult question to answer because of difficulties in measuring information transference between communicating parties, and in practice people adapt their communicative strategies to the context, making a single metric of how much is non-verbal inappropriate. Furthermore, there is no particular value in determining these proportions, other than the clear fact that non-verbal behaviour is important. Even so, significant attention has been devoted to attempting to determine what proportion of communication is non-verbal.

Birdwhistell (1971) claims that up to 65% of communication is non-verbal, while Mehrabian (1971) is often misunderstood in saying that "*55% of the meaning of communication is body language, 38% is in tonality, and 7% rests in the words themselves*". In fact, Mehrabian's proportions apply only to how much is contributed towards *liking* or *disliking* a person when that person is displaying incongruent messages. In other words, his proportions do not apply to general communication, and if applied to general communication by extrapolation would mean that 93% of meaning is lost when reading a book.

2.2 Verbal versus non-verbal behaviour

In computer science much attention has been paid to developing computational models of speech and language using a variety of statistical and symbolic processing approaches. Speech synthesis systems using computers have been demonstrated since the early 1960s (Bell Labs, 1997) and are now present in a large variety of products, including toys, car information systems, screen readers for people with visual impairment, and websites converting written news to speech. The most recent speech synthesis technologies are close to human grade speech (Aylett & Pidcock, 2007).

In contrast, speech recognition, language understanding, knowledge representation systems, language generation systems, and dialogue systems are less mature, posing

greater technical challenges because of background noise, variations in speakers' voices (both differences in voice for one speaker and differences in voice between different speakers), complex prosodic (intonational) aspects and ambiguity in speech. However, speech recognition applications for limited domains are presently available in, for instance, telephone menu systems, some computer games, and information services. Various approaches are used for speech recognition including dynamic programming (Bellman, 1957), knowledge bases and neural networks (Katagiri, 2000), with the Hidden Markov Models (Rabiner, L.R., 1989; Rabiner, L. R. & Juang, 1986) being the most widely used underlying technology. Speech recognition attempts to find the most likely sequences of words given the variations in speech and high levels of background noise.

Language understanding uses knowledge of the structure (syntax) of language (Chomsky, 1957), together with the meanings both of component words (lexical semantics) (Pustejovsky, 1995) and of component word combinations (compositional semantics) (Sauerland, 2007), to determine the meaning or meanings of a whole utterance. Furthermore, the appropriate detection of polite and indirect language (pragmatics) (Levinson, 1983) is important to understand overall meaning. Once overall meaning has been established, appropriate responses can be determined, possibly resulting in responsive speech, and thus creating conversation. Within a conversation there is much discourse convention about which party speaks when, and how transitions of turns and topics are made. Finally, all the previous systems must be reversed to generate the final speech signal.

In summary, the language knowledge required to engage in complex language behaviour comes in six categories (Jurafsky & Martin, 2000):

Phonetics and phonology	The study of linguistic sound.
Morphology	The study of meaningful components of words.
Syntax	The study of the structural relationships between words.

Semantics	The study of meaning.
Pragmatics	The study of how language is used to accomplish goals.
Discourse	The study of linguistic units larger than a single utterance.

In practical systems the language understanding system can feed back to the speech recognition system what words are more likely to occur given the previous words and their structure, enabling more effective speech recognition.

Computers are currently able to understand and generate speech within certain domains. This 'Natural Language Processing' (NLP) focuses mainly on plain text input and output, and usually just that of written language – which is much more consistent and coherent than spoken language. NLP tends to expect complete, structured sentences, whereas natural speech introduces many complexities such as discontinuities, corrections, and higher rates of errors and ambiguities.

In addition to the difficulties in understanding the language, speech recognition techniques give poor results in a general domain, with low word accuracy and low correct sentence meanings, so they are only effective in limited domains. They also consume a significant amount of computational power. With slow, clear speech, low noise environments and constrained domains (words, phrases, or meanings) speech recognition techniques perform well and show much promise, enabling wide use in automated telephone systems.

Verbal behaviour is almost entirely intentionally communicative. Exceptions to this are speech practice (babies babbling, adults practising presentations); speech when in a different or imagined world – sleep, children with toys, a variety of mental disorders; or speech to assist other cognitive processes – speaking while performing a task or developing a concept. In fact this latter is also true of non-verbal communication – there is evidence to suggest that gestures, for example, assist cognitive processes – "*Gestures, together with language, help constitute thought*" (McNeill, 1992).

Understanding non-verbal behaviour is more complex than understanding verbal behaviour for a myriad of reasons. In contrast to verbal behaviour it is not clear (possibly inherently) when or which aspects of non-verbal behaviour are communicative in a specific situation and when not. Straightforward cases when non-verbal behaviours are communicative exist, such as shaking the head to say no, or pointing somewhere while saying 'it's right there'. Non-verbal behaviour uses anatomical elements that are also used for other behaviours (such as swinging a stick), thus introducing a filtering problem, especially as non-verbal behaviours may occur at the same time as action-based behaviours. Non-verbal behaviour also has much more freedom of form and much less structure. It should be noted that while speech can mostly be understood without the non-verbal behaviour (if it has been removed or hidden), the opposite case does not hold – much non-verbal behaviour requires some understanding of the accompanying speech to make sense.

2.3 Classifications of non-verbal behaviour

Non-verbal behaviours include body movements, eye behaviours, facial behaviours, and non-verbal utterances (grunts, etc.), and serve a whole variety of purposes within an interaction. The next section provides in-depth detail on the various types of non-verbal behaviour, their forms, when they occur, and what purposes they serve.

2.3.1. Kinesics

Kinesics includes all body movements that are not performed merely for action-based purposes. Kinesics is one of the most commonly discussed types of non-verbal behaviour, frequently referred to as 'body language'. 'Body language' is in fact a misnomer (Bavelas, 1996), as a language is constrained to a syntax –rules that determine how words or other symbols combine into phrases and sentences – while most kinesic behaviour, or even most non-verbal behaviour generally, does not have this constraint. A small subset of hand gestures behave as symbols (emblematic gestures), but do not have rules to combine them into a more structured meaning, and the addition of grammatical

rules to the symbols creates a sign language. Using a strict meaning of 'verbal' of 'of or concerned with words', sign language is in fact a verbal behaviour, not a non-verbal behaviour. A more clear distinction could be made by using the terms grammatical behaviour and non-grammatical behaviour in place of verbal and non-verbal behaviour respectively, but this terminology is not used.

There is a large cultural variation in kinesic behaviour, so common movements from one culture may be not understood or misinterpreted in another culture (possibly offensively). Kinesics, as with all much non-verbal behaviour, has beauty (or any interpretation) in the eye of the beholder (Hungerford, 1878).

There are a number of classes of kinesic behaviour – self-adaptors, object-adaptors, gesture. Self-adaptors are actions to alter the self and object-adaptors actions to alter objects or the environment. Self-adaptors and object-adaptors are action-based movement, but are frequently intended or interpreted to mean something and so are not *merely* action based movements and are relevant in the context of this thesis. Gesture is difficult to define.

The word 'gesture' and 'gesticulation' are both used in common speech as well as technically. One definition is "*a gesture may be defined as a physical movement of the hands, arms, face, and body with the intent to convey information or meaning*" (Cerney, 2005, p. 29), but 'intent' implies awareness or desire to communicate with the body, when it happens more spontaneously and some non-verbal behaviour certainly occurs without intent. The best definition is a negative one by Ekman and Friesen (1969) – *"[gesture is] all hand movements that are not classified as self-adaptors or object-adaptors*". Self-adaptors and object-adaptors are discussed within the context of gesture.

Gesture

People gesture a great deal, and while speaking they gesture almost constantly (McNeill, 1992) to emphasise or confirm the spoken word and also as word or phrase replacement

('the finger' etc.). Both (Duncan & Fiske, 1977) and (McNeill, 1992) give a gesture frequency of around one gesture per second while speaking.

Gesture is tied closely with the speech accompanying it, both temporally and contextually, and reflects the underlying concept that a person is speaking about (McNeill, 1992) and in fact gesture appears to reflect that underlying concept more accurately than speech. That is, mistakes are more common in speech than in gesture. McNeill gives examples where people are talking about a direction and gesture left, but say 'right', and then correct their speech to match the gesture, so matching with the underlying concept. In McNeill's studies matching with the underlying concept was possible because the subjects were describing something known – the Looney Tunes cartoon 'Canary Row' starring Sylvester and Tweetie Pie (Freleng, 1950).

Cerney (2005, p. 29) states "*gestures may be identified by their function, their linguisticity, and their role in communication*" – see Figure 2-1 below (Cerney, 2005, figure 2-1).

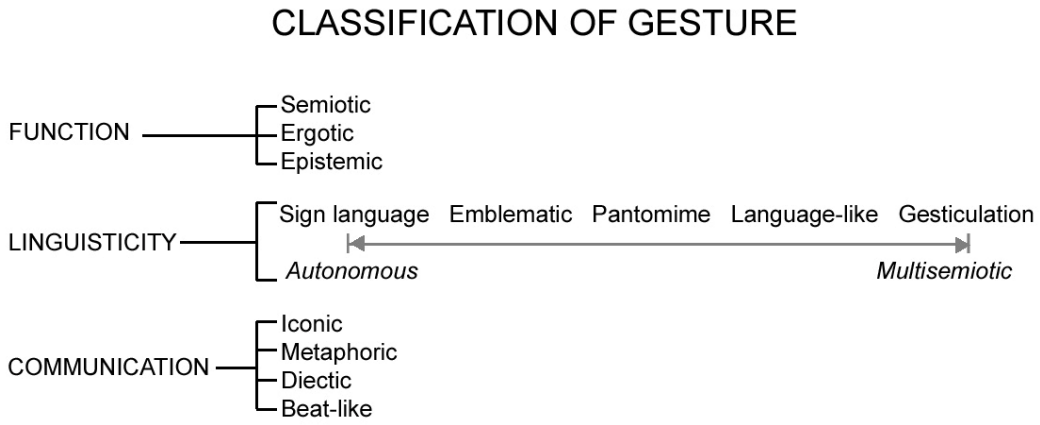


Figure 2-1 Classification of gesture (Cerney, 2005)

Gestures classify functionally into three groups:

Semiotic gesture	Gesture that conveys information, either on its own or in conjunction with other forms of communication (e.g. speech), for example, waving good-bye.
Ergotic gesture	Gesture that manipulates the physical environment, such as opening a door.
Epistemic gesture	Gesture that discovers information about the environment, such as weighing an object by holding it, or feeling the surface to find its texture.

Within the definition of non-verbal behaviour within this thesis, it is only semiotic gesture that is relevant.

Gesture classified linguistically sits on the 'Kendon continuum' – see Figure 2-2 (McNeill, 1992) below. This continuum describes the linguistics properties of gesture along with its degree of conventionality (increasingly from left to right) and whether speech is obligatory with the gesture (decreasingly from left to right).

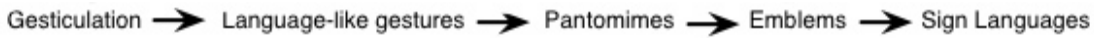


Figure 2-2 Kendon's continuum (McNeill, 1992)

McNeill (McNeill, 1992) writes "*As we move from left to right: the obligatory presence of speech declines; the presence of language properties increases; and idiosyncratic gestures are replaced by socially regulated signs*". Gesture accompanies speech, and according to the (McNeill Lab, 2003) is "*non-conventionalized, is global and synthetic in mode of expression, and lacks language-like properties of its own. The speech with which the gesticulation occurs, in contrast, is conventionalized, segmented and analytic, and is fully possessed of linguistic properties. These two contrasting modes of structuring meaning co-exist in speech and gesture, a fact of profound importance for*

understanding the nature of thought and language in general, and how they function in communication."

Most gesture occurs with speech and this is termed 'spontaneous' gesture – meaning that it occurs spontaneously with speech. Gesture that occurs *without* speech appears much more intentional. Spontaneous gesture is generally made with the head or hands, or if that is not possible then any available body part (or even the whole body), for example, pointing with a foot when one's hands are full. Recent research on spontaneous gesture indicates a link to other cognitive processes, termed 'growth points' (McNeill, 2005; McNeill Lab, 2006). That is, both speech and gesture come from a single underlying conception.

Spontaneous gesture occurs synchronously with speech. That is, each specific gesture occurs with the word that it relates to. Furthermore, the 'stroke' (the semantic, or meaningful, component) of a gesture coincides with the *peak phonological stress* – the most emphasised phoneme – of the speech stream. Spontaneous gesture can be complementary, supplementary, or contrastive to the speech. In other words, gesture can re-iterate or emphasise the speech, add information to the speech, or communicate something contradictory to (or slightly different from) the associated speech.

In contrast to speech, gesture has few constraints on how it is constructed. As McNeill (McNeill, 1992) notes "*the important thing about gestures is that they are* ***not*** *[original emphasis] fixed. They are free and reveal the idiosyncratic imagery of thought*". Gesture is highly context-dependent and often is related to the whole idea rather than to a specific word or syntactic structure.

Another method of classification uses four dichotomies: act-symbol, opacity-transparency, autonomous semiotic-multisemiotic, and centrifugal-centripetal. For more see (Cerney, 2005), or (Nespoulous et al., 1986), the latter being the original source.

Finally, gesture can be categorised by its role in communication and, for the purposes of this thesis, this is the most useful classification. These categories are emblematic, iconic. metaphoric, deictic, emphatic and cohesive.

One gesture can often be placed into multiple categories, or even is a combination of more than one gesture – especially beat-like gestures that often occur overlaid on other gestures – so the borderlines between these categories are grey. This categorisation is due mainly to (McNeill, 1992), but his work was based on that of Efron (1941); Freedman and Hoffman (1967); and Ekman and Friesen (1969), though McNeill often ignores emblematic and cohesive gesture.

Emblematic gesture

Emblematic gestures, or emblems, are simple symbolic gestures, which are culture specific and have a defined meaning within a culture. An example is the 'thumbs up' in Western culture (except in Sicily, where it has a different standard cultural meaning from the rest of Western culture). These semi-standardised gestures are the starting point for development of sign languages. Some emblematic gestures clearly have roots in other forms of non-standardised gesture (usually iconic) and were therefore at some point grounded in the real world, such as the 'come here' gesture, but other no longer appear to have any grounding in the real world and are effectively arbitrary symbols.

Iconic gesture

Iconic gestures are pictorial or animatorial representations of an object, or an action, serving to describe some facet of that object or action. These gestures are grounded concretely in the physical world.

To represent objects or actions though gesture some concept of the object's (or action's) shape, movement, or affordances are required and affect the gesture. This leads to great power and variability among iconic gesture. For example, Alice talking about a teacup in context of drinking tea may perform a gesture of lifting a teacup by its handle, while if

she was discussing the size of the cup, the gesture would be involve some illustration of the cup size, such as 'the cup was *this* big'.

Metaphoric gesture

The counterparts of iconic gestures are metaphoric gestures, which represent abstract concepts or metaphors and are concretised (made into objects) by the gesture. For example, Alice might say 'I had this *great idea*' and inscribe a sphere with her hands; the sphere representing the 'whole idea' concept.

Deictic gesture

Deictic gestures are pointing gestures and refer to something concrete, imagined, recalled, abstract, or temporal. While frequently using the hands, deictic gestures can also use any element of the anatomy or the motion of an element. For example, jerking the head towards an object. Deictic gestures serve to reference an object (possibly abstract) or to specify a referent in speech (such as 'I picked *this* up'). For example, in a conversation Alice may uses a deictic gesture to indicate left when talking about Bob – setting up that space as representing Bob, indicating that in the 'world space' that she has in her mind Bob is on the left. She may also indicate right when talking about Charlie – setting that space for Charlie. Later in speech, Alice can refer to those spaces as with other spontaneous gesture to be complementary, supplementary, or contrastive. So, she may refer to the space on the left while talking about 'him', thus adding supplementary information (the 'him' on the left) and resolving an ambiguity.

Emphatic gesture

Emphatic, beat-like, or 'baton' gesture is gesture providing emphasis. This form of gesture can overlay any other gesture type, or be a simple bi-phase gesture (up/down, left/right, etc.). The emphasis can be on a phoneme, syllable, word, phrase, or section of speech. For example, emphatic gesture can provide the difference between the following two phrases 'I want you to *go* now' and 'I want you to go *now*' (with the transition between phases occurring on the emphasized word). Usually these gestures would also

correspond to emphasis from the vocal stream. Emphatic gesture has little variation in form other than the scale and speed of the phase transition, with larger, faster transitions indicating more emphasis (within an individual). Emphatic gesture can, and frequently does, utilise all body parts, especially the head and hands, but additional movement of more of the body provides further emphasis. This form of spontaneous gesture is distinct from the previous forms in that it can overlay any other gesture as it indexes a section of speech rather than providing semantic content (though it is also used independently).

Cohesive gesture

Cohesive gesture serves to connect related sections of discourse that are temporally separate. It can use any other type of gesture, or just any movement. The cohesion is provided by repetition of the *same* gesture form. For example, when listing items people often provide an emphatic gesture on each item. The emphatic gesture marks each item, while the repetition of the same gesture form connects them together to say 'here's *one*, and *another*, and *another*, and *another*'.

2.3.2. Oculesics

Use of the eyes is an important component of human-human communication. Kendon (1967) identifies four functions of gaze behaviour (in addition to looking at specific items for information gathering), with Knapp and Daly (2002) building on this to classify five functions of gaze:

Regulating the flow of communication

Monitoring feedback

Reflecting cognitive activity

Expressing emotions

Communicating the nature of an interpersonal relationship [added by Knapp and Daly (2002)]

The regulation of communication flow, gazing briefly at another person (specifically at the face) establishes an obligation to interact. Further and longer gazing shows a desire to increase the level of interaction; while decreased and shorter gazing desires a decrease in the level of interaction. Studies using biosensors (skin galvanic response, heart rate) have shown that extended gazes increase general arousal ((Kleinke & Pohlen, 1971; Nichols & Champness, 1971) cited from (Anderson, 1985)), which can lead to highly intense encounters – both positive (intimacy between lovers or between mothers and babies) and negative (aggression between tense parties).

During an interaction eye glances serve as turn-taking signals and also highlight grammatical breaks, conceptual unit breaks, and the ends of utterances (a sequence of speech separated from another by a marked gap), while (as discussed above) the length of gaze shows a desire to change the level of interaction. These glances also allow feedback on the interaction by monitoring the reactions of the other person.

Gaze can also be used to convey some elements of the internal state of a character. Cognitive load (trying to process difficult or complex ideas) can lead both listeners and speakers to look away, the averted gaze reflecting a shift in attention from the external to the internal. There is evidence that the eye gaze direction under this condition changes with different forms of cognitive load, linked to the active hemisphere of the brain (Ehrlichman & Weinberger, 1978; Weisz & Adam, 1993; Wilbur & Roberts-Wilbur, 1985).

Basic emotions such as surprise, fear, disgust, anger, happiness or sadness can be expressed through the eyes, though in fact it is the facial areas around the eyes that displays the emotion, not the eyes themselves (Ekman & Friesen, 1975; Ekman et al., 1971). People are adept at detecting emotional state from the eyes (faces). "*We associate various eye movements with a wide range of human expressions: downward glances are associated with modesty; wide eyes with frankness, wonder, naïveté, or terror*" (Knapp & Daly, 2002). There is some evidence (Hess & Goodwin, 1973) to suggest that people are capable of detecting responses from the actual eyes alone, specifically pupil dilation,

which increases with more aroused states, especially fight or flight responses (Cannon, 1929).

As with pupil dilation, a wide variety of eye behaviour exists that occurs during normal interactions, but there is currently little evidence to suggest this affect the interactant. For example, people tend to disproportionately look at their interactant's right eye with their own right eye (MacDorman et al., 2005; Minato et al., 2005).

Additionally, "*recent findings suggest that perceptual and oculomotor mechanisms that are biased toward the upper field (which disproportionately represents radially distant space) are activated during complex mental operations, ranging from semantic processing to mental arithmetic and memory search*" (Previc et al., 2005) – in other words there exists a relationship between eye movements and cognitive activity (Raine, 1991). It is suggested that higher-order cognition in humans is, in contrast to how it is generally viewed, "*closely entwined with the brain mechanisms mediating more basic perceptual-motor interactions*" (Previc et al., 2005). In practice, this means that a variety of motor actions occurs with higher-order cognition and these movements may, in fact, assist in the cognition. This latter point also adds support in relation to gesticulation, for which there is evidence indicating that gestures assist in word recall and speech flow, and the disruption of the ability to gesture disrupts speech flow and increases error rates (McNeill, 1992).

Finally, eye gaze can also communicate the nature of an interpersonal relationship. Gazing and mutual gazing is found most in conversations. When interacting with a very high-status addressee moderate mutual gaze occurs, while maximal mutual gazing occurs when interacting with a moderately high-status addressee, and is minimal with a very low-status addressee (Efran, 1968; Hearn, 1957).

2.3.3. Proxemics

Proxemics is the use and arrangement of the self in the physical world – "*...the study of man's transactions as he perceives and uses intimate, personal, social and public space*

in various settings while following out of awareness dictates of cultural paradigms" (Hall, E. T., 1974). Hall (1966) also describes a set of measurable distances called 'reaction bubbles' between people as they interact (in US/UK Culture) as:

Public speaking	12 feet or more	4m
Conversation among acquaintances	4-12 feet	1-4m
Conversations among good friends	1.5-4 feet	50-100cm
Embracing or whispering	6-18 inches	15-50cm

These distances vary significantly between cultures. Cultures with lower population densities, or those where individualism or privacy are highly important tend to have larger distances for the set of reaction bubbles. In cultures where the reverse is true, maintaining these larger distances can be taken as unfriendly or rude, although the distances for the reaction bubbles vary between cultures the same set of instances still occur. Interactions that are so close as to be touching cross over in to the discussion of haptics, although note that these distances apply to standing conversation-type interactions.

Proxemics is closely related to the idea of territories in human sociological behaviour, including, in addition to the staking out or ownership of an area of land, objects, relationships, jobs, schools, abstract and symbolic objects and ideas such as religion, value systems, and includes abstract spaces such as the space around a person. Violation of appropriate personal space has powerful responses similar as with other territorial violations, and can have serious adverse consequences on an interaction. "*...it seems we are forever conscious of our intimate zone and its violations. Examples: the butler who doesn't listen to the conversations of the guests, the pedestrian who avoids staring at an embracing couple, or the person who becomes preoccupied with a magazine during another's nearby telephone conversation. They all show some awareness of communication property rights and will adjust both their body language and proxemics to relay that message.*" (Katie, 1997)

2.3.4. Haptics

Haptic behaviour (also known as tacesic or touching behaviour) is behaviour using the touch (or the lack therefore) and can considered a proxemic behaviour – "*body contact and touching are proxemic phenomena*" (Harper et al., 1978). Touch may be the most basic or primal form of communication and strongly conveys aspects of basic states – love, affection, hostility, anger, presence. The absence of touch can also be a strong signal. Haptic behaviour can be approximately categorised into five levels of intimacy (Heslin, 1974):

Functional/professional

Social/polite

Friendship/warmth

Love/intimacy

Sexual arousal

One of the most widespread haptic behaviours is the hand shake, but this comes with large variation across cultures and levels of intimacy. Haptic behaviours are more common in some cultures than others. Remland and Jones (1995) found that touching while communicating was relatively rare in some countries (England (8%), France (5%) and the Netherlands (4%) compared to other countries (Italy (14%) and Greek (12.5%)).

Jones and Yarbrough (1985) determined seven types of touch, with a total of 18 different meanings:

Positive affect	Express positive emotions, with meanings of support, appreciation, inclusion, sexual interest or intent, and affection.
Playfulness	Signals to make an interaction less serious. This includes both playful affection and playful aggression.

Control	Attempts to influence the behaviour, attitude of state of another in the form of compliance, attention-getting, or announcing a response.
Ritual	Easing transitions in (greeting) and out (departure) of interactions.
Hybrid (mixed)	Combinations of other touches, for example, affectionate greeting.
Task-related	Directly associated with performing a task in the forms of reference to appearance (or simple reference), instrumental ancillary (doesn't assist in task), instrumental intrinsic (assist with task).
Accidental touch	Unintentional and without meaning.

Inevitably the lines between categories of touch are somewhat vague and many touches may fall into different categories. Furthermore, perception of touch is highly variable and in the same way a proxemic behaviour, violations (or perceived violations) will generate negative responses. For example, an accidental touch may be perceived as a positive affect, possibly sexual, generating a negative behaviour in response. Touch is a powerful form of non-verbal behaviour, but with that power comes the risk of negative effects or responses – "*...in power is also joined an awe-inspiring accountability to the future"* (Churchill, 1946).

2.3.5. Paralanguage

Trager (1958) defines paralanguage as "*elements of vocalization not typically included in the phonological description of language. e.g. intensity (stress), duration of syllable, laughing, uh-huh uh-uh*". In other words, paralanguage refers to the elements of vocal behaviour other than the specific words that are spoken. Paralanguage includes pitch,

volume, speed, rhythm, intonation, along with non-word vocalisations and is an inevitable aspect of all speech.

Other than non-word vocalisations, these aspects of speech are also called prosody – relating to: changes in syllable length, loudness, pitch, and formant structure of speech sounds (acoustic changes); changes in velocity and range of motion of articulators such as the jaw and tongue, and changes of quantities such as air pressure in the trachea and tensions in the laryngeal muscles (speech articulator changes); and changes in rhythm, tone, intonation, and lexical stress (phonological changes).

Prosody is an important facet of speech and has been demonstrated to correlate with both verbal behaviour and other forms of non-verbal behaviour, such as head nods (Munhall et al., 2004) and gesture (Loehr, 2004). Prosody can assist the verbal stream in a lexical manner, such as providing emphasis or accents to words or syllables. Typically the perceived pitch of speech (f0 – the fundamental frequency) peaks in amplitude on the stressed syllable. Prosody also serves in a non-lexical manner to, for example, change sentences from declarative to questions by raising pitch towards to utterance. Prosody provides a discourse function by emphasising new information or topics and can provide other, more complex, discourse function, such as a person being sarcastic, ironic, caustic, satirical, or sardonic. These more complex forms are not purely prosodic behaviours and it has been shown, for example, that "*prosody alone is not sufficient to discern whether a speaker is being sarcastic*" (Tepperman et al., 2006).

Prosody can also convey emotions – "*emotional arousal affects a number of (relatively) easily observed behaviors* [sic]*, including speech speed and amplitude*" (Ball & Breese, 1998). In general, increased arousal increases both the speed and amplitude of speech. Finally, prosody reflects the underlying physical system, so, for example, prosody strongly indicates gender. It is suggested that prosodic signals are evolved patterns, rather than learned conventions, due to little evidence for either personal idiosyncrasies or cultural differences (Frick, 1985).

2.3.6. Olfactory

Olfactory communication is highly important in nonhuman animals – communicating emotional states through changes in body odour, but it is not clear that this is also true for humans. It has been shown that people can identify odours of people through the smell of swabs from them when a variety of emotions are induced (Chen & Haviland-Jones, 2000), but studies have not shown that people can obtain this information at the distances of normal interactions, or how fast changes in emotional states can be detected. It does seem that even if olfactory communication does occur in humans it is not as powerful as in nonhuman animals.

2.3.7. Observed behaviour

Observations of people's behaviour when not interacting with them can provide important information about those people and affect future behaviour. For example, witnessing a person commit some violent act would instil more caution than if that person had been doing something less fear inducing. This is true even with less obvious behaviours such as mere conversation – it has been shown that ordinary people listening to 20-second sound clips of doctor-patient conversations can strongly predict whether those doctors (surgeons) will be sued for malpractice, even when the frequencies that make speech intelligible are removed (leaving in the prosodic elements) (Ambady et al., 2002). This concept also applies within interactions and is frequently termed 'thin-slicing' – "*the ability of our unconscious to find patterns in situations and people based on very narrow slices of experience*" (Gladwell, 2005).

2.3.8. Chronemics

The effect of time in non-verbal behaviour is termed chronemics and involves the way time is perceived, structured, and the reaction time can cause. Time behaviours, such as punctuality and willingness to wait, provide information about an interactant at a high level. As such, a person who is consistently late is providing signals that could indicate a lack of value of a meeting. Time affects can also be overlaid on other verbal and non-

verbal behaviours to provide local context. For example, long gazes indicate increased arousal or a desire to increase arousal.

2.4 Spatial-task context

Communication can be considered to occur in four different task contexts: cooperation, coaction, competition and conversation (Knapp & Daly, 2002). In other words, communication occurs in order for some number of parties to: perform a task together (cooperation), to exist in the same vicinity (coaction), to perform a task at the expense of another (competition), or to entertain or pass on information (conversation). Communication and the forms thereof vary across these different contexts and also with the physical proximity of the communicating parties. Non-verbal behaviour provides information as to the beliefs, desires, and intentions of another person, or alternatively it can be considered as providing indicators as to that person's cognitive, emotional, physical, intentional, attentional, perceptual, interactional and social status. The set of non-verbal behaviours used varies distinctly across both the task context spectrum and over spatial distance, creating a spatial-task context as illustrated in Figure 2-3. Computer games provide a good illustrative example because interactions between characters and between characters and the player occur across the full range of spatial-task context. Within a game scenario non-verbal behaviour can be modulated by the spatial-task context. Examples of the spatial-task context are shown in Figure 2-4 using screenshots from Half-Life 2 (Valve Corporation, 2004), though it should be noted that at present computer games do not have a concept of spatial-task context.

Figure 2-3 maps out the range of spatial and task context for these examples. In conversations the movement of the other conversational party (both body and face) is visible in detail and furthermore, people are highly attuned to interactions in intimate, personal, and social spaces and are sensitive to many subtle cues and nuances in non-verbal behaviour. At further distances less detail of a person's behaviour is apparent. There is a significant transition in non-verbal behaviour from situations where intimate

verbal communication is possible to those where it is not. The sensitivity of non-verbal behaviour to proximity is due to a number of factors, including the more public nature of non-verbal gesture in open spaces, and the requirement on particular physical behaviour to carry the full communicational load (e.g. subtleties in gaze and facial expression are not visible at a distance).

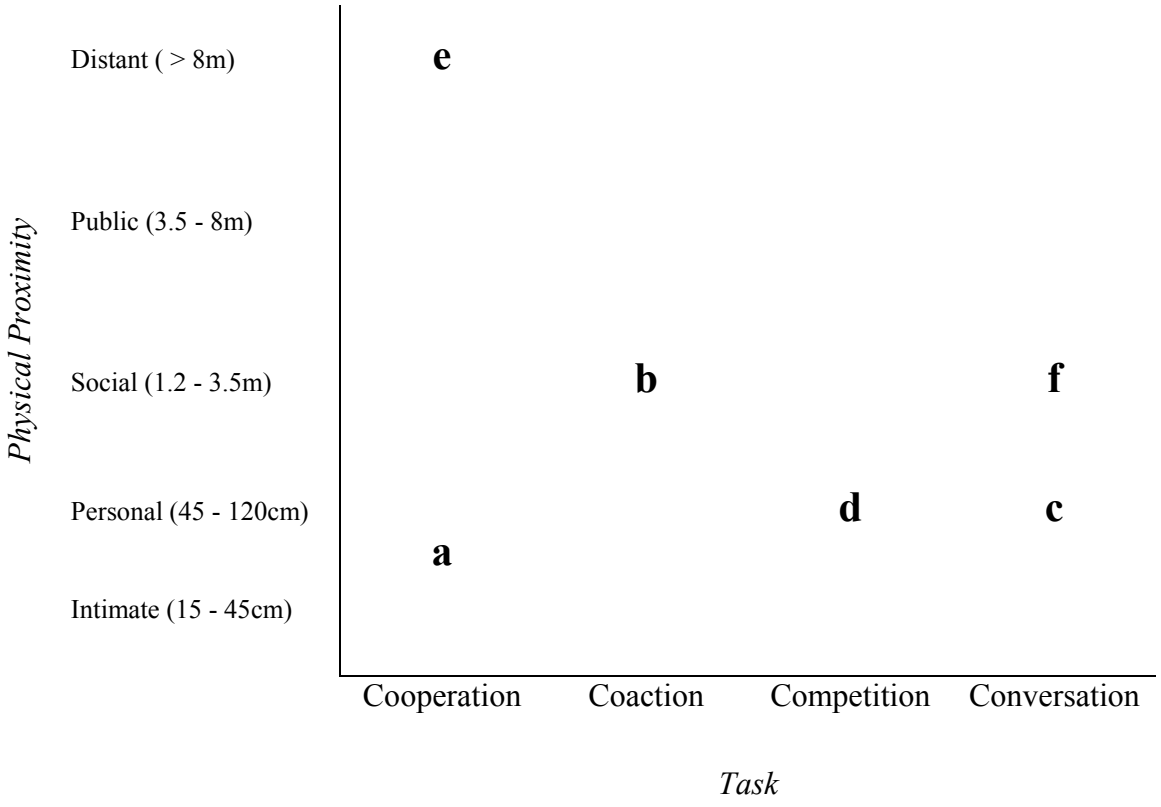


Figure 2-3 Task and spatial context

Figure 2-4(a) shows an example of cooperation in intimate space. The male character demonstrates his attentional state – that he is attending to the female character – with his body orientation, face orientation, and gaze direction. Of course, people are rarely static, but different non-verbal channels (e.g. face orientation, body orientation, gaze direction, body position) are closely coordinated in demonstrating attention. Thus, the male character could look away but still communicate his attention sufficiently through his body haptic and proxemic behaviour. In an interaction between unfamiliar subjects,

however, strong or constant facing or looking at a person is widely considered an aggressive signal. It is considered rude, or at least off-putting (Knapp & Daly, 2002). Figure 2-4(a) also illustrates non-verbal behaviour using facial expressions and kinaesthetic (touching) behaviour.

Figure 2-4(e) illustrates a situation at the other end of the spatial scale, cooperation at a distance between the player and a non-player character (in fact, navigation and negotiation, a subset of cooperation). The non-player character shown and the player will collide if they do not arrive at an agreement as to how to pass one another and communicate this – the characters must cooperate through the use of non-verbal behaviour to resolve a potential conflict. In the real world, people in this situation use a range of subtle non-verbal mechanisms such as gaze and body turning to initiate and mutually negotiate space. Non-player characters in Half-Life 2 will avoid the player, but will not exhibit non-verbal behaviour in doing so and simply move around the players as they approach. Without non-verbal behaviour it is difficult for players to decide which way to move out of the way (indeed they do not need to) and it is this absence of social conventions (and the ability to break them, to invite conflict) that both undermines the engagement of players with the game and limits their expressivity.

a – cooperation in intimate space

b – coaction in social space

c – conversation in personal space

d – competition in personal space

e – cooperation at a distance

f – conversation in social space

Figure 2-4 Half-Life 2 task and spatial context examples (Valve Corporation, 2004)

Between the proximal and distant spatial scales are social spaces, Figure 2-4(f) is an example of conversation in a social space. Here non-verbal behaviour facilitates a number of aspects of the interaction (and the dialogue in particular) including the mediation of conversation flow, such as whose turn it is to speak (interactional state). Turn-taking mediation is a complex coordination of behaviours, but in simple terms speakers provide opportunities to allow the listener to take a turn (such as, a slightly prolonged pause, or a look up into the eyes), at which point other listeners can, if they choose, take a turn. If not, then the speaker will continue. Additionally, others can

indicate that they might like to speak, with signals such as increased eye contact, leaning forward or standing taller (Duncan & Fiske, 1977). Turn-taking mediation is not required in Half-Life 2 because the game developers have not allowed the player to speak, but it is potentially a very important component of computer systems that hope to include natural language interactions (particularly spoken interaction) between real people and characters.

Finally, Figure 2-4(b, c, and d) illustrate the remaining task contexts: coaction, conversation and cooperation, and competition. Characters sharing the same approximate area of space engage in coaction behaviour, corresponding to mutual monitoring – this can be interpreted as communication by virtue of the fact that watching a character implies that you might react to it – that is, there is an implied reason (intention) for watching. Coaction can be considered the default task context, which develops into the other contexts. Competition contexts give rise to distinctly different forms of non-verbal behaviour from other contexts, but these still serve to communicate internal states. In Figure 2-4(d) the raised baton serves to communicate 'you have crossed a line – back off or I will hit you'.

2.5 Managing interaction

Non-verbal behaviour also plays a major role in managing an interaction. For an interaction to 'work', different parties have to take turns speaking, as people find it almost impossible to listen and talk at the same time, and non-verbal behaviour helps to mediate who should speak when. These are known as turn-taking behaviours.

Sacks, Schegloff, and Jefferson (1974) found that "*overwhelmingly, one party talks at a time, though speakers change, and though the size of turns and ordering of turns vary; that transitions are finely coordinated; that techniques are used for allocating turns ... and that there are techniques for the construction of utterances relevant to their turn status, which bear on the coordination of transfer and on the allocation of speakership"*. They also observe, among other things, that "*occurrences of more than one speaker at a*

time are common, but brief"; "*transitions (from one turn to a next) with no gap and no overlap are common*"; and that "*repair mechanisms exist for dealing with turn-taking errors and violations; e.g. if two parties find themselves talking at the same time, one of them will stop prematurely, thus repairing the trouble*".

Sacks, Schegloff, and Jefferson liken turn-taking behaviours to an economic model wherein "*turns are valued, sought, or avoided. The social organization of turn-taking distributes turns among parties*" and suggest that this organization "*will affect the relative distribution of turns among parties*". Furthermore, Sacks et al. provide a set of rules "*governing turn construction, providing for the allocation of a next turn to one party and coordinating transfer so as to minimize gap and overlap*". Turn taking behaviour is universal, occurring in all known languages and cultures, between parents and infant, and within sign-language communities. Another model of turn-taking and the associated verbal and non-verbal behaviours is to view the conversation timing patterns as governed by "*endogenous oscillators in the brains of the speaker and the listeners*" that "*become mutually entrained on the basis of the speaker's rate of syllable production. This entrained cyclic pattern governs the potential for initiating speech at any given instant for the speaker and also for the listeners (as potential next speakers). Furthermore, the readiness functions of the listeners are counterphased with that of the speaker, minimizing the likelihood of simultaneous starts by a listener and the previous speaker*" (Wilson & Wilson, 2005). In other words, the patterns of turn-taking are cyclic patterns, with each interactant with their own internal representations of the point within those patterns, with the verbal and non-verbal behaviours serving to bring and keep together those internal representations.

Non-verbal behaviour is important not only for specific purposes, such as managing interaction, but also for providing engagement and realism. The complexities of non-verbal behaviour allow much expressive power and an increase in realism, engagement and affective purposeful behaviour and, while highly challenging, building ECAs with non-verbal behaviour is worth the challenge.

3. Embodied Conversational Agents (ECAs)

In developing new ECA technologies it is important to have an understanding of what ECAs are, how they are built, their present abilities and how they can be evaluated. Starting with a brief overview of ECAs and how they may have social influence over people, this chapter defines ECAs and gives some examples of where they occur, followed by a discussion of the historical focus of ECAs on goal-based abilities and ECAs being built as 'deliberative' systems focused on text processing. An overview is given of the present state of the art in ECAs noting the attention that has recently begun to be given to creating non-verbal behaviour for ECAs, and also the difficulties that arise in developing ECAs due to groups working with ECAs generally building their own in-house ECAs.

The new challenges introduced by both the complexities of non-verbal behaviour and the volume of incoming data in a non-verbal stream are then covered. General approaches to evaluating ECAs and the difficulty of such evaluation are considered, followed by some past evaluation methods used for ECAS. ECAs involve a large set of disciplines and there has been an obvious desire to built complete agents rather than agents implementing only certain aspects or components, but with the complexities involved this has inevitably led to the development of agents which are highly functional in some areas, while highly limited in others. The difficulties of establishing criteria for evaluating ECAs are discussed as is the difficulty of comparing different ECAs. Many previous evaluations of ECAs have had a predominant focus on the users' experience, with limited attention given the behaviour and performance of the ECA, and have been extremely thin on objective, empirical methods. Previous studies evaluating ECAs are critiqued, along with discussing those which use more solid scientific approaches.

Given the importance of non-verbal behaviour in human-human interactions as discussed in Chapter 2, one would assume that non-verbal behaviour would also be important in human-ECA interactions. It has previously been shown by Reeves and Nass (1996) that people tend to "*treat computers, television, and new media like real people*", and as ECAs are a form of media, that would imply that people also relate to ECAs as real people, or 'social actors' – again, people or things that perform social actions within

an interpretive sociological perspective (Weber, 1978). In other words: people tend to be polite to computers; can view computers as team-mates and may respond to praise from them; usually like computers which have personalities similar to their own; more often describe masculine-sounding computers as extroverted, driven and intelligent; and expect feminine-sounding computers to be more knowledgeable about love and relationships.

It is not entirely clear why people treat various forms of media as real people. It may be that the complexity of interaction with these media forms requires people to use an internal model of similar complexity, and the most readily available model on which to draw is that of real social actors (people). Alternatively, the view could be taken that while the high-level cognitive components of the human brain are aware that these forms of media are not real people, the lower-level components – R-complex/Reptilian Brain and the Limbic System/mammalian brain – cannot make this distinction. As those components of human brains guide much of human behaviour, it is reasonable to expect much the same sort of behaviour between real people and complex media. In the case of ECAs, we would expect this even more so as ECAs are even closer to real people.

It has been discussed previously that communicating with ECAs is in many circumstances easier and more natural for people than communicating with computers in other ways. Interaction with (high-fidelity) ECAs would require no additional learning of interaction techniques and would be highly efficient, though ECAs are not without their drawbacks as an interface. It should be noted that interactions with present-day ECAs do require additional interactional learning due to the limitations of those ECAs. In order for interactions with ECAs to proceed effectively, an ECA needs to understand the verbal and non-verbal cues that people it interacts with portray (and may portray), and also to generate appropriate verbal and non-verbal cues in response.

The development of ECAs both within game scenarios and within more academic/business pursuits (such as information and advice agents) has maintained a predominant focus on the specific goals of that scenario. For example, non-player

characters in games have focused on high-level coordination of behaviours to achieve game goals such as assisting with or competing against a player's in-game goals, and ECAs giving information, such as directions, have mainly been focused on achieving those goals. There has been less focus on the social interactions of ECAs with both users and/or other non-player characters/ECAs. This focus on developing ECAs to achieve specific goals has strongly influenced the structure and design of those agents. Specifically, such ECAs have relatively well-developed high-level cognitive behaviours (goal-oriented behaviours), but much less well-developed low-level (simple/social) behaviours, especially non-verbal behaviours. Given the previous discussions of the importance of non-verbal behaviour within human-human interactions, present ECAs are missing some important aspects. That said; present day ECAs are still highly complex and have highly effective behaviours, both in real world and game scenarios.

3.1 Anatomy of an ECA

An ECA is a complex system involving many different interacting components. ECAs have some form of 'embodiment'. This is most frequently a graphical representation of a human (or non-human), but can also be a physical representation – a (possibly humanoid) robot. The conversational component of an ECA usually involves generated speech (through speech synthesis or speech splicing), though it can also include text-based speech output. The words spoken (or displayed) may be generated dynamically or statically from some form of lookup. Finally, the ECA must have some form of agency. That is, it must have some kind of input that has some significant influence over the conversational behaviour (or output) of the embodied character. The non-player characters in computer games can only sometimes count as ECAs, as usually they fail to fulfil the second or the third criteria at the same time – no speech, or speech is not significantly responsive to external inputs such as player characters (such as within cut-scenes where a player's behaviour doesn't significantly affect the dialogue).

Present ECAs can be classed in classical artificial intelligence (AI) terms as deliberative systems. Minsky (Minsky, 2006) suggests that a 'deliberative' system has the ability to select the best action or behaviour from a set of alternatives – which has long been studied in the theory of games and decisions. He opens this further by including options in a payoff matrix, or values in a continuous interval. More generally, deliberative systems have a straightforward sense, process, act cycle – the system takes in a set of inputs (sense), processes this data to determine an appropriate action, then performs that action. It is this form of intelligence that has been the focus of AI for decades – the high-level cognitive intelligence. So the 'process' component of present ECAs is mostly focused on high-level symbolic processing, and generally takes input in a highly limited form – usually purely text input. We should note at this point that non-player characters in games tend not to be deliberative systems – they tend to be reactive systems where each behaviour is a simple reaction from (simple) inputs. More complex behaviours are usually guided by pre-computed solutions. For example, non-player characters do not usually perform route planning (deliberative), but merely look up a route from a pre-computed solution.

Within the deliberative realm, the focus of ECAs has been on conversation, specifically to understand text or speech input, and to generate natural language responses. Natural language processing (NLP), as this is known, is a relatively mature field within computer science and is effective within constrained domains – usually where complete, grammatically correct sentences are used. NLP tends to struggle significantly more in more open domains and with more natural speech – incomplete sentences, corrected sentences, ambiguous sentences, paralanguage, etc.

Both understanding and generating non-verbal behaviour is a much less mature field within computer science. This immaturity is due historically to computers being restricted to predominantly text-based input and output (through a keyboard and screen) and the dramatically increased data input size and complexity involved with non-verbal behaviour input (audio streams, video streams, motion data) and output (computer graphic characters). It is only recently that computing technologies have advanced

sufficiently for all the required components for building an ECA to exist, and combining all these together is state-of-the-art. Much focus has been put on developing and implementing mark-up languages and transducers (XML or otherwise) in order to define what non-verbal behaviour should occur (Cassell, Justine et al., 2001; DeCarlo et al., 2004; Kranstedt et al., 2002; Noot & Ruttkay, 2003). While there has also been development on gesture generation and connection to speech (Kopp et al., 2004), 'gesture understanding' has also been given significant focus as a form of interaction, though usually on using gesture as a form of (explicit) control, rather than as an aspect of normal conversation such as with ECAs.

There are few standard approaches to building ECAs (though the XML approaches are attempting to help with this). Each group working with ECAs has generally built its own in-house ECA with strengths and weaknesses in various areas according to the targets of the research group. With this in mind, the main focus of research using ECAs has been on simply trying to build an ECA with some of the required abilities. Much less attention has been given to evaluating the performance of ECAs once they are built.

3.2 Evaluating ECAs

Up until recently the evaluation of an ECA has predominantly been as simple as 'Was it built?' because of the large challenges involved in achieving just that. However, in order to establish ECAs in useful roles, some more significant forms of evaluation are required. Evaluation is important both to determine if one ECA is better than another for a specific role, and also in order to guide the future development of ECAs – it is not clear in which ways or areas present-day ECAs do particularly well or badly in interactions with real people, and it is not clear how 'good' ECAs could potentially be in the future.

Evaluation is inevitably largely dependent upon the roles in which an ECA is envisaged, though some evaluation can be performed independently of the role an ECA is built for. To date few methodologies for evaluating ECAs have been presented, and furthermore,

those evaluations that have been presented are predominantly post-interaction methodologies, as discussed in more depth in section 3.3. Methodologies that can provide evaluations during conversation would provide additional and powerful information.

Only recently have ECAs begun to fulfil their promise of providing useful roles or services to people, be that in computer games, sales environments, education, or other areas. As previously stated, the challenges of designing and building ECAs have meant that research focus has been mainly "*on some specific problems which are prerequisites for developing full-fledged multimodal ECAs*" (Ruttkay & Pelachaud, 2004) with less focus on evaluating full systems. As Ruttkay and Pelachaud go on to state "*the evaluation of single modalities often cannot be done without taking into account the (unwanted) influence of other modalities*", and even now ECAs are limited in their use of the full set of modalities that humans routinely use. Furthermore, evaluation is complicated, as each implementation of an ECA is made for a specific role and as such not easily comparable to others. The complexity of human interactions and differences among people, their ways of behaving, their subjective values, and many other factors make any evaluation highly challenging, even without the limitations and non-comparability of present ECAs.

More fully-fledged ECAs have been developed (André et al., 1998; Badler, 1997; Hayes-Roth et al., 1996; Isbister et al., 2000; Stone & Lester, 1996; Trappl & Petta, 1997) but still the focus of even full system development has been on specific limited areas of an ECA rather than its full behaviour. To evaluate these limited ECAs, methodologies have inevitably been tuned to the positive characteristics of each particular ECA and as a result such approaches to evaluation do not extend well to ECAs in general.

In the development of highly functional ECAs one must also pay due notice to research on humanoid robots that indicates that as ECAs become more visually realistic, they may encounter a so-called 'uncanny valley' (Mori, 1970), where users' acceptance of

ECAs drops significantly as the visual realism approaches a level that is indistinguishable from an actual person. That is, as a humanoid robot, or an ECA, approaches the visual realism of a human, people then judge it as a real social agent (rather than as computational agent) and critique it as that – the ECA is a real social actor that is exhibiting subtly strange behaviour and/or strange visual characteristics/abnormalities.

For ECAs to be effective in their target environment they need people to treat them like real people, and methodologies to measure the extent to which people consider them as social entities will help in this development. Human-human interactions follow many conventions (within and across cultures, gender, ages, social hierarchies, etc.) and these conventions lead to social contracts and breaking these contracts is taboo (though, of course, that does not mean it does not happen).

3.3 Existing ECA evaluation methods

As discussed above, evaluating ECAs is hard – for a myriad of reasons including the following (Isbister & Doyle, 2004):

ECAs are highly complex – they aim to have human levels of behaviour and interactivity, and therefore inherit human levels of complexity.

ECA development builds on an extremely large set of disciplines and research areas including: agents architectures, artificial intelligence, synthetic speech, natural language processing, motions, interface design, sociology, psychology, art, drama, and animation.

The obvious desire to build 'complete' agents rather than agents implementing just certain components or aspects of humans, but

> inevitably due to the complexities outlined above and limited development resources these agents are highly functional in some areas, but highly limited in others – "*one system may have excellent facial animation; another a flexible emotional model; a third may be adept at handling social interactions*".

It is difficult to even establish criteria for the evaluation of ECAs – "*there are no formal, widely-accepted definitions of core terms such as believable, social, or even conversational*".

As mentioned previously, most ECAs have been developed for a specific purpose, or for a specific research area, and each evaluation has inevitably focused on that specific target area, leading to evaluations that are not comparable across a variety of ECAs. With all the above in mind, it is understandable that evaluation methods for ECAs are still in their infancy and easy to see why many methods to date have been limited. For a more detailed discussion of empirical studies conducted to evaluate ECAs, see work by both Dehn (Dehn & Mulken, 2000) and Ruttkay (Ruttkay & Pelachaud, 2004).

Evaluations of ECAs must focus on the users' experience, behaviour, and performance while interacting with the ECA. The predominant focus of most evaluations has been only on the users' experience, with limited attention given to their behaviour and performance, and many are thin on empirical, scientific methods. Many existing studies use limited empirical approaches that fail either to identify objectively measurable variables or to adequately explore the impact of the low level of functionality of the agents on the study. For example, Bernsen and Dybkjaer (2004) merely gather subjective data of users' perceptions of interactions with an ECA through structured interviews and presents this data using conversational analysis. The analyses were not tested for inter-rater consistency and no quantitative metrics were taken or calculated. Furthermore, only a small number of experiments were performed and multiple variables were varied across control conditions. Overall, this makes the conclusions distinctly

untrustworthy and of limited use. This level of evaluation is not uncommon at present. Studies such as those performed by Abbattista, Lops, Semeraro, Andersen and Andersen (2002) show more promise, with both quantitative and qualitative measures taken, but unfortunately the analyses presented focus on the qualitative measures with no results given from the quantitative measures. The qualitative measures were taken using the usual combination of interviews and questionnaires – measuring the users' subjective measure or descriptions of the experience. Research by Rickenberg and Reeves (2000) demonstrates that more empirical methods can be used effectively with both subjective and objective measures. Data was obtained through well controlled experiments with subjects matched across conditions and standard psychological scales used to measure anxiety. These studies also, in contrast to those previously mentioned, measured task performance. The data was subjected to thorough statistical analysis giving results that are both reliable and repeatable. Bente (Bente, Krämer, Petersen et al., 2001; Bente, Krämer, Trogemann et al., 2001) builds on this work to create the 'Development and Evaluation Platform for Animated Characters' (DEPAC), where "*systematic variations of specific non-verbal cues can be incorporated to test their particular effects on person perception and impression formation*" (Bente, Krämer, Trogemann et al., 2001).

Direct objective measures of subjects' behaviour and reactions have been taken in a variety of studies (Bers, 1996; Cassell, J. et al., 1999; Essa, 1995; Grammer et al., 1997; Thorisson, 1996) to inform the development of ECAs in a general way, but only a very few studies have used direct objective measures to evaluate ECAs for how well they perform.

4. Persuasive potential of ECAs: introducing synthetic ECAs

The potential social influence of ECAs is largely unknown. Present ECAs have limited abilities, and limited social influence. Establishing the potential social influence of ECAs beyond those which a presently able to be build provides motivation to build more sophisticated ECAs. The concept of synthetic ECAs was developed for this purpose. A synthetic ECA appears to be a real ECA, but is in fact audio and video of a real human transformed to give the appearance of an ECA. In other words, a synthetic ECA is a synthetic synthetic human – it pretends to be a thing that pretends to be a human, similar to a play within a play, or an actor portraying another actor. This enables the evaluation of the potential social influence of highly sophisticated ECAs. This chapter introduces synthetic ECAs in more depth, along with their use to determine the *persuasive* potential of ECAs. The reasons for using persuasion as an evaluation measure where an ECA is acting as a service agent to bring about behaviour change is discussed, while keeping in mind that other evaluation measures could be used in that and, more importantly, in other contexts. Persuasion is also given as an example of an evaluation approach usable across different ECAs – it is not dependent on the specific manner in which an ECA is built. The question of the persuasive potential of ECAs or how persuasive an ECA could ultimately be is introduced and compared to the limited persuasiveness of present ECAs.

A specific scenario of an ECA discussing a charity and charitable giving and then providing an opportunity to donate money to that charity is introduced as a evaluation metric to determine the persuasive potential of an ECA (within a specific context) and to elucidate how important some aspects of non-verbal behaviour are for a persuasive effect. Previous work studying the social influence of ECAs is discussed and critiqued, along with the advantages that using synthetic ECAs may introduce for determining directions for research.

How the synthetic ECA was implemented is discussed, showing previous work cartoonising video, and introducing the approach taken in this thesis. The fact that the synthetic ECA can be used for further research by other groups without significant expense or complexity due to it using only consumer hardware is also highlighted. The recent (since studies were performed) availability of cartoonising functions in software

packages is recognised and their advantages over the approach used in this work are noted – namely they are simpler and more generic. As the synthetic ECA is not a real ECA, it was necessary to verify that people did nevertheless believe it to be a real ECA. The reasons for this are discussed in more detail, along with details on the study to provide this verification.

Within this thesis persuasiveness is used as an empirical and objective evaluation measure for ECAs. The level of persuasiveness of an ECA is by no means the only evaluation measure that could be used, but it is an appropriate one for the specific context where ECAs may be used as service agents to effect behaviour change, and persuasive effects are common in human-human interactions. Behaviour change is, in fact, the real measure of persuasion, and people try to effect behaviour change in others with many of their normal interactions. ECAs with the capability to persuade real people to change their real behaviour would have significant value both over other ECAs and in general. The set of arenas where persuasive ECAs could have value includes: service agents (agents providing advice, information, guidance, or education on specific tasks or areas); in-game agents (agents that persuade game players to interact and value them, leading to the development of more complex games); advertising agents (agents that persuade people to buy specific products).

It should be noted that though this thesis has a focus on persuasiveness as the evaluation metric of ECAs within empirical studies, other evaluation metrics of ECAs are important, such as subjective perceptions of ECAs. Persuasion is used as one example of an evaluation metric that can be used across a variety of different ECAs. Other evaluation metrics include concepts such as believability, engagement, trust, realism, intelligence, use for specific task(s), friendliness, beauty, and many more. Generally, this is the same set of evaluation metrics that people may apply to real people and/or tools.

When considering persuasive ECAs, the questions immediately arise: “How persuasive could an ECA be?” and “how persuasive can an ECA be compared to a real person?”. The term ‘persuasive potential’ is employed to mean how persuasive an ECA *could*

ultimately become – in other words, the potential of ECAs to be persuasive. As shown earlier the behaviour of present-day ECAs is limited in comparison to real humans, especially with respect to non-verbal behaviour, and consequently it would be expected that the persuasiveness of present-day ECAs would also be limited – not achieving their full *persuasive potential*.

Previous research by Bailenson and Yee (Bailenson & Yee, 2005) indicates that ECAs have social influence, and Reeves and Nass (Reeves & Nass, 1996) have shown more generally that computer interfaces (such as ECAs) are treated as social actors. That is, for the most part people treat computer interfaces as real people – for example, people like it when computer interfaces compliment them; people like compliments, even when they know the computer is lying; people like computer interfaces that compliment other people or other computer interfaces. These results, as stated by Reeves and Nass, are the same as for real people – people like it when other people compliment them; people like compliments even when they know the other person is lying, etc.

ECAs that can intentionally persuade humans, or effect behaviour change, raise important ethical issues which are beyond the scope of this enquiry.

4.1 Empirical evaluation of persuasive potential

With the focus on non-verbal behaviour within interactions, two questions are posed:

Within a specific context, how persuasive can a specific ECA be compared to a real person?

What role does non-verbal behaviour play in the persuasive effect? Specifically, within the context, if there is no non-verbal behaviour, does this affect the persuasive effect, and is the link between the non-verbal behaviour of the two parties important for the persuasive effect?

A specific context is used merely to restrict the scenarios sufficiently to allow for evaluation, and it is presumed that results within that specific context may be extrapolated for other contexts, or rather, provide a foundation to demonstrate similar results in other, more general, contexts. The evaluation of the role of non-verbal behaviour is important to determine whether ECAs in the future should use non-verbal behaviour and more importantly if this non-verbal behaviour should be linked to the non-verbal behaviour of the ECAs interactant person.

As discussed in Chapter 2 non-verbal behaviour in human-human interactions is highly complex, and also highly important. Furthermore, non-verbal behaviour in human-human interaction has a strong linkage, or is closely coupled, to the subject. For example, Alice's non-verbal behaviour is strongly linked to Bob's non-verbal behaviour, both temporally and in form and meaning. Given that computer interfaces, such as ECAs, are treated as social actors (like other humans) this suggests that non-verbal behaviour between a human and an ECA should also maintain this linkage, which will eventually be highly complex and important to the interaction. This leads to the hypothesis that in an interaction between a human and an ECA, close-coupled non-verbal behaviour is important for persuasion.

Given that the behaviour of present-day ECAs is limited compared to real humans, especially with respect to non-verbal behaviour, the notion of a synthetic ECA is introduced and used to empirically evaluate the persuasive potential of an ECA within a specific context, and further used to evaluate whether the close-coupling of non-verbal behaviour between an ECA and human would be important. These evaluations are important for motivating the development of non-verbal behaviour within ECAs.

The next section discusses social influence in ECAs along with direct measures of behaviour change; what synthetic ECAs are and how they are implemented; followed by full details of the new empirical studies in this research, demonstrating the validity of using synthetic ECAs, and evaluating the persuasiveness of a synthetic ECA in the specific context of the ECA presenting information about a charity through a web-chat-

style interface, using a direct measure of behaviour changes. Finally, the results of these studies are discussed, along with their limitations and meanings for the future development of ECAs.

4.2 Social influence in ECAs

Published previous work on persuasion and social influence (Bailenson & Yee, 2005; Baylor, 2006; Blascovich, 2002) primarily uses metrics based on self-reports of attitudes and belief. Only a limited number of empirical studies have measured behaviour change directly. Bickmore et al (2005) used a 'relational agent' – "*computational artefacts that build and maintain long-term social-emotional relationships with users*" (Bickmore, 2003) – in the role of an exercise advisor to encourage older adults to meet the minimum level of physical activity currently recommended, and used a combination of questionnaires and direct behavioural empirical measures. These behavioural measures took the form of number of steps walked as recorded by a pedometer. Results demonstrated that relational agents increased the amount of physical activity (i.e. number of steps) five times faster than the control group over the duration of the study (Bickmore et al., 2005) and that difference was highly significant. The control group in this study used non-interactive paper-based materials which undermines the inference that it is the ECA alone that explains the persuasion effect (for example, any interactive system might have a similar effect). Bickmore's initial studies demonstrate that current state-of-the-art ECAs have persuasive potential and whilst it is fair to assume that present state-of-the-art ECAs are unlikely to be as persuasive as real people (due to their limited cognitive and communicative capacity) studies comparing synthetic ECAs with other forms of interactive media, and with real people, would provide both stronger evidence of their utility and a justification for further technical development.

ECAs provide a relatively new and unexplored medium for interacting with computer and information systems. Modern computer hardware and software make it possible to build ECAs with high visual and auditory acuity, which are highly customizable, but the

largest limitation of present ECAs is not their appearance, but their behaviour. That said, even with their limited behavioural acuity, present-day ECAs can be demonstrated to have significant social influence (Bailenson & Yee, 2005; Fogg, 1998; Reeves & Nass, 1996). As stated by Zanbaka, Goolkasian, and Hodges (2006), "*in order to successfully exploit virtual humans for these ... applications ... researchers must first determine if there exists a measurable similarity between a person's response to a virtual character and that person's response to a real person.*" Like Zanbaka (2006), this study investigated persuasion as an aspect of social influence, with the aim of measuring that similarity in order to demonstrate the utility of ECAs. Previous studies have focused on indirect measures of persuasion – the effect of attitudes and beliefs – mainly through the use of content-related agreement questionnaires, rather than by measuring the effect of persuasive intervention against a real behaviour. For example, in measuring whether social perception of human speech and computerized text-to-speech was affected by gender of voice and listener, a study by Mullennix, Stern, Wilson, and Dyson (2003) assessed listeners on attitude change and on their perception of various voice qualities, while Stern, Mullennix, Dyson, and Wilson (1999) measured perceived favour towards a variety of different voices. It is important to measure beliefs and attitudes, but the present study maintains that it is more important to measure the actual desired outcome – the desired change in behaviour. Thus, the experiments were designed to measure behaviour change directly. For the purpose of the experiments in this thesis 'persuasion' means the change of interactant B's behaviour caused by interactant A.

4.3 Synthetic ECAs

A synthetic ECA appears to the viewer be a real (computer generated) ECA, but is in fact the movement and sound of a real human transformed in real-time giving the appearance of an ECA – the behaviour is that of a real human, thus resolving the behavioural limitations of present-day ECAs. With this human-level behaviour (both verbally and non-verbally) of synthetic ECAs, they can be used to evaluate the persuasive potential of ECAs in a specific context.

4.4 Implementation of a synthetic ECA

Synthetic ECAs use a real human (termed a *wizard* - likened to the wizard in the Wizard of Oz who controlled a disembodied and imposing head from behind the scenes (Fleming et al., 1939)) – see Figure 4-4 for a resulting image - for the behavioural functionality and can be implemented either 1) by driving a real ECA directly (from motion capture and speech recognition of the Wizard), or 2) by transforming video and audio of the Wizard in real-time. Whilst requiring expensive real-time motion capture equipment the first approach is straightforward, but it does have drawbacks because of difficulties in obtaining both facial motion capture and upper body motion capture concurrently. Furthermore, this approach introduces subjective beliefs about which aspects of movement and which body/face elements are important in human interaction (due to the motion capture and character animation limitations and capabilities). The second approach, transforming video and audio of a Wizard to give the appearance (aurally and visually) of an ECA, avoids these issues of introducing subjective beliefs and utilizes only commodity hardware. This latter point is important as it allows synthetic ECAs to be used in other studies evaluating potential affect of ECAs amongst more groups of researchers – the expense and complexity of using motion capture equipment places that approach beyond reach of most research groups.

Both these approaches suffer from risks that due to behavioural and/or visual acuity people could come to believe that the character must (in their opinion) be driven or controlled by a real human – as a computer system could not (in their opinion) provide that high level of behaviour. Thus, it was important to verify that people believed that a synthetic ECA was a real ECA. Therefore, as a precursor to studying persuasion a study was run to validate the synthetic ECA approach. This study discussed in section 4.5.3 concluded that subjects did believe the synthetic ECA to be a real ECA.

Present day ECAs all appear visually to be computer generated. That is, they do not approach photo-realism. Photo-realistic ECAs would have a natural advantage in terms of persuasive potential over present day ECAs as people could be led to believe they

were real people, but with otherwise present-day technology these ECAs would be let down by their behaviour. A photo-realistic *synthetic* ECA (i.e. an ECA driven by a real person and appearing photo-realistic) would be indistinguishable from a real human (in appearance and behaviour) and would therefore not be of use to investigate the potential effects of ECAs. Given the aim to investigate the persuasive potential of ECAs, the synthetic ECA agent needs to appear as an ECA to support the belief that it is an ECA. Conversely, it would make no sense for the synthetic ECA to be photo-realistic as then combined with the realistic behaviour from the Wizard it would be the same as a real human and only the effect of beliefs about a character that was otherwise identical would be tested.

4.4.1. Cartoonising video

Creating the appearance of an ECA from video of the Wizard is achieved by 'cartoonising' the video in real-time. Previous work by Fischer and Bartz (2005)) cartoonised video streams for augmented reality purposes (see Figure 4-1), to prevent users from being able to distinguish between real and computer generated artefacts. High fidelity was important in this work and significant attention was given to running the cartoonising process on GPU (Graphical Processing Unit) hardware. For this present study, however, high fidelity was not so important thus the challenges of targeting GPU hardware were avoided.

Figure 4-1 Cartoonising for augmented reality (Fischer, J. & Bartz, 2005)

Similarly, Wang et al (2004) report on cartoonising video as an offline process with a focus on spatio-temporal coherence and a variety of video styles, exemplified in Figure 4-2 (Wang et al., 2004). Again, these factors are less important for this present work while real-time performance was a definite requirement.

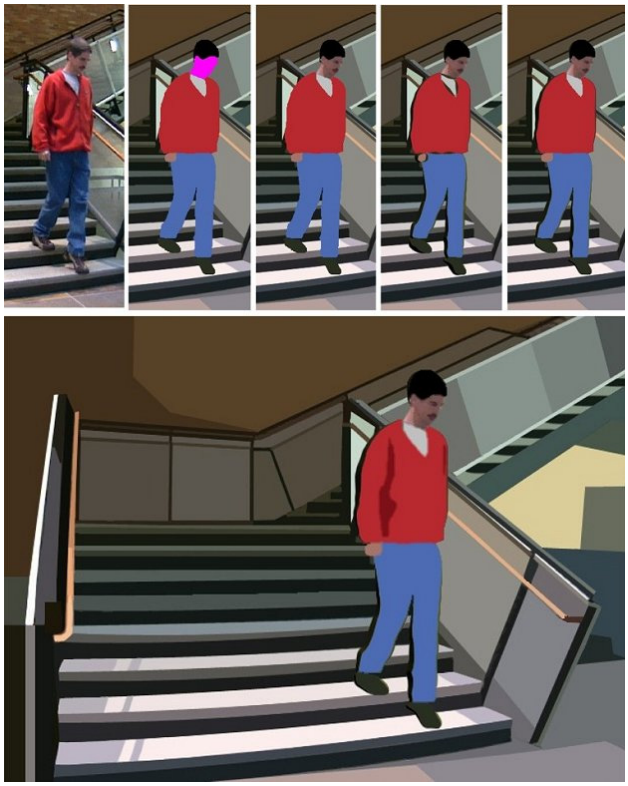


Figure 4-2 VideoTooning (Wang et al., 2004)

Cartoonising filters are available in many photo-editing packages, but they do not run in real time. The open source photo editor 'the GIMP' (Kimball & Mattis, 2006) has cartoonising code available so the video transformation algorithm used in this study was based on their code, with modifications to run in real-time. The transformation algorithm was implemented using EyesWeb (Camurri et al., 2004) – a rapid application development environment built on top of the computer vision library OpenCV (Intel Corporation, 2005), as shown in Figure 4-3. This figure shows the flows of video frames from the webcam (left) to the screen (right), along with the transformations that are

applied to those frames. Initially, the frames are converted to RGB format, and then passed to separate sections (top path and bottom path). The top path performs some noise filtering of the frames, then reduces the colour depth (right bit shift followed by left bit shift). The bottom path converts to greyscale, and then performs edge detection using filters including Gaussian smooth and a hi-pass frequency. The edge detection frames are then alpha blended with the colour depth reduced frames with the edge-detection on top. These frames are then mirrored and delayed so they appear on the screen correctly.

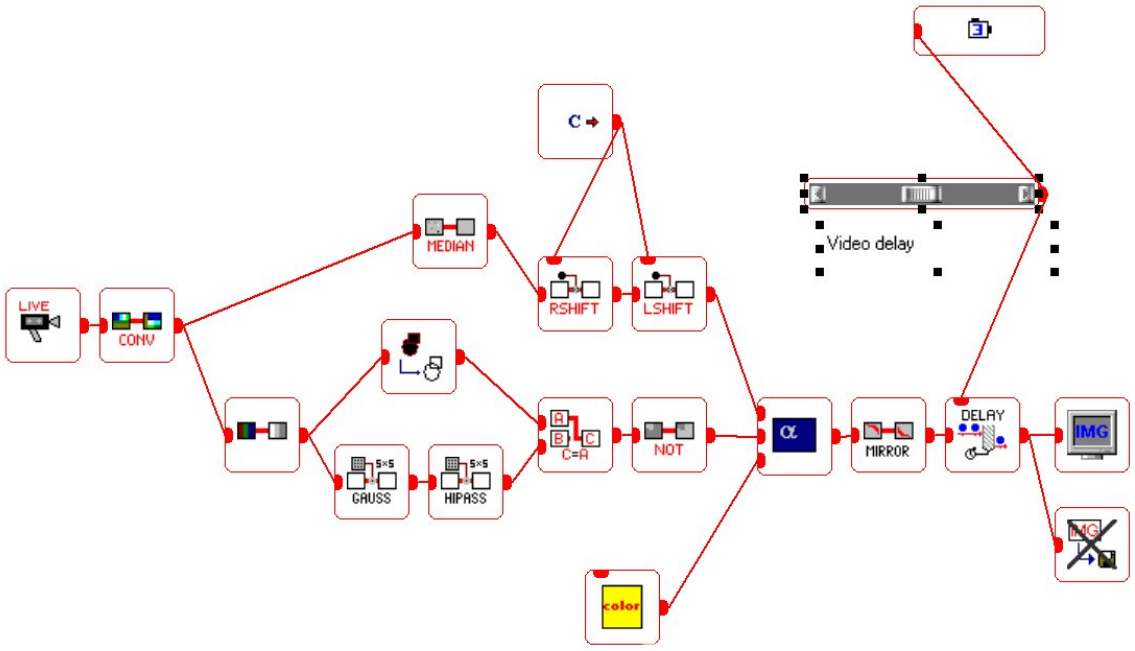


Figure 4-3 Custom cartoonising filter in EyesWeb (Camurri et al., 2004)

Video transformation runs at 25 frames per second (the frame rate of the webcam) at a resolution of 320 x 240 pixels, but is still limited by the level of computing resource available, and increased visual acuity, frame rate, or resolution would require additional computing resources. Visual acuity is below that of the previous work on cartoonising video, but is sufficient for the purpose and has the advantages of working in real-time and not requiring specialist hardware or specialist software development. Figure 4-4 below shows an example frame of how a Wizard appears as a synthetic ECA, which was

shown (during development) to a variety of non-computing science undergraduate students who believed it was computer character, not a real human.

Figure 4-4 Synthetic ECA

Since the verification and synthetic ECA persuasion studies were performed in 2007, cartoonising algorithms have become more commonplace. Specifically, the open source media player VLC (Cellerier, 2005) now implements real-time cartoonisation (settings→preferences→video→filters→distort video filter).This approach would be simpler and more generic to use in further studies.

The audio of the Wizard was transformed using commercial voice transformation software MorphVox (Screaming Bee, 2006), as shown in Figure 4-5. The audio and

video were synchronized to play together, by delaying the audio to match the longer delay of the video introduced by the processing thereof.

Figure 4-5 MorphVox Voice Changing Software (Screaming Bee, 2006)

4.5 Verification of validity of synthetic ECA

As it has been widely observed that users have a tendency to treat computer interfaces (such as ECAs) as social actors (Reeves & Nass, 1996) it has been hypothesised that users should respond to a fully functional ECA in similar manner as they would to a human conversant. However, for studies of the persuasive potential of ECAs, it is

important to establish that people, in fact, treat a synthetic ECA as a computational artefact and not as they would a human (or we would simply be studying the persuasive potential of people themselves). A verification study was designed, aimed to establish that: 1) Subjects believe the synthetic ECA to be a real ECA 2) Subjects behave differently towards synthetic ECAs than towards real people. The first aim is important, so that it can be argued that any effects of a synthetic ECA can also be applied to the real ECA, while the second aim is important because if people treated a synthetic ECA the same as a real person one wouldn't expect a difference in behaviour between the two.

Subjects interacted with a real human in a video conference. The real human (the Wizard) appeared to subjects either as a human (condition H) or as a synthetic ECA (condition E). The Wizard asked questions and responded to the user according to a simple script. The Wizard participating in the video-conference was unaware of whether they were either directly projected (human condition **H**) or appearing as a transformed image (ECA condition **E**). Aim 1 was established through the use of a post-interaction questionnaire, while aim 2 was established using two approaches. The first approach used the amount a subject would, with the character, break the 'social contracts' that are a natural component of dialogue with another person. In other words if the subject considered their conversational partner (the synthetic ECA) an intentional agent, they would be less likely to break the 'social contracts'. The second approach used eye gaze as a measure of social engagement. As discussed in Chapter 2, gaze and eye behaviours are important features of human-human interactions, especially in conversations – serving a variety of purposes beyond simply gathering information (as noted before: regulating the flow of communication; monitoring feedback; reflecting cognitive activity; expressing emotions; communicating the nature of an interpersonal relationship) – and have a complex set of social norms and social contracts.

Eye tracking technology enables continuous high-precision tracking of where people are looking, while being minimally invasive and totally objective. Eye behaviour when interacting with non-social entities is significantly different from that while interacting with social entities. These differences can be used to measure the extent to which people

consider ECAs as social entities, and furthermore that same eye behaviour can be used as a continuous, real-time, on-line metric for evaluating social behaviour in ECAs, though this latter point is not the focus of this work. Suffice to comment that this means that methodologically sound empirical evaluations of ECAs could be performed using eye tracking and as interactions could also take place with real people, eye-gaze behaviour while interacting with ECAs can be directly compared with the target ideal of real human interactions.

The verification study uses qualities of the subject's gaze behaviour as a measure of the maintenance of the social contract, with an expectation of difference between the two conditions of interacting with a synthetic ECA and a real human to demonstrate the difference in attribution of intentionality towards the synthetic ECA.

To force subjects to break their 'social contracts' they were requested to attempt a visual counting task at the same time as interacting over a video conference. This visual counting task required them to break their social contract and the characteristics of these breaches were measured using eye tracking technology – specifically the Tobii x50, illustrated in Figure 4-6 (Tobii Technology AB, 2006a). The eye tracker measured where on the screen a subject was looking, so it could then be determined when the subject was looking at the character or the distraction task or elsewhere.

Figure 4-6 Tobii x50 Eye Tracker (Tobii Technology AB, 2006a)

The visual counting task involved an image on the same display as the video conference, but at a different location. This image has a number of items to count and a set of numbers to click to indicate how many items were present. These images and their order are given in Appendix A2, and an example of an image with the counting question is shown in Figure 4-7.

Figure 4-7 Example page of distraction task

The transformation of the audio and video in the *synthetic ECA* condition introduced a small delay and for consistency this delay was also introduced into the human condition.

For consistency, the interaction with the Wizard was highly scripted. The Wizard asked open-ended questions that were independent of previous context, which allowed the majority of the talking to be done by the subject. Questions required detailed answers

about specific aspects of the 'subjects' life, for example, "*I'd love to know about your house. Could you describe it for me? How many rooms there are? Who do you live with? Where is your house?"* See Appendix A1 for the full script. The attentional and cognitive resources required by the counting task conflict with the demands of maintaining the conversation, so in addition to differences in gaze behaviour between the two conditions a reduction in performance on the counting task in the human condition was also predicted – arising from the higher sense of obligation to maintain the social contract.

4.5.1. Data collection and measures

The conversation, taken from the subjects' viewpoint (both of the character and the counting task), was recorded using screen capture. The spatial and temporal properties of each subjects' gaze were recorded using the eye tracker at sample rate of 50Hz. Video of the screen was captured at only five frames per second due to the technical limitations imposed by the computational load of the image processing for the Wizard, though this was sufficient for analysis. The performance on the counting task was measured from the screen capture of the session, including whether the subject counted correctly, and the time taken to complete each counting task.

Task performance was measured in terms of the accuracy, time taken counting, and the total number of images counted during the conversation. Additionally, subjects completed a post-experimental questionnaire on their opinions on the character and the interaction. See Appendix A3 for the questionnaire

4.5.2. Subjects

The study involved 19 subjects, mostly undergraduate and postgraduate students at Newcastle University. Nine subjects were randomly allocated to the human condition (H) and ten to the synthetic ECA condition (E). They were neither age nor gender matched. The human condition had 9 subjects, 4 male, 5 female. The synthetic ECA

condition had 10 subjects, 3 male, 7 female. Subjects' ages ranged from 18 to 36 with a mean of 22.

4.5.3. Results

The eye tracking data was analysed into fixations, yielding a location on the screen, and each fixation was automatically tagged as being either on the character, on the counting task (separate window on the display), or elsewhere on the display. A summary of the eye tracking data for one subject illustrates the character of the data collected. Figure 4-8 shows the complete set of fixations (number), with larger dots representing longer fixations. Figure 4-9 shows this same data summarised, highlighting where the majority of fixations occurred.

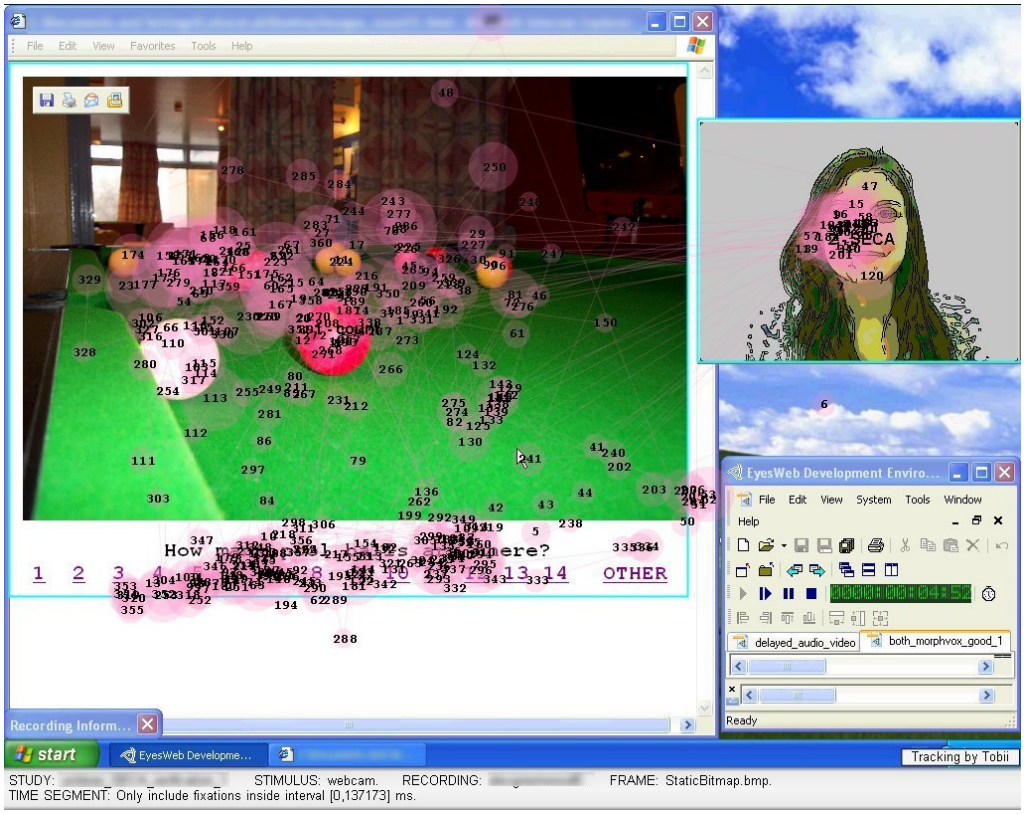


Figure 4-8 Eye fixation points

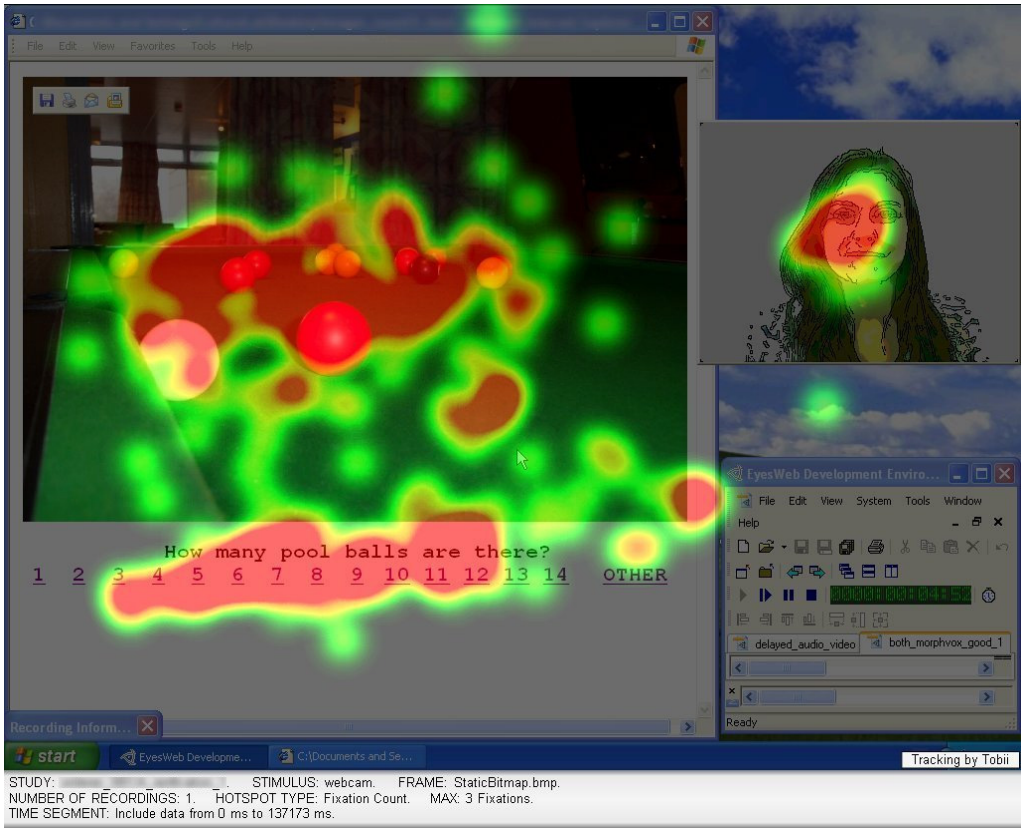


Figure 4-9 Eye fixation point summary

As is typical in visual search tasks, fixations occurred widely over the stimulus, with a higher concentration on the numbers. Fixations also focused specifically on the face of the character and slightly more towards the character's right eye, which corresponds with observations that people tend to look at each other's right eye (MacDorman et al., 2005). The reasons why people look at each other's right eye more than the left are, at this point, unclear.

Most of the questionnaire questions concerning subject perceptions of the interactions showed no significant differences between the two conditions (human **H** and synthetic ECA **E**). It was no surprise to find that subjects differed significantly between the two conditions in their rating of how human the character was. Subjects were convinced **E** was not human despite its actually being a transformed image of a real person. Ratings on a Likert scale from -2 (strongly disagree) to +2 (strongly agree) for the proposition

"*The character's body was human*" were 1.7 for **H** condition compared to 0.2 for **E** (t=2.83, p=0.012). Subjects were generally not convinced that **E**'s speech was human: – for the proposition "*The character's speech was human*", the **H** condition gave an agreement of 1.6, while the **E** condition gave an agreement of 0.8 (t=1.97, p=0.065). Subjects were overall very convinced that the character **H** was a human and that the character **E** was not: – for the proposition "*The character was a human*" agreement ratings were 1.22 for **H** and 0.1 for **E** (t=2.50, p=0.023).

When asking many questions there is a danger of finding a 'significant difference by chance' with multiple paired comparisons and it could be suggested that a Bonferroni correction should be used to lower the significance level for such questions. This is because as the number of statements increases, the chance that the existing data set shows significance just by chance for one of the statements also increases. This is definitely true, but the Bonferroni correction tends to massively over-estimate this chance. For this study with a total set of 39 independent tests (on 2 conditions) the Bonferroni correction would reduce the significance level from 0.05 to 0.05 divided by 39, or 0.00128. The Bonferroni adjustment is used to minimise Type I errors, but does so by increasing the probability of accepting the null hypothesis when the alternative is true – a Type II error (Morgan, 2007). For this, and further studies, a Bonferroni correction will generally not be applied.

The common sense argument against using Bonferroni corrections says *"Bonferroni adjustments imply that a given comparison will be interpreted differently according to how many other tests were performed"* (Perneger, 1998). In other words, if another 50 statements had been given for agreement/disagreement, it would be even less likely that any of the first 50 statements would be correlated with the condition. Bonferroni corrections were developed for statistical tests aiding decision-making, not for assessing evidence in data. Generalised alternatives to Bonferroni corrections have not at this point been established, but it has been suggested that Bayesian methods (which can incorporate a priori beliefs) would be more appropriate – *"The integration of prior beliefs with evidence is best achieved by Bayesian methods, not by Bonferroni*

adjustments. In summary, Bonferroni adjustments have, at best, limited applications in biomedical research, and should not be used when assessing evidence about specific hypotheses" (Perneger, 1998). Further discussion of the appropriateness of Bonferroni corrections is beyond the scope of this thesis, though further readings on multiple testing can be found in (Bauer, 1991).

There was no significant difference between the two conditions for accuracy, time taken for the counting task, or in the total number of images counted. However, total conversation length was significantly different between the conditions. Average conversation length was longer with a synthetic ECA: 163 seconds compared to 141 seconds with a human (t=2.14, p=0.047). A variety of reasons to explain this could be theorised, such as subjects felt less social pressure to stop talking when talking with the ECA. The reason is not important for the work within this thesis.

Eye tracking data showed highly significant differences between the two conditions, specifically with respect to the proportion and the mean length of total fixation time on the character. When interacting with **H**, subjects spent on average 20% of their total fixation time on the character, while for interactions with **E** this proportion increased to around 45% of the time (t=-2.46, p=0.025). This inevitably left less time under the human condition for attending to the counting task, though the difference does not attain statistical significance: – 52% c/f 75%; t=1.91, p=0.074, and did not affect accuracy or speed. The most significant metric of the gaze behaviour was the mean length of each fixation – when subjects were looking at **H** they spent on average about 625 ms on each fixation, whereas when looking at **E** it was approximately half that at 346 ms (t=2.69, p=0.015). There was also a trend for the mean number of fixations on **H** character to be higher than on **E** (70 c/f 42), although this difference did not reach statistical significance (t=-1.85, p=0.082). These results are summarised in Table 4-1 (significance value in bold are those below 5% chance).

	H	E	
Metric	Mean	Mean	Sig.
Questionnaire statement agreement with "The character was a human"	1.22	-0.1	**0.023**
Questionnaire statement agreement with "The character was computer generated"	-1.11	0.20	**0.024**
Number of fixations on counting task	216.80	209.78	0.911
Number of fixations on synthetic ECA	42.20	70.33	0.082
% fixation time on counting task	0.7445	0.5174	0.074
% fixation time on synthetic ECA	0.1979	0.4503	**0.025**
Mean fixation time on counting task	289.64s	242.89s	0.086
Mean fixation time on synthetic ECA	624.79s	346.62	**0.015**
Number of images counted	11.40	10.00	0.475
Image counted per minute	4.42	4.28	0.868
Conversation length	163.30	140.67	0.047

Table 4-1 Summary of synthetic ECA verification study metrics

4.5.4. Discussion and conclusions

As hypothesised the results of the verification study show that subjects did not believe the synthetic ECA was human, and although task performance was not different between the two conditions, gaze behaviour in the two conditions showed a marked difference. In

considering gaze behaviour as one aspect of social contract maintenance (in conversation) a significant difference was found both in the average length of each fixation (shorter fixation on the synthetic ECA), and in the time spent looking at the character (less time on the human). This suggests that significantly different social protocols are in operation under the two conditions. Both the questionnaires and the gaze behaviour indicate that subjects were unaware that the synthetic ECA was in actuality a real human and subjects appeared to interact with the two in a distinctly different manner. This suggests that the concept of synthetic ECAs is valid and appropriate for exploring various potential qualities and the evaluations thereof of ECAs, specifically the persuasive potential of ECAs.

5. Persuasive effect of synthetic ECAs

The concept of synthetic ECAs introduced in Chapter 4 and the verification of synthetic ECAs enables using synthetic ECAs to evaluate the persuasive potential of ECAs. This chapter establishes this persuasive potential empirically within a constrained charitable giving scenario. The importance of using a direct measure of behaviour change is discussed, and full details of the experimental design are given. The experimental conditions used and the aspects of non-verbal behaviour they were intended to elucidate are discussed. The specific procedure for each subject and the stages they complete are given, along with the measures taken during the study. Finally, the results of the study are presented, showing that the most persuasive synthetic ECAs are those which have visual information on their interactants; the meanings of this result and consequences for the future development of ECAs are then discussed, along with the limitations of the study.

The persuasive effect a synthetic ECA has on people can be used as an estimate of the persuasive potential of an ECA. As discussed previously, most evaluations of ECAs, whether for evaluating persuasiveness or other social effects, have been based on questionnaires or structured interviews (Bailenson & Yee, 2005; Keeling et al., 2004) – measuring persuasion indirectly. There appear to be no studies that have evaluated the persuasive effect of ECAs using a direct measure of persuasion – as defined as a difference in behaviour over a set of conditions, or any other definition.

It was believed that it is important that being able to see one's interactant during an interaction is important in order to modulate one's non-verbal (and possibly verbal) behaviour in accordance with it. This leads to the assumption that being able to see the interactant and therefore their non-verbal behaviour enables modulation of the ECA's behaviour and thus increase persuasiveness (or alternatively, that not modulating behaviour in response to an interactant's non-verbal behaviour decreases persuasiveness).

5.1 Direct measure of behaviour change

The novel approach of measuring behaviour change directly involves giving each subject the opportunity to donate money from their £10 payment for participating in the experiment to charity after an interaction with the synthetic ECA. When compared with the control condition of interacting with a real human in a web-chat format (human-level persuasiveness) the amount donated is a direct measure of the persuasive effect of the synthetic ECA. It would have been ineffective to measure behaviour change for *each* subject, as asking them beforehand to donate or asking how much they *would* donate would influence the later donation, but it is possible to measure behaviour change over a group of subjects – the prediction was that under different conditions the subjects would on average donate differing amounts.

5.2 Experimental designs

Subjects interacted with a character under four conditions:

A The character consisted of transformed video and audio of the Wizard, and the Wizard had video and audio feedback on the subject (synthetic ECA *with* video condition).

B The character did not appear, though real audio of the Wizard was presented, and the Wizard had only audio feedback on the subject (audio only condition).

C The character was real video and audio of the Wizard, and the wizard had video and audio feedback on the subject (human condition).

D The character was transformed video and audio of the Wizard, and the Wizard had only audio feedback on the subject (synthetic ECA *without* video condition).

The human condition C is the control condition – how persuasive a 'real' human can be. Conditions A and D reflect the persuasive potential of ECAs (utilizing a synthetic ECA) – the difference measuring the importance of visual feedback for persuasiveness. Condition B was included for completeness. It was assumed that each group would on average be the same pre-experiment as subjects were put into each group at random.

For consistency, audio and video were delayed across all conditions due to the delay introduced by video transformation. The human Wizard was unaware during each interaction (and in fact for the duration of the whole study) that they were sometimes appearing to subjects as an ECA. They were under the impression that they were only engaged in a video conference.

5.3 Subjects

Subjects were recruited from the local area through a readily available subject pool. There were 76 subjects – 44 female and 32 male, with 21 subjects exposed to the ECA with video condition A, 18 subjects exposed to the audio only condition B, 19 to the human condition C and, 18 to the ECA without video condition D. Subjects were randomly assigned to one of the four conditions, the variation due to some subjects not turning up to experiments. Subjects interacted with the character (whether human or synthetic ECA or audio only) though a webcam and computer screen, with audio provided with headphones with an in-built microphone. They were able to see the head and shoulders of the character. Under all conditions and subjects the Wizard was a female. Due to logistical reasons the study was performed in two sections, with a different female Wizard for each section. Other than using a different Wizard the only difference between the two sections was a change of room. This change was checked in the statistics of each group for differences, and differences were not found and

additionally the significant results were found within both sets of subjects from each room independently.

5.4 Wizard behaviour

In all conditions the Wizard was presenting information to the subjects about a specific charity, and giving the subjects the opportunity (anonymously from the Wizard's perspective) to donate to the charity – the Wizard was not actively seeking to persuade the subjects, merely to present information about the charity.

5.5 Procedure

Each experiment consisted of a series of steps for each subject. Each step gave instructions to and for the next step, and additionally the experimenter gave subjects the full set of instructions on all steps at the start. For the duration of the experiment, subjects were self-guided.

The first step was a Myers-Briggs (Quenk, 2000) personality type test that took to majority of the time. This was to distract subjects from being focused on the interaction with the character as the main important section of the study. The personality type data was not used.

The second step was the interaction with the character, under one of the four conditions. When the subjects started they found themselves able to see (except under the audio only condition) and hear the character and were instructed to say 'Hello' to start the interaction (see Appendix B1). The Wizard then asked some general questions about the subject (such as their name) then went on to present information about the charity, allowing questions and interacting non-verbally with the subject. Finally, the Wizard explained that after the interaction the subject could donate some of their £10 payment for the study to the charity if they chose to. The subject then terminated the interaction

and if they felt so disposed donated some of their payment to the charity. The general introductory script and ending script are shown in Appendix B2, while the information used for the informative section is given in Appendix B3.

The final step of the study was a follow-up questionnaire (paper-based) consisting of the same set of statements as the verification study above used to open the interaction, again using a 5-point Likert scale (Likert, 1932) ranging from -2 (strongly disagree) to +2 (strongly agree), with an opportunity for open-ended comments at the end, as shown in Appendix B4.

5.6 Measures

The main measure was the amount each subject donated to charity. Subjects could donate any amount from zero to a maximum of £10 (the amount they were paid) in 50p increments. The hypothesis was that subjects would donate most under the human condition, with reduced amounts under the other three conditions, and also that the synthetic ECA *with* video feedback condition would have more donated than the synthetic ECA *without* video feedback – reflecting the postulated importance of visual feedback in persuasion. The follow-up questionnaire was included for completeness and verification of the study, but like the Myers-Briggs questionnaire did not include measures that were directly relevant to the actual study – merely concerning the nature of the interaction and the subjects' beliefs as to the computer-generated or human nature of the character they interacted with.

5.7 Results

The average donation for all subjects was £2.60, with a standard deviation of £3.17. The minimum donation was zero, while the maximum was the maximum possible of £10. The standard deviation of the amount donated was large across all conditions.

The mean amount donated and standard deviation for each condition is shown in Table 5-1 below, and illustrated in Figure 5-1 and Figure 5-2, where it is clear that subjects were less persuaded to donated money to charity under condition D.

Condition		Mean	N	Std.Dev.
A	Wizard appears as ECA and CAN see subject	£3.50	21	£3.17
B	Wizard is not shown and CANNOT see subject	£2.94	18	£3.67
C	Wizard appears as HUMAN and CAN see subject	£2.47	19	£3.58
D	Wizard appears as ECA and CANNOT see subject	£1.36	18	£1.62
	Total	£2.61	76	£3.17

Table 5-1 Amount donated to charity versus condition

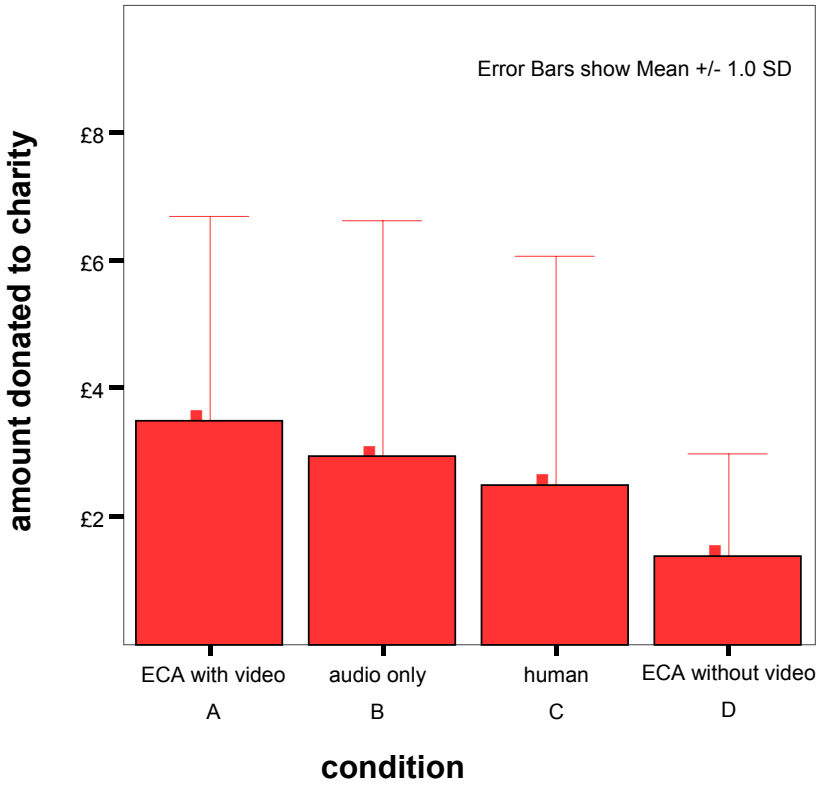


Figure 5-1 Amount donated to charity versus condition

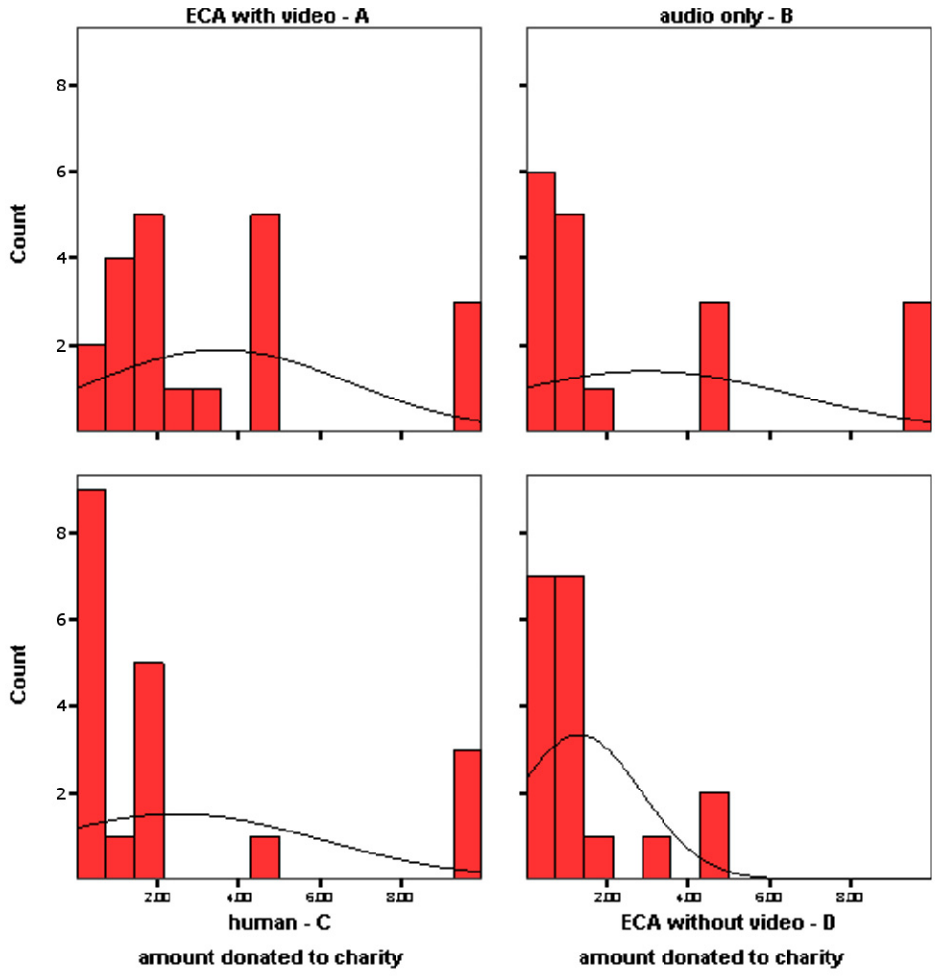


Figure 5-2 Amount donated versus condition (histograms)

Statistical analysis showed non-normal distribution of donations making both ANOVA and t-Tests invalid. Instead a non-parametric Kruskal-Wallis test for correlations was performed, for which the test statistic (Chi-squared) was 7.754, equating to a probability of 0.051. The 0.05 boundary is an arbitrarily chosen number, and in this case it was chosen to proceed with further comparisons of the means even though the probability was (very slightly) above 0.05.

A Wilcoxon test was run to compare non-normally distributed means, which found the probability for the difference between synthetic ECA *with* and *without* video feedback to

be p=0.003. There is a danger of finding a 'significant difference by chance' with multiple paired comparisons, so a Bonferroni correction was applied giving the significance criteria as 0.05/6 = 0.0083.

In summary it is concluded that there is a reasonably robust significant difference between the synthetic ECA with and without video feedback groups, and therefore that ECAs with visual feedback on the interactant have a greater persuasive potential than ECAs without visual feedback.

No significant difference was found between any other pairs of conditions, so the experimental results cannot support the hypothesis that a synthetic ECA is less (or more) persuasive than a real human, and the large variances preclude concluding that they are equally persuasive.

5.8 Discussion and conclusions

The experimental studies show that when interacting with what seems to be an ECA (even though it is really a real person with video and audio transformed so as to appear as an ECA), real people are more persuaded (using an absolute measure) when the ECA can see them than when it cannot – validating the assumption of being able to see the interactant being important to maximise persuasiveness. The conclusion being that being able to 'see' the user is important for ECAs to be effective.

This study was focused on exploring how *synthetic* ECAS can be used in measuring performance of ECAs, specifically on measuring persuasive effects. Using a direct measure of persuasion, it wasn't possible to draw any hard conclusions about how persuasive *synthetic* ECAs are as compared to humans, but the study did find that visual feedback was important in the persuasive effect of *synthetic* ECAs. As subjects were not aware that the synthetic ECA was not a real ECA, it can be concluded that visual feedback will also be important in the persuasive effect of *real* ECAs. This is an

important conclusion for demonstrating the utility of future work involving using visual feedback to inform the behaviour of ECAs.

The development of *synthetic* ECAs enables experiments evaluating ECAs with human-level behaviour before those high behavioural quality agents have been developed. These experiments may be useful not only in informing the development of future ECAs, but also for approaching some of the ethical, personal and societal issues.

Finally, the introduction of a methodology using a direct measure of persuasion may encourage future work also using direct measures. It does not necessarily follow that from changes in attitudes and beliefs, actual behaviour is affected, and as it is this final effect on behaviour that is important in many arenas where ECAs may be used, direct measures of persuasion are important.

5.9 Limitations of this work

The results of this study are limited to the interactions within a relatively simple environment (a webcam interface) and may not generalize to more realistic or complex environments. The study also does not address agents that may attempt to be more pro-actively persuasive – using more persuasive language, non-verbal communication, and other persuasion methods. Furthermore, different people and different types of personality have differing amounts of persuadability. This fact could have been used to restrict the subject set to people who would likely be more persuadable and therefore increase the general donation level and presumably increase the differences between conditions, but the overall target of ECAs are to interact with all people, and in this case to have a persuasive impact on all people, so it was felt that focussing on groups of people who were more persuadability would detract from the meaning towards this target.

The visual acuity of the character could be increased, especially the resolution of the displayed image, though the resolution was the same across conditions and is normal for

video conferencing. Additionally, the complete set of possible conditions was not performed.

The quantization of monies given to subjects and the exact denominations may have had an effect on the amounts donated, and the large variances involved with the amounts donated require studies with larger numbers of subjects for more conclusive results.

6. Behaviour-based architecture(s)

ECAs need to both use non-verbal behaviour and react to non-verbal behaviour at interactive rates, as shown in the studies presently in Chapter 5. The architectures of present ECAs are not designed with this in mind. The concept of behaviour-based architectures in robotics provides a solution to this architectural problem, and corollaries can be drawn with the historical development of robot control system and the present state of ECAs to suggest that a behaviour-based architecture would be appropriate. This chapter overviews behaviour-based architectures and examines the corollaries mentioned above in more depth, before discussing how behaviour-based architectures may provide ECAs with non-verbal behaviour that responds constantly and quickly to the non-verbal behaviour of an interactant.

The result from Chapter 5 indicated that in order for an ECA to be most persuasive it was important for it to have responses to the non-verbal behaviour of an interactant, and that these responses should happen promptly and constantly

Introducing input into an ECA system about the non-verbal behaviour of an interactant creates a significantly larger amount of more complex information than present ECA systems are designed to work with. A similar problem was found in the development of robot control systems when they moved from simple, controlled simulated worlds to the complex, noisy, uncontrolled real world. In order to resolve this problem the concept of behaviour-based systems was created and eventually progressed into the three-layer or hybrid architecture that is seen in most robot control systems at present. It is postulated that the development of ECA systems can learn from this history and use behaviour-based systems and hybrid architectures to endow ECAs with interactive non-verbal behaviour.

A streaming architecture for building ECAs is introduced as a proposal for an implementation of a hybrid architecture based on the history of robotics. Streaming architectures view the world as a set of data streams, and modules that perform processing on those streams. It is proposed that not only can the simple, low-level non-verbal behaviours be implemented in this manner – connecting inputs to outputs with a

minimum of processing and delay in between – but medium-level sequencing behaviours such as a conversation state, and high-level 'cognitive' functions such as determining what should be said next can also be integrated (though with increased processing and delays).

The results of studies of real people interacting with synthetic ECAs in Chapter 5 indicate that it is important for ECAs to react non-verbally to the non-verbal behaviour of their interactants. In order to test the value of ECAs reacting to non-verbal behaviour, a prototypical ECA must be developed. The functionality of this ECA may be constrained and the context limited, but it must be sufficient to demonstrate an increased value (persuasiveness) over the same ECA without interactive non-verbal behaviour; therefore it is important to be able to evaluate this ECA. Evaluating a real ECA under the same paradigm as the studies in Chapter 5 – discussion of charitable giving – is appropriate and provides a concrete and limited scenario and a solid (and comparable) metric. Although not fully functional, this ECA will have some interactive non-verbal behaviour, while also presenting information about charitable giving.

It has already been discussed that present ECAs mainly employ a deliberative architecture, mostly focused on natural language processing and usually with text as the only input (possibly from speech recognition). They process the text to understand a meaning, search data sources for answers, and then generate grammatical responses. This high-level intelligence involves symbolic processing and search and has a high and variable latency. This chapter overviews and draws corollaries with the development of robot control systems where researchers found that deliberative systems struggled with the volume, complexity and noisy nature of real-world data. This suggests that deliberative AI is also not appropriate for all aspects of an ECA's architecture. In other words, much non-verbal behaviour requires faster and more timely responses than verbal behaviour; the data sources for non-verbal behaviour are much greater in volume and complexity than text input and at this point deliberative processing on those inputs is intractable. In contrast to natural language the models for non-verbal behaviour (from psychology and psycho-linguistics) are highly limited and are not computational models

– they are predominantly descriptive models and do not at present enable the identification of appropriate responses in a given situation.

The use of more reactive or reflexive systems – behaviour-based systems – could provide an ECA with faster, more timely responses within the computing power available. In practice, for ECAs to utilise both verbal and non-verbal behaviour effectively a hybrid system of deliberative and reactive systems would be required, similar to that found in robotic systems – a hybrid of reactive and planning modules. To date this form of technique appears to have had little attention within the ECA community. The strongest example is that of Liu et al. (2003) who used a subsumption architecture to provide an ECA with a real-time motion control system so it could independently navigate a virtual world and quickly make responses to the environment, while also performing task-planning to realize more intelligent behaviours. This work did not focus on the ECA in a conversational scenario, nor employ much in the way of non-verbal behaviour (interactive or otherwise), but did generate realistic real-time motion for the character.

The present-day example of non-player characters in computer games employ significantly different approaches to traditional ECAs as a timely response is more important than in most ECA scenarios. In other words game characters focus on timely responses at the cost of sophisticated behaviour, while traditional ECAs do the opposite. These game characters are required to provide involving game play and generally do not engage in verbal interactions except during cut-scenes or as minor responses to clear events, such as when hit by a bullet etc. They are required to react strongly in real-time – delayed reactions are not acceptable to players. This has meant that the AI approach for games characters is almost entirely opposite to that of traditional ECAs – characters are highly reactive; do little planning; have highly limited sensory input and limited output mechanisms (just some pre-animated action), but have a fast behaviour loop – in the order of hundredths of seconds. Table 6-1 below compares the AI approaches of traditional ECAs with that of game characters.

	Traditional ECAs	***Games characters***
Deliberative vs. reactive	Deliberative systems – sense, process, act	Reactive, very limited planning
Sensory input	Limited sensory input (frequently text only)	Limited sensory input
Forms of processing	Processing: natural language processing and understanding; speech recognition; knowledge reasoning	Processing: If-Then, conditions; Route planning
Interaction style	Highly turn-based and discrete	Not targeted for conversation, close interaction (mainly)
Behaviour loop speed	Slow behaviour loop – SPA loop ≈ seconds	Fast behaviour loop – SPA loop ≈ $seconds^{-2}$
Non-verbal behaviour abilities	More complex, but still limited non-verbal behaviour	Limited non-verbal behaviour

Table 6-1 AI structures: traditional ECAs versus game characters

The history of robotic AI systems shows strong similarities to the present development of AI for ECAs. Early robotics used classical symbolic AI methods – sense, process, act – and were found to be effective in highly simplified (and simulated) environments. In other words, with limited sensory input and limited output options (frequently turn-based and/or discrete) they could generate appropriate responses when time wasn't a strong constraint (slow behaviour loop) (Fikes & Nilsson, 1971). In those early days of robot

control systems it was the view that internal abstract modelling was the important aspect of intelligence. The original proposal for the Dartmouth Summer Research Project on Artificial Intelligence in 1955 – arguably the start of AI as a specific research field – reads that an intelligent machine "*would tend to build up within itself an abstract model of the environment in which it is placed. If it were given a problem it could first explore solutions within the internal abstract model of the environment and then attempt external experiments"* (McCarthy et al., 1955).

Reviewing this proposal four decades later, Arkin writes that "*this approach dominated robotics research for the next thirty years, during which time AI research developed a strong dependence upon the use of representational knowledge and deliberation reasoning methods for robotic planning. Hierarchical organization for planning was also mainstream: A plan is any hierarchical process in the organism that can control the order in which a sequence of operations is performed*" (Arkin, 1998). Arkin argues that "*behavior-based robotics systems reacted against these traditions*", with Brooks taking an opposite approach to developing behaviour-based systems claiming that "*planning is just a way of avoiding figuring out what to do next*" (Brooks, Rodney A., 1987). It is also evident that at this time advances in robot and sensor technology made it feasible for the first time to test these control systems in the real world.

Arkin (1998)also noted that "*the inception and growth of distributed artificial intelligence (DAI) paralleled these developments*" with the Pandemonium system (Selfridge & Neisser, 1960) generating coherent behaviour from a set of competing or cooperating processes (or agents). By 1986 Minsky progressed the idea of multi-agent systems as the basis for all intelligence – from multiple simple agents interacting, more complex intelligence can emerge (Minsky, 1986). This leads to the concept of emergence as a whole – "*the appearance of novel properties in whole systems*" (Moravec, 1989), "*Global functionality emerges from the parallel interaction of local behaviors*" (Steels, 1990). Figure 6-1 (Arkin, 1998, Figure 1.10) below illustrates the changes from purely symbolic deliberative systems like those used presently for ECAs through to purely reflexive system introduced to robotics by Brooks.

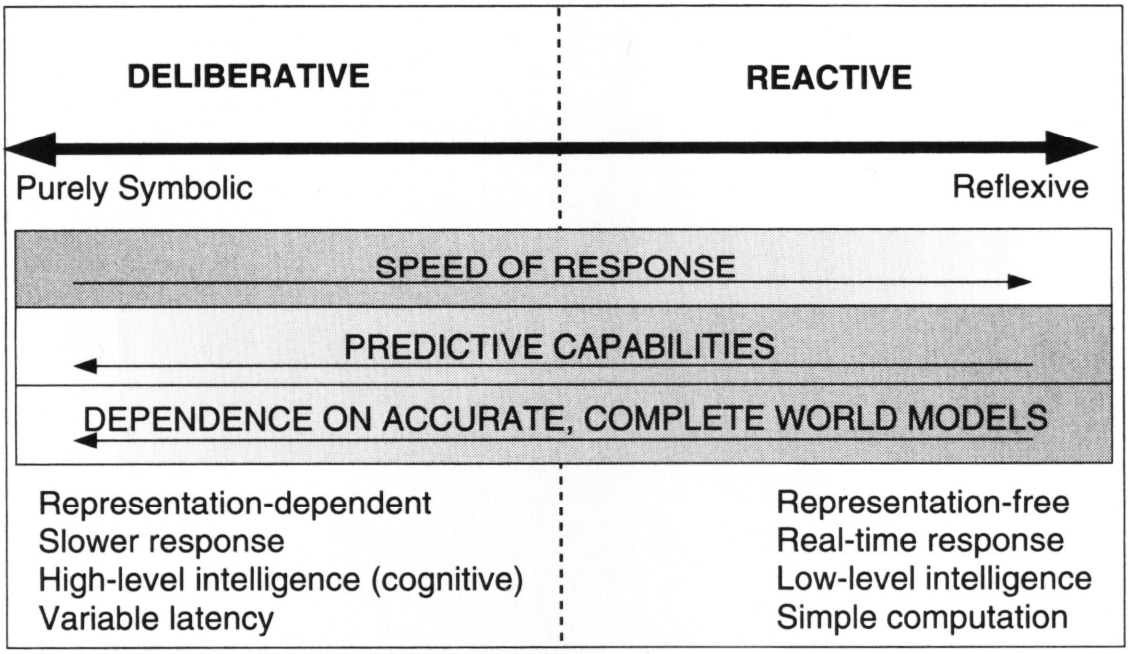


Figure 6-1 Robot Control System Spectrum (Arkin, 1998, Figure 1.10)

Further, Arkin states that "*behavior-based roboticists argue that there is much that can be gained for robotics through the study of neuroscience [study of nervous system's anatomy, physiology, biochemistry, and molecular biology], psychology [study of mind and behavior], and ethology [study of animal behavior in natural conditions]"*, though the goals of robot control systems and those of ECA systems are generally different from such fields – robot and ECA systems do not necessarily require a satisfactory explanation of human level intelligence. That said, an awareness of the major brain subdivisions (Arkin, 1998) clearly suggests that the main attention of the development of ECAs has been on the equivalent of the neocortex.

The terminology of behaviour-based systems causes some confusion as a 'behaviour' in behaviour-based systems means simply 'a reaction to a stimulus', while within common usage and the area of non-verbal behaviour 'behaviour' has a more complex meaning.

While discussing behaviour-based systems, it should be noted that within this thesis the simple 'reaction to a stimulus' definition is what is meant unless otherwise stated.

Behaviour-based systems have behaviours as the basic building blocks for actions, usually "*a simple sensorimotor pair, with the sensory activity providing the necessary information to satisfy the applicability of a particular low-level motor reflex response*" (Arkin, 1998), and avoid abstract representational knowledge in favour of simple reaction to events in the world as soon as they occur – "*Constructing abstract world models is a time-consuming and error-prone process and thus reduces the potential correctness of a robot's action is all but the most predictable worlds*" (Arkin, 1998).

This approach to designing a control system results in a naturally modular system, where new behaviours (in the behaviour-based systems meaning) can simply be added to extend or increase competency. This point is important for the use of behaviour-based systems for ECAs, as it enables building ECAs that have very limited behaviours and competencies and adding to them incrementally over time, without having to (re)design an entire new system.

Behaviour-based systems focus on the challenges of determining what the basic behavioural building blocks are; how those behaviours are implemented or grounded in sensors and actuators; and how the behaviours can be coordinated effectively. Behaviour-based systems are most frequently compared by the way they approach coordination – through arbitration (choose one), subsumption (choose highest priority), action selection, or other forms of competition or cooperation. Generally, as Maes states, "*coordination functions are in effect behaviors that modulate the action of other behaviors*" (Maes & Brooks, 1990), and as such could be changed, or replaced, and/or further modulated by other behaviours (modules). Brooks first used behaviour-based systems to control robots moving around in real rooms with real obstacles and found it allowed a "*robust and flexible robot control system*" (Brooks, R. A., 1986). Brooks goes on to suggest that the behaviour-based system and its focus on behaviours which each connect input, processing, and output individually are more effective than systems

composed of "*independent information processing units which must interface with each other via representations*" (Brooks, R. A., 1991).

A more common example of a behaviour-based system is that used for flocking behaviours for boids (Reynolds, 1987), where three simple behaviours – separation, alignment, and cohesion – combine to create the complex flocking behaviours similar to that seen in the real world – see Figure 6-2. The boids example also illustrates quite clearly (though not in a simple image) some of the limitations of pure behaviour-based systems – notably difficulty with sequential tasks and the lack of planning or goals.

Figure 6-2 Flocking boids (Reynolds, 1987)

These limitations were recognised by a variety of groups (Bonasso, 1991; Connell, 1991; Gat, 1991), which independently came up with similar solutions – namely combining a behaviour-based system, with its advantages in reactivity and robustness, with a more traditional planning type system, and a middle 'sequencing' layer to connect the other two (Gat, 1998). These new architectures are hence termed three-layer architectures, or hybrid architectures. The three layers can also be defined by the content of their state –

the reactive layer has no state; the deliberative layer contains predictions about future state; the sequencing layer contains a history of previous state. Gat noted that "*the architectural guidelines that govern the design of the three-layer architecture are not derived from fundamental theoretical considerations. Instead, they are derived from empirical observations of the properties of environments in which robots are expected to perform, and of the algorithms that have proven useful in controlling them*", following with "*robot algorithms tend to fall into three major equivalence classes: fast, mostly stateless reactive algorithms with hard real-time bounds on execution time, slow deliberative algorithms like planning, and intermediate algorithms which are fairly fast, but cannot provide hard real-time guarantees*".

It can be seen from the developmental history of robot control systems that the development of ECA control systems is following a similar trajectory for similar reasons, and that it reasonable to suggest that ECA control systems will eventually also require hybrid systems in order to function effectively in the real world. Hindsight means that some of the steps in the development could be skipped – it is clear now that ECA systems should use hybrid architectures. In fact, behaviour-oriented and hybrid systems have already been suggested and developed for virtual humans (Bryson, 2003), though this is still in its infancy and has not been used for managing/processing complex input from the real world.

6.1 Proposed architecture

It is evident from normal human interactions and from literature that non-verbal behaviour is highly interactive. That is, people constantly and rapidly react to others around them, whether in conversation or just walking down a street. And as has been shown in Chapter 5, visual feedback is important in a conversational paradigm with a visual character in order for conversation to be persuasive. In other words, within a conversation characters should constantly and rapidly respond verbally and non-verbally

to their interactants. Present ECAs are significantly limited in their non-verbal response, especially the linking to their interactants' non-verbal behaviour. This follows both because they do not have the (complex) inputs available on which to base a reaction and also because they have an architecture that is not designed to react in a sufficiently rapid manner.

Also, much of non-verbal behaviour is semantically context-free (from the specific meaning of conversation) – much non-verbal behaviour occurs in a manner independent of *what* the conversation is actually about. For example, whether talking about the weather, what happened in the football match last night, or the state of the government, the majority of non-verbal behaviour in the interaction is the same – people still make appropriate eye contact, still provide and respond to turn-taking signals, and still nod along to provide encouragement. These behaviours that occur while the other interactant is talking, encouraging or discouraging or other modulating the interaction are called back channel behaviours (Yngve, 1970).

This second point makes the development of ECAs that have interactive non-verbal behaviour seem more tractable, as most of the behaviour will be the same whatever the conversation the ECA is involved in is about – the majority of non-verbal behaviours can occur without knowledge or awareness of the meaning of the conversation. In other words, the interactive non-verbal behaviour system may not need to keep track of more than the basic conversation state (who's talking, etc.), which fits nicely into the behaviour-based architectures paradigm. Higher-level behaviours can still be implemented as (more complex) behaviours within that same paradigm to create the desired hybrid or three-layer architecture. In the case of an ECA conversing about giving money to charity, the three layers would correspond to:

Reactive layer	Reacts to nods, eye contact, etc.
Sequential layer	Which conversation state the character is in (ECA talking, subject talking, etc.).
Deliberative layer	Decides what and when to say.

There are many different behaviours that occur during conversation, both verbal and non-verbal. The focus in this thesis is on non-verbal behaviours, and specifically those which appear to perform some form of function in conversation or portray (intentionally or otherwise) salient information. These behaviours include (aggregated from various sources including Knapp and Daly (2002), McNeill (1992), Efron (1941):

Speaking

Spontaneous gesture (including eye flashes, eye-brow flashes, head nods, speech emphasis (loudness, pitch, etc.))

Request turn

Accept turn

Deny turn

Maintain turn

Give up requesting a turn

Give a turn

Barge into conversation (verbal behaviour, though usually accompanied by non-verbal)

Give up barging into conversation

Gaze at (attention to face)

Mutual gaze (look at where interactant is looking)

Attention to object

Attention to element of interactant's body

Thinking

Expect turn

General attention

Mimicking/mirroring

Positive back channel (nods, paralanguage, simple language, facial expressions)

Negative back channel (shake paralanguage, simple language, facial expressions)

There are of course many other behaviours that may not serve any conversational or communicative role but that may add to the realism of a character, such as:

Self-adaptors
Attention to movement
Attention to bright things
Attention to flashes
Attention to noise (directional or otherwise)
Sway/minor movement
Blinking (though modulated by stress levels among other things)
Breathing (modulated by a variety of things)
Sighing (could be communicative)
Lip-licking (also modulated by stress)

Implementing all of, or even a significant proportion of, the above behaviours would be extremely challenging. The aim of developing a prototypical ECA with interactive non-verbal behaviour is merely to demonstrate that some (or more) of these behaviours may have an impact on the evaluation of an interaction. Furthermore, with the modular design of the behaviour-based system in such a hybrid architecture, further behaviours could be added incrementally and experimentally, adding to the non-verbal competency of an ECA. 93

Conversation state can be modelled quite simplistically as just, for example, Alice is talking, Bob is listening versus Bob is talking, Alice is listening, through to a more complete model such as that in Figure 6-3 below, where the various turn-taking states and their transitions are complex. This model was created using a series of Gedanken (thought) experiments of two people talking with each other. For example:

> Alice is waiting at a bus stop, while there she keeps an eye on the environment around her (*monitoring state*). She notices as Bob, whom she knows, walks up and she looks at him as he does (*attending*). Bob says "Hello", then asks how

Alice is (*Alice is listening (attentive)*). The end of Bob's sentence offers Alice a turn with Bob making eye contact (*turn offered*). Alice accepts the turn with her own eye contact (*accepting turn*), then speaks in response (*speaking (listener attentive)*), but after a while Bob isn't paying much attention (*speaking (listener inattentive)*) and soon Alice stops talking and reverts to just looking at Bob (*attending*), before finally returning to keeping an eye on the world around her (*monitoring*).

This model is by no means complete and is not presented as the only model that could be generated from a Gedanken experiment, but it does serve to both illustrate the complexity of turn-taking behaviours in human-human interactions, and provides a model of turn-taking that could be used for providing an ECA with more sophisticated turn taking behaviours.

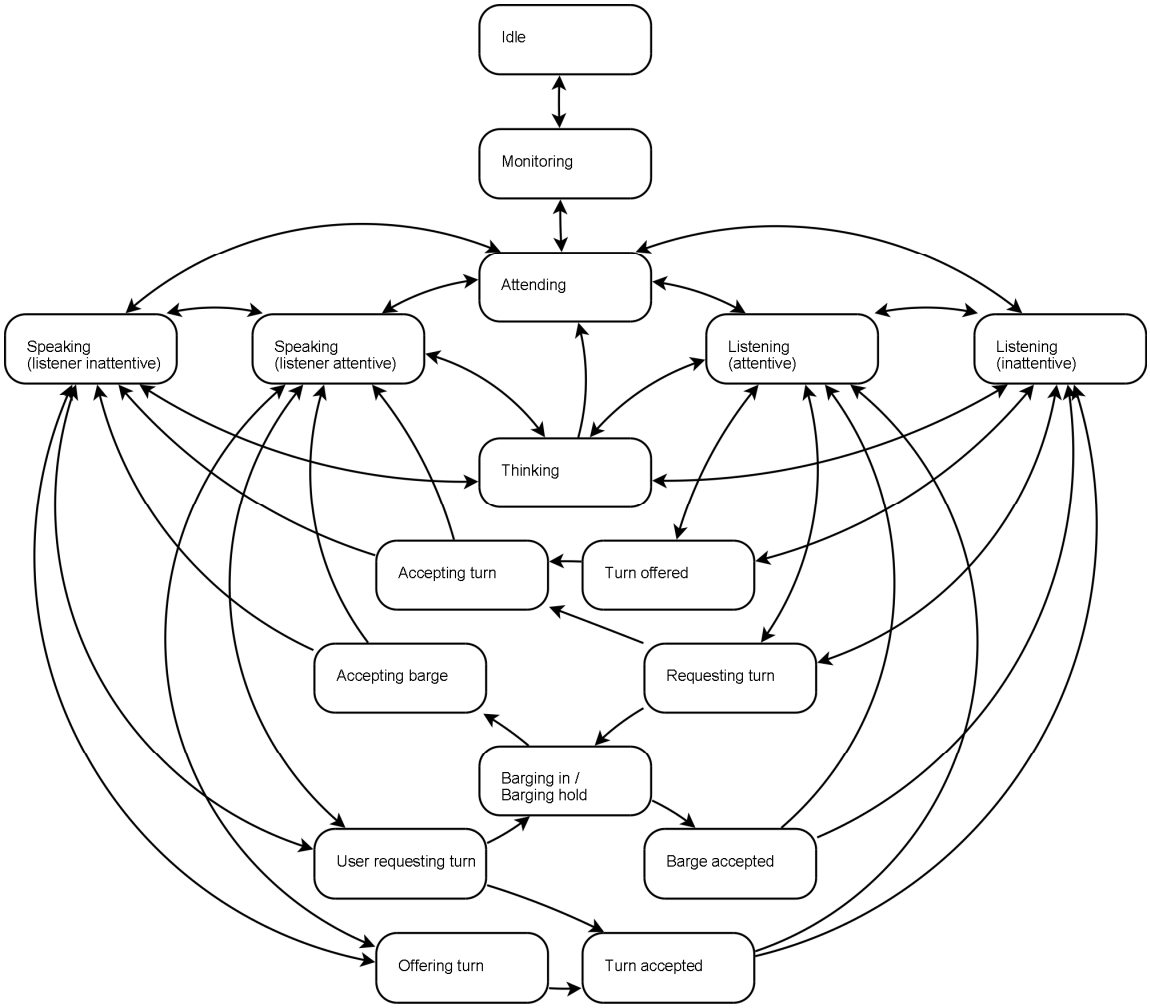


Figure 6-3 Conversation state diagram (complex)

Implementing a highly complex conversation state diagram is not appropriate for a prototype, and also the evaluation methodology introduces constraints on the conversation that simplify the state-transition diagram in Figure 6-3 significantly – not just reducing the number of possible states, but also significantly reducing the various transition causes to be detected and output behaviours needed.

In order to implement an architecture, a more definite picture of how that architecture will work is required. In this case, the important factors in implementation are the requirement of *fast* responses and multiple complex real-time inputs, and complex real-time outputs. These requirements correlate very strongly with those of multi-media

systems. In both multi-media systems and an envisaged interactive ECA system there is some form of streaming data inputs (audio, video, motion capture, etc.) which is then processed and/or combined in some form, before creating some other streaming output (audio, video, 3D graphics, etc.). These forms of architectures are called streaming architectures (also known as pipeline architectures, or filter graphs), and presently exist in a myriad of forms, such as DirectShow (Microsoft Corporation, 2007), GStreamer (Freedesktop.org, 2007), EyesWeb (Camurri et al., 2004), PureData (Puckette, 1996), Max/MSP (Cycling74), Isadora (Troika Tronix, 2008), vvvv (Meso).

Streaming architectures consist of a 'pipeline' of modules (also termed elements or filters) linked together so that they are collectively a process that transforms the input into a desired output. Katafiasz (Katafiasz, 2006) describes this in more detail with specific focus on GStreamer. More strictly this pipeline is a directed graph of modules, in which 'media' flows from input to output.

To date, streaming architectures have been specifically focused on processing audio and video media streams to create new audio and video streams. The approach suggested in this work is to extend the view of streaming architectures beyond audio and video into other forms of media – motion capture, 3D graphics, speech recognition, speech synthesis, etc. – to enable the development of more interactive ECAs (and other more complex forms of media). In that regard, some of the presently existing implementations mentioned above, such as EyesWeb and PureData, are already on this path. For example, EyesWeb includes many modules for vision processing that could be used directly to detect areas of skin in images etc. This raises the question of whether or not a prototype interactive ECA architecture should be developed using such a pre-existing architecture. This is discussed more fully in Chapter 7 on the ECA implementation.

The rest of this chapter discusses streaming architectures, their implementations, and some of their pros and cons in more detail, providing a stronger insight into how an ECA architecture using behaviour-based or hybrid systems may be implemented using these streaming architectures.

6.2 Streaming architectures up close

As mentioned previously, the terminology in streaming architectures has so far not reached a consistent standard. Furthermore, many of the terms are used differently in the same and/or other fields. This present discussion will use the words module, pin, link, and pipeline as defined in the following paragraphs.

A module is an object (inherited from a 'module' base-class) that takes some set of input, performs some processing on that input, and generates an output. With this in mind, there may any number of instances of a specific module type, such as two instances of a video rendering module, each of which creates a window on the screen with some video within.

Each module type defines a set of communication pins, each of which is either an input pin or an output pin, similar to the audio or video sockets on a piece of audio or video equipment. Each pin has a specific data type (such as image frames, audio buffers, integer value, Boolean values, text strings) that it will accept as input or create as output. From the abstract perspective, a module is not required to have either input or output pins, and may also read or write data from some other source. For example, a 'video source' module may have no input pins, and only a single output pin, which streams out data read from a video file. In practice, as with the previously mentioned streaming architectures, some specific input or output pins may always be required. For instance, this may be an 'active' input that controls (though a Boolean value) whether or not the module is active – i.e. creating any output.

The output pin of a module may be connected to the input pins of another module creating a link. Output pins may have multiple links (the stream can just be copied to multiple modules), but each input pin can/must have only a single link (a combining module would need to be used to combine multiple streams appropriately if needed). The type of an input pin must match that of the output pin it is connected to, and input pins may have default values that are used if they are not connected to any output.

Some streaming architecture implementations allow the dynamic creation of new pins and this creates the opportunity for 'magic' input pins that accept multiple data types. In practice, when linked these 'magic' pins create a new, appropriately typed pin (similar to generic functions in object-orientation), and it is this new pin that is linked to. Dynamically created pins also allow the possibility of modules with an arbitrary large set of inputs. For example, an audio mixing module could be created that mixes equally the audio input from however many audio streams are linked to it. The creation of new input pins may also create new output pins. For instance, a generic buffer module will create a specific type of input pin when its input is linked to another module (such as an image type), and will at the same time create an identically typed output pin (reflecting the buffered up previous images).

The set of specific modules and their links creates a pipeline. Frequently this is the finished product, but it should be noted that a pipeline is abstractly, and usually also in practice, a module itself – it has a set of input pins and a set of output pins. One of these 'meta'-modules can therefore be used to create more complex pipelines, which are also modules, ad infinitum.

Each module, in all the examples mentioned above, is generally connected to others at run time. Although this is not a requirement of streaming architectures, it makes them much more powerful and useful in practice, and some existing architectures allow the connection, disconnection, or reconnection of modules while they are active and (may) have data flowing through.

Modules may have parameters that they use to alter the processing or transformation that they perform, such as how much to blur an image, and/or internal or descriptive parameters created during processing, such as the size of an image. Most often these parameters are exposed to other modules as additional input pins (with default values), or output pins.

The implementation (in code) of non-meta modules may be arbitrarily simple or complex. A module may merely pass the input through to the output, or may perform

some highly complex processing. Anything that can be performed in code, with any additional libraries or external data or processing resources, is acceptable unless in some way restricted by the streaming architecture or the underlying system. Additionally, a module may maintain some form of history (such as a buffer module), or it may predict something about the future. This approach enables all the required layers of a three-layer architecture, namely, a reactive layer with no state, a deliberative layer with predictions of future state, and a sequencing layer with a history of previous state.

Trivially, one could view the already existing ECAs as a single module with no input or output pins, and could easily imagine some separation of one of those ECAs into some input modules taking input such as text, processing that input in a single 'cognitive' module exactly as it is now, and generating 3D graphics in response in a third module. In fact, it could be argued that many present ECAs already have this form of separation, but that is not specifically a streaming architecture, and therefore the power of having many smaller modules that may be combined into many different pipelines is not available with all the significant processing within a single module.

The envisaged system is, in fact, quite similar to the above view of present ECA systems, with the addition of some simpler modules that process video, audio, or other forms of more complex inputs, and use this also to drive a character. The complex 'cognitive' module influences the function of the lower-level modules (though their input pins) as and when appropriate.

The issue of one module affecting the function of another module introduces an important question beyond the scope of this thesis about priority and security of modules – which modules should be able to link to which other modules, should some modules be able to break links with other modules (possbily so they themselves can be linked), should modules be able to view the whole graph of modules and their links, etc. Given that at present most streaming architectures run on a single machine and that most modules employed are used over and over again and which ship with systems or software packages, this has not been a major issue except in terms of media copyright

protection. The problem with copyright material is that if an audio or video stream is decoded within a streaming architecture from an encoded source, then as soon as a module decrypts that data in a non-encrypted form for use by other modules, then that non-encrypted data becomes available to any module that may link to that decryption module. Of course, the decryption module could not create any output pins, it could directly send the video or audio to the graphics or audio card, but this would defeat the advantages of a streaming architecture. Without streaming, the decrypted video would still be available in RAM, but not trivially accessible as it would be in a streaming architecture. Similarly, a 'rogue' module could insert itself between two linked modules (moving the relevant links to itself) and manipulate the stream, such as by inserting advertisements into a video stream. The question of how various forms of security should be managed in streaming architectures is unresolved, especially in the area where modules may not all reside on the same computer. Further discussion of these security and access management issues is beyond the scope of this thesis, but the issues have significant importance if modules implementing various sections of ECAs or other systems are to be shared to aid the development of new and better systems.

In summary, a streaming architecture is proposed to allow ECAs to have more interactive non-verbal behaviour. Some presently existing streaming architectures already support arbitrary data streams, such as motion capture data or positional estimates of objects from video streams; others would need to be modified to allow this. Streaming architectures are nothing new. However, it is new to use streaming architectures as an integral aspect of ECAs to enable rapid responses to complex data, while allowing higher-level 'cognitive' module or modules, such as those that presently exist, to modulate lower-level modules. This latter point is neither a constraint nor an addition to streaming architectures, merely an approach to building them using a three-layer or hybrid approach.

To date, streaming architectures have not been used to control 3D characters other than directly to drive the position of a 3D character from computed 3D positions of various

elements of a subject's anatomy. In other words, beyond puppetry, streaming architectures have not been used as an aspect of a 3D character control system (or brain).

ECAs with interactive non-verbal behaviour could be developed without employing a streaming architecture: a streaming architecture is by no means required. However, it provides a simpler and more manageable approach to building interactive ECAs where the focus is more clearly upon creating interactive behaviours and enabling module re-use.

7. Implementation of architecture and of behaviours

The implementation of a prototype ECA using the proposed architecture to demonstrate the benefits that architecture provides is discussed in this chapter, along with the implementation of a set of behaviours to drive an ECA in that style. Architectures enabling interactive non-verbal behaviour have not been built before, and streaming architectures have not been used for ECAs before, so the prototype is designed to demonstrate and evaluate these options. Specifically, an ECA is prototyped to perform and be evaluated in the same 'giving money to charity' scenario as was used in the evaluation of synthetic ECAs. The advantages and disadvantages of using one of the various existing streaming architectures are discussed, and the chapter concludes that the next stages of development would be better integrated into an existing streaming architecture, though which architecture is not specified. Like the synthetic ECA, the prototype ECA using a streaming architecture was designed to work on standard consumer hardware. An overview of the data flow with the prototype is given, followed by a discussion of the implementation of data flow between modules.

The key non-verbal behaviours identified for implementation in the prototype are presented – namely, nod mimicry, conversation state control through affirmations (nods and short utterances) and interruptions (long utterances) – along with a variety of other design decisions. Each of the modules implemented is described with an indication of the events which each creates or responds to – specifically, the Wizard of Oz module sending events to the character modules at a users request; the speech detection module detecting speech in the audio stream using an energy thresholding approach; the eye tracking module integrating with the Tobii eye tracker SDK to determine presence and detect nods from the subject's eyes; the face detection module using the Haar classifier in OpenCV to determine the location of faces in video frames in order to determine presence and detect nods; and finally the character module that embodies a complex set of functionality.

The character module is not itself a streaming architecture, though it fits into one and could be converted into one, and as a whole maintains the state of the conversation and embodies both the high-level 'cognitive' planning of speech, the state of conversation,

and the low-level character animation with lip-sync. A brief overview of the script that controls the 'cognitive' behaviour of the character is given, along with the simple conversation state maintained by the character. Finally, the specifics of how the character is implemented are covered – the character rendering using OpenGL, the character animation using Cal3d including multiple animations for each behaviour and various background animations, the speech synthesis with lip-sync component of the character and its interface to the speech server through a caching proxy.

This chapter discusses the practical implementation of a simple ECA architecture in the style of streaming architectures, and the implementation of a set of behaviours to drive an ECA in that style. The previous chapter discussed behaviour-based systems and the need for an ECA architecture that supports interactive non-verbal behaviour. In other words, there is a need for an architecture that endows an ECA with the ability to respond rapidly and constantly to real-world inputs. Furthermore, the previous chapter suggested the use of streaming architectures to enable this, where the system is viewed as a set of data streams, and various modules that constantly process those data streams, to create output streams that drive a character. It was also discussed that some of the behaviours implemented by some of the modules may be modulated by higher-level 'cognitive' modules that perform more abstract functions on a slower timescale.

Architectures enabling interactive non-verbal behaviour have not been built before, and streaming architectures have not been used for ECAs before. The aim in this work is to demonstrate and evaluate these options through the implementation of a simple prototype. The aim is not to prototype streaming architectures for ECAs, nor is it module re-use. However, it is suggested that at further stages this would enable more rapid and effective development.

The aim is to build a system using the ideas from streaming architectures, and a pre-existing streaming architecture if suitable, to show that this style of approach to building ECAs is appropriate and/or effective. More specifically, the aim is to prototype an ECA to perform the same scripted conversation about giving money to charity as previously

used in evaluating synthetic ECAs, and then to use this same evaluation criterion to determine the efficacy of such a prototype. For this prototype the use of a strict pipeline architecture is therefore not necessarily required, merely the concept thereof. That said, the use of an existing streaming architecture could enable more rapid development and more code re-use, and allow more experimentation with the structure of various modules. The aim in this thesis is not to evaluate the various existing architectures, nor to determine which, if any, would be best for implementing a prototype, or further systems. Such an evaluation would be appropriate before future development took place.

For the prototype ECA it was decided early in the design process that a pre-existing streaming architecture would not be used. This was for several reasons. None of the existing architectures have modules for 3D characters or graphics. It was considered that a module or modules for this could be integrated into one of the architectures. The architectures which include video/audio processing (EyesWeb, PureData, vvvv, to name some) are those which are hard to integrate with, so it was decided the development effort required to build modules was too high compared to the perceived benefits. Both the GStreamer and DirectShow architectures are easy to create modules for, but at this point neither has pre-existing video/audio processing modules and the effort to build both video/audio modules along with a 3D character module was also too great. Given the aim for a prototype, not a generic system, overall it was decided that building modules for an existing architecture was excessive. Finally, the set of possible target architectures is reduced as some of them no longer appear to be under active development. The various architectures can be highly complex and therefore a simplified streaming style architecture was deemed most appropriate. The rest of this chapter discusses the implementation of this architecture, with focus on the development of the various modules.

7.1 Implementation

The prototype ECA system was targeted to run on normal consumer hardware in general, without specialised additional equipment such as motion capture, with the possibly exception of eye tracking hardware. The latter exception is due to the availability of eye tracking hardware in the development area, and also to recent progress in eye tracking from standard consumer webcams (though not to the quality of specialised hardware) (Chau & Betke, 2005; Li & Parkhurst, 2006; Li et al., 2005). Input on the human interactants (referred to as subjects here onwards) would be through audio (microphone) and video (webcam) of the subject, and the output would be a fully animated 3D ECA (referred to as character from here onwards), with lip-synched speech synthesis. The design also defined the option to have a Wizard of Oz to guide some of the interaction – a person behind the scenes who could control the behaviour of the ECA if required. In practice, for the experiments this Wizard of Oz functionality was not used.

Figure 7-1 following shows an overview of the data flow in the architecture – with data flowing from the Wizard, a microphone, a webcam, and the eye tracker, through various behaviour analysis modules, to a character animation module that generates the animation and speech synthesis. The character animation module in the figure is itself made up of a number of more specialised sub-modules, but was not implemented in a streaming architecture as the focus was on using a streaming architecture for interactive behaviour rather than the complexities of character animation (though a streaming architecture would also be appropriate for character animation and would allow better module reuse and aid collaborative ECA research).

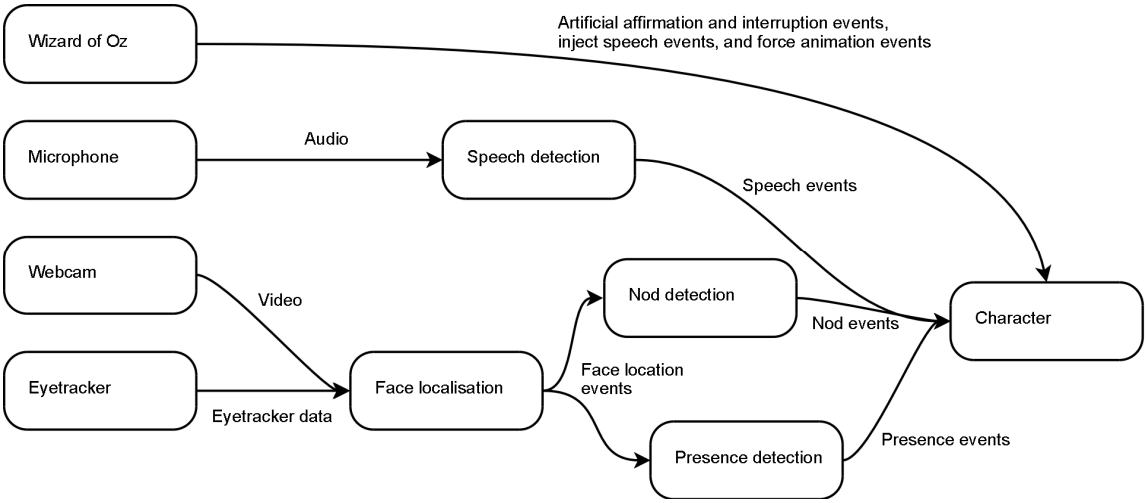


Figure 7-1 Data flow in prototype

In actual implementation the modules performing the first stage analysis (e.g. speech detection) were directly acquiring the appropriate input on the subject. In other words, the output-only modules, such as the microphone module, and the first analysis module were built as a single unit.

During design it was apparent that the various analysis modules created a heavy computational load. With this in mind and with network sockets being one of the easiest ways of communicating between separate programs on a single computer, it was decided that the various modules would send data to each other over (possibly local) network connections using UDP packets. Using UDP packets creates the advantage that each module can run without others being active (though it may just send data out into the ether, or have no incoming data to react to) or with dummy modules sending or receiving data. This makes development and debugging simpler. Furthermore, it makes the system more robust; as if one module fails the others merely stop receiving data from it. Using UDP data also means that the various modules can if required run on separate computers (and even on separate operating systems), thus spreading the computational

load. The matches with a streaming architecture with the UDP packets representing the stream flow along links between modules.

Similarly to the studies on synthetic ECAs previously described the prototype ECA was designed to talk through a script with a subject, giving information about a specific charity and charitable giving. At the design stage it was decided that:

The voice of the ECA would be created through high-quality realistic speech synthesis.

The character's mouth/lip movement would be synchronised with the speech.

In addition to movements and reactions to the subject, the character would have continuous background movement of head, eyes, arms and torso.

It was assumed that the subject would always be directly in front of the display and the camera and therefore it was not required for the character to look anywhere except straight ahead (where the subject was assumed to be). Adding a behaviour to track and look at a subject would be straightforward, but was simply not required in this context. With the experimental context it was also determined that the subject should start the conversation (rather than the system through presence detection); so that the character would not start the conversation while a subject was settling into the seat and getting comfortable. Without using speech recognition this would be difficult, so it was decided simply to use a button to start. Therefore, while it would have been straightforward to incorporate, presence detection (and response) was not implemented. Though the character state machine does have an inattentive state that would be used when subject is present, it is the transition from this inattentive state that is altered – occurring on a button press rather than on a subject becoming present.

In order to produce a working prototype, it was decided to focus on a few key features of interactive non-verbal behaviour, without attempting to develop a large set. These three key features were simple. Firstly, the character should be able to nod, with a variety of

different nods, to mimic any nodding behaviour of the subject, similar to previous work on the chameleon effect (Bailenson & Yee, 2005; Chartrand & Bargh, 1999; Lakin et al., 2003).

Secondly, the ECA's speech flow should be modulated by the subject's utterances and nods, mapping to a simple conversation state diagram (Figure 7-2). In other words, subject's short utterances or nods should be taken as affirmation, long utterances as interruption. While character is speaking affirmative behaviours should do nothing, while interruptions should cause the character to stop speaking. Once the subject stops interrupting the character will start speaking again, repeating the last phrase in progressively shorter versions. The character will also be able to ask questions expecting a few words of response. In this case, once the character has asked a question, the speech of the subject will cause the character to continue again (though once the subject is finished speaking), or wait for a timeout before continuing. Overall, the state of the conversation will influence the behaviours the character will perform, and behaviours by the subject will cause reactive behaviours by the character as well as possibly transitioning between states.

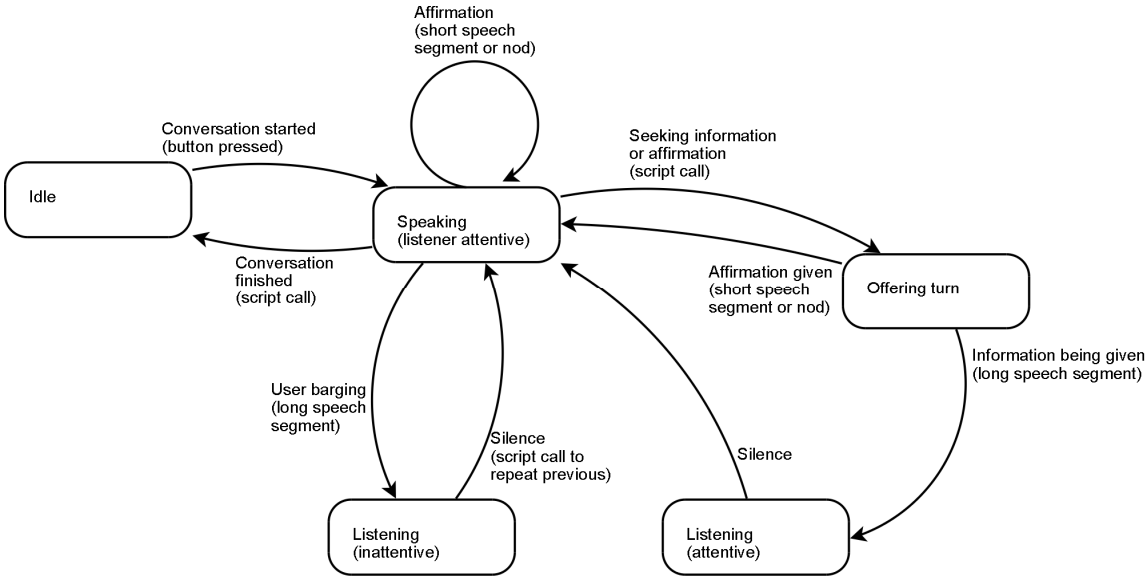


Figure 7-2 Conversation state diagram (simple)

The third and final requirement for the ECA was that it should be able to perform a clear finish animation to indicate the end of the interaction to the subject (in addition to this being stated through speech).

The rest of this chapter describes the implementation of each module in turn with some discussion of how modules could be enhanced. In this implementation and in most other streaming architectures, streams are actually implemented using an event based system – when a module creates a new output, it sends events to the linked modules to notify them. In this case the events are sent as UDP packets and include the relevant data (the event receiver does not need to collect the new data in response to an event).

7.2 Wizard of Oz module

The Wizard-of-Oz module is the simplest module developed. It is a simple user interface to send events to the character. It can send a speak text event to make the character speak custom text, affirmation events, interruption events, and events to cause animations (from a list) to play. Figure 7-3shows a screenshot of the Wizard-of-Oz interface.

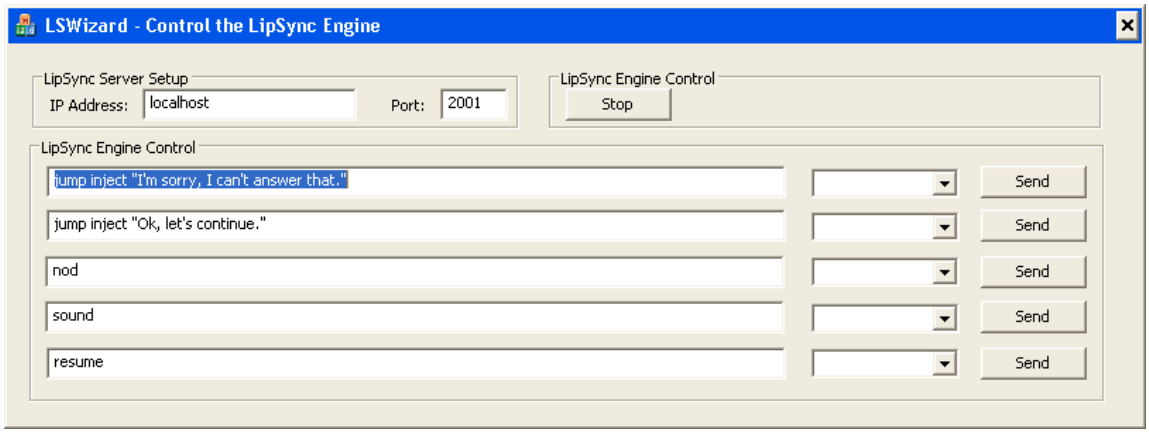


Figure 7-3 Wizard of Oz interface

7.3 Speech detection

Two different forms of speech detection were developed, both taking input from a desk-mounted microphone – a headset was considered, but dismissed because of the added impact on subjects. The first uses simply the energy threshold on the audio stream, while the second integrates with the Microsoft Speech API (Microsoft Corporation, 2008; Rozak, 1996) to use its speech recognition capabilities to identify utterances and also to identify entire phrases. This latter ability is not used, but shows that it would be straightforward to implement in the future. It should be noted that speech recognition technologies presently perform with a fairly high error rate, especially on general speech, from untrained voices, and in environments with significant background noise. Speech input was not a requirement for the prototype, only speech detection. It was found that the speech detection capabilities of SAPI were reliable, but that they used considerable computational power and had a longer delay than and were no more reliable than an energy threshold approach in the constrained environment such as was the target for the prototype.

The energy threshold approach calculates the root-mean-square (RMS) energy of a sliding window on the incoming audio stream to determine the energy of the sound at that time. Background noise creates a constant (though mildly varying) energy level which is below the threshold level. Speaking increases the energy level above the threshold and a speech event is created. Of course, this approach also detects other forms of energy in the audio stream and so moving a chair loudly or clapping can also create speech events. These issues are resolved in two ways. Firstly, the target context does not create much opportunity for additional sound other than speaking and does not encourage a subject to clap, and secondly, only speech sounds of certain durations creative affirmation or interruption events. Short sounds of around 100ms to 500ms create affirmation events, while longer sounds of 500ms to 2000ms create interruption events. Shorter sounds, such as claps, do not create any events that get passed on to the character.

More advance techniques speech detection techniques beyond simple energy in the audio stream could be used to increase robustness, such as restricting the frequency range to the 300 to 3400Hz used by most human speech. After a point, this becomes exactly the same approach employed by speech recognition engines to detect speech before trying to recognise it – known as voice activity detection. In the future it would be expected that using a speech recognition engine such as Microsoft Speech API or CMU Sphinx (Carnegie Mellon University, 2008a) for both speech detection and recognition would be more appropriate and that the delay in detection in speech recognition engines would be shortened.

Figure 7-4 shows an example of the command window output of the speech detection module using the sound energy technique showing the detected sound events.

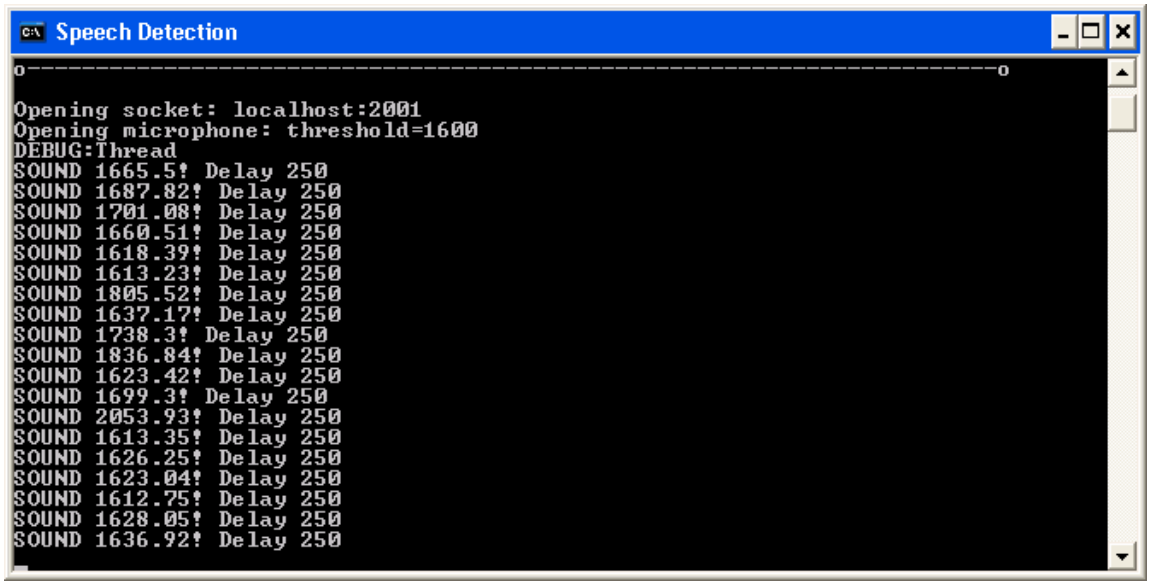


Figure 7-4 Speech detection command window

7.4 Eye tracking

The eye tracking module was developed using the software development kit (SDK) provided with the Tobii x50 eye tracker (Tobii Technology AB, 2006b) to interface with

that eye tracker. Once calibrated the eye tracker determines where a subject is looking on the screen 50 times a second, along with the 3D position of the eye. The eye tracking module uses this data both to determine that a subject is present (though this is not used) and to detect nods (turning points in the vertical position of the eyes) and then to generate events appropriately. In practice the eye tracker was not used because the nod detection was redundant given the webcam nod detection (see Figure 7-13 below) and the equipment was not standard consumer hardware, but it was developed at initially it was not clear that it would not be used and it also demonstrates the relative ease of which new modules may be developed.

7.5 Face detection

The open source computer vision library OpenCV (Intel Corporation, 2005) comes with many useful computer vision functions, specifically including a Haar classifier. This classifier determines a set of regions within an image that match a given Haar cascade (model), which is usually created from example images. For face detection, OpenCV already provides a set of Haar cascades to match faces from both a full-frontal view and a profile view (the latter was not needed in this research). The face detection module uses the Haar classifier of OpenCV to find faces in the video stream, and given the restrictions of the experimental area it was found safe to assume that it would only find a single face in the stream. It was therefore unnecessary to determine which face it should use or compute for multiple faces. The position of the face in the image is used to detect nods by detecting turning points' in the vertical position on the face within the image, reflecting vertical movement of the actual face. Haar classifying in this circumstance does not determine the position or orientation of a face with 3D, merely where in the image a face is, so the nod detector is only sensitive to vertical movement of the face, not rotation of the face around the neck which is a more significant change during a nod. Using only the position of the face to determine nods was found to be effective when the face was near the camera as it would be during the experimental interactions.

The face detection module generates two forms of events: face presence events, when a face is detected, and nod events when a nod is detected. An example frame of the face detection module is shown in Figure 7-5 below, with an example set of the command output from the face detection module in Figure 7-6.

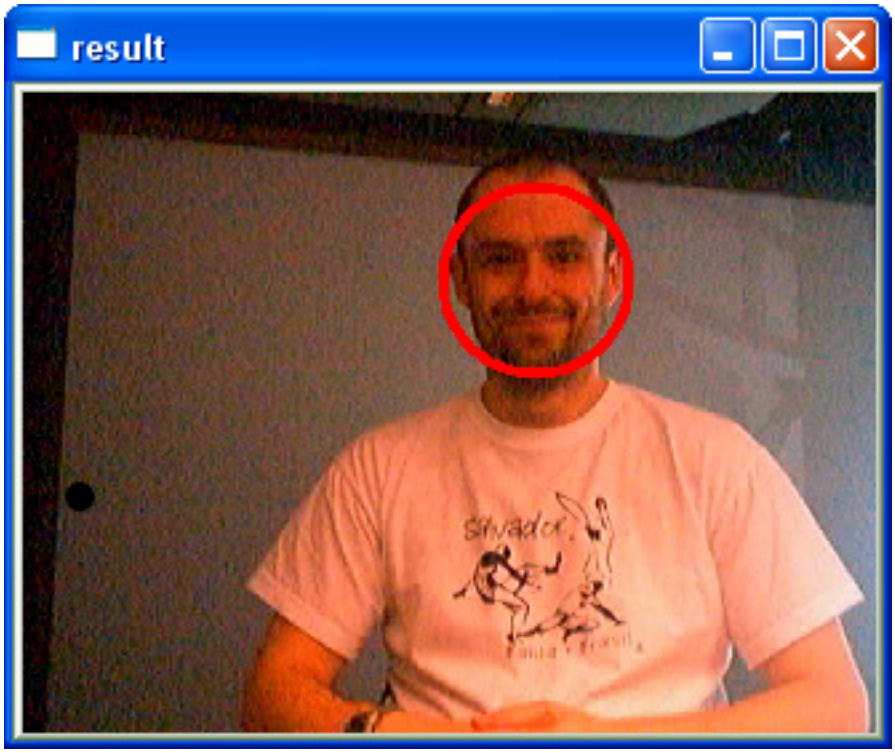


Figure 7-5 Face detection module

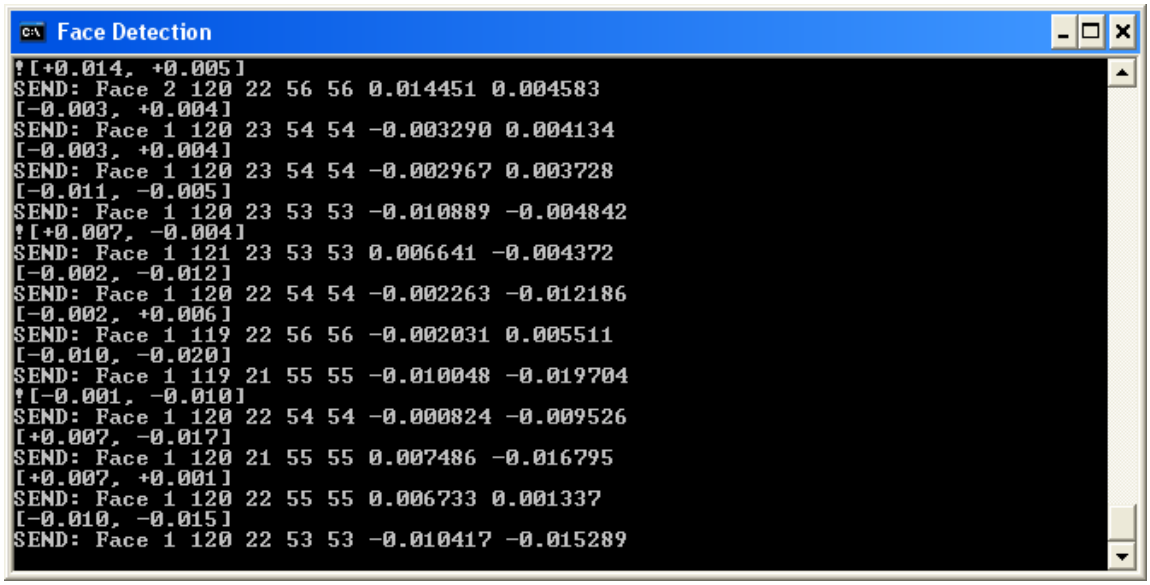


Figure 7-6 Face detection command window

Other forms of face detection and tracking were considered, specifically a simple skin finding technique that relies on the fact that skin *hue* is consistent across different lighting conditions and pigmentation. In other words, areas of skin can be consistently detected in images for all people and across varied lighting conditions. This approach was not used however, partly because the face detection technique is easily confused by neck, hand and arm skin regions and also because of the easy availability of Haar classifier in OpenCV.

The Watson (Morency, 2006) face detection and tracking library was also considered for use for the prototype, but it was found that while it was highly effective in detecting and tracking faces, including determining the full 3D position and orientation of faces accurately, it was relatively unstable and would tend to crash after 2 to 5 minutes of operation.

Marker-based tracking using a system such as Vicon (Vicon, 2005) or ArtTrack (Advanced Realtime Tracking GmbH, 2008) or using a coloured hat or the like was also considered but required the use additional non-consumer hardware such as IR cameras and markers (including hat). The area of marker-less motion capture or tracking is

presently an area of strong research focus and it is expected that it will be significantly more effective in the near future (Organic Motion, 2007). Marker-less motion capture has the advantage that subjects do not have to be augmented with special equipment, and frequently uses standard video streams, so less specialised equipment can be used. Furthermore, marker-based motion capture usually has to occur in a controlled environment, while marker-less motion capture can take place in the natural environment.

7.6 Character

The character module used in this research consists of a variety of sub-sections but is not strictly a streaming architecture. The character module maintains the state of the conversation and embodies both the high-level 'cognitive' planning of speech, the state of conversation, and the low-level character animation with lip-sync. These could be separated into separate modules to be closer to a streaming architecture, but the focus was on the behaviour modules as a streaming architecture rather than on the character animation. The character module accepts a variety of events from other modules, including affirmation and interruption events to affect the conversation flow, presence detection (not presently used), and requests to speak specific text or perform specific animations (from the Wizard of Oz). The character module assumes the subject is directly in front of the display (and the camera) and therefore does not use any position of the subject's face to affect where the character looks. Details of the various components of the character module will be discussed in more depth in the following paragraphs.

The 'cognitive' behaviour of the character is determined by a script (using a simple custom scripting language) that defines what the character will say, what states it will go into, and which events it will await to transition between states, and how long timeouts should be to transition between states if no event occurs. It also shows what should be said if speech is interrupted: when interrupted the character moves on through a list of

ever-shorter versions of the same statement or paragraph, repeating only when the last one is reached. This means that if the subject interrupts the character it doesn't repeat what it just said, rather it says it again in a shorter and shorter form as real people do. The character does not track how many times it has been interrupted and has no emotional model so it doesn't get angry or exasperated when it is frequently interrupted, but it would not be a major effort to create this form of interaction. An emotional model would simply be another module that takes appropriate inputs (for instance, interruptions making it less happy, and affirmations making it more happy), and the internal emotional state then used to create appropriate outputs affects other modules, such as the expression on a character's face, or the volume or rate of speech. The emotional state could even affect the conversation state by, for example, moving the character to a 'sulking' state. Emotional modelling is not within the scope of either the prototype or this thesis overall, but is obviously highly relevant to interactions with real people, and would add important realism and complexity.

An example section of the script (fully given in Appendix C2) used for the experimental scenario is given in Figure 7-7. It should be noted that this script does not control how the non-verbal behaviours work, what they respond to, or what animations they trigger, but only what the character will say with some control of the conversation state. The conversation state, in turn, determines which of the behaviours will be active, with each behaviour generating events when it senses appropriate input.

```
# Set script delays based on input choice
if $inputs = 0 set sectionWaitTime 1.1
if $inputs = 1 set sectionWaitTime 1.5
if $inputs = 0 set speakWaitTime 0.5
if $inputs = 1 set speakWaitTime 0.5


# (introduction)
call say "Hi, my name is Alfie what's yours?" "Sorry,
what was your name?"

# (wait for response)
state 1
```

```
delay 0.2
if $inputs = 0 wait 4
if $inputs = 1 wait 4

state 2
call say "Hi there." "Hi"
call say "I'm here to talk with you about donating
money to charity." "I'm going to talk about donating
money to charity."
```

Figure 7-7 ECA script sample

Figure 7-8 shows a screenshot of the character command window, showing the loading of various components of the animated character, while Figure 7-9 and Figure 7-10 show that same command window while a subject is interacting with the character, with the character ignoring and reacting the subject respectively.

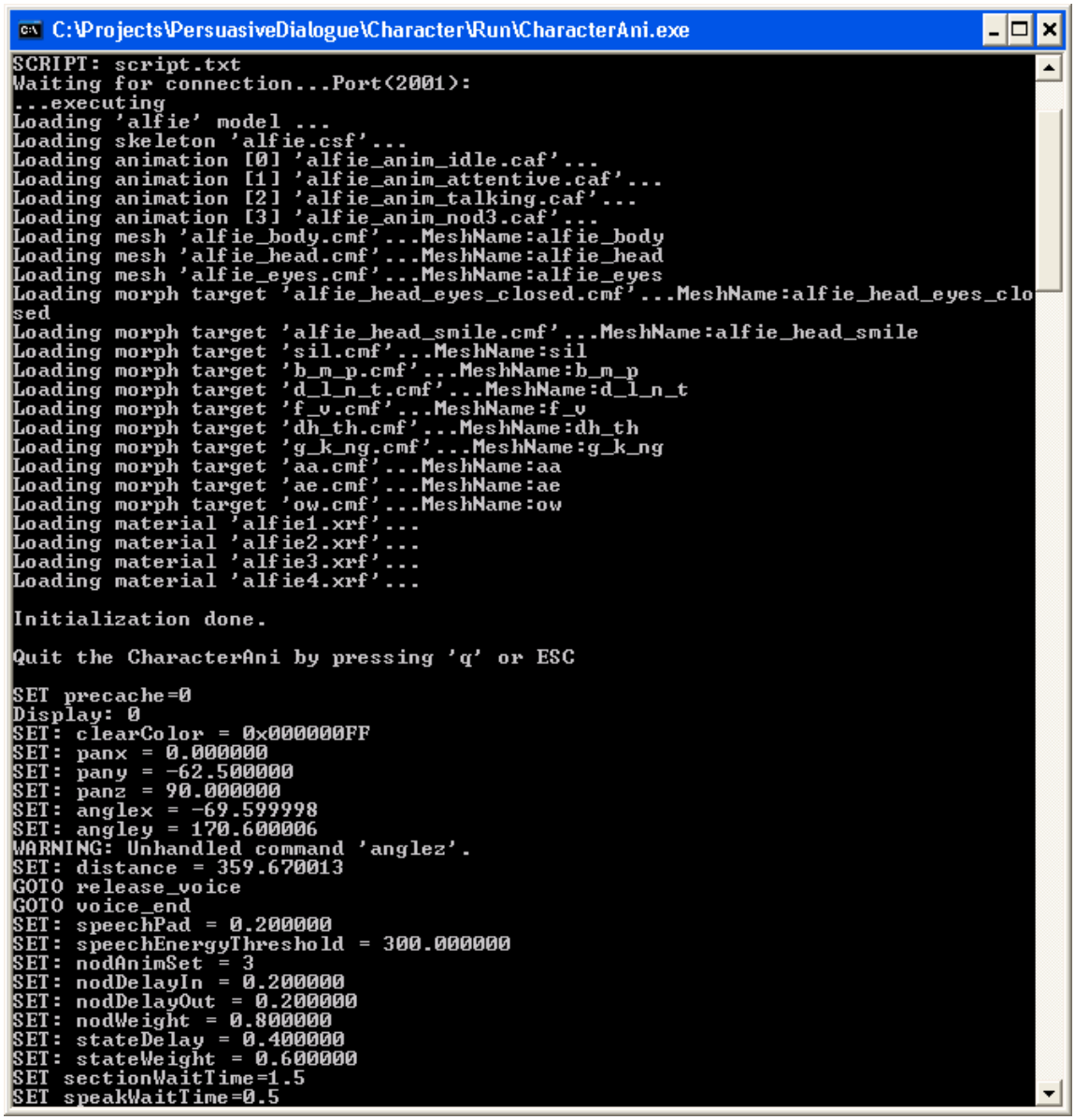


Figure 7-8 Character command window (loading)

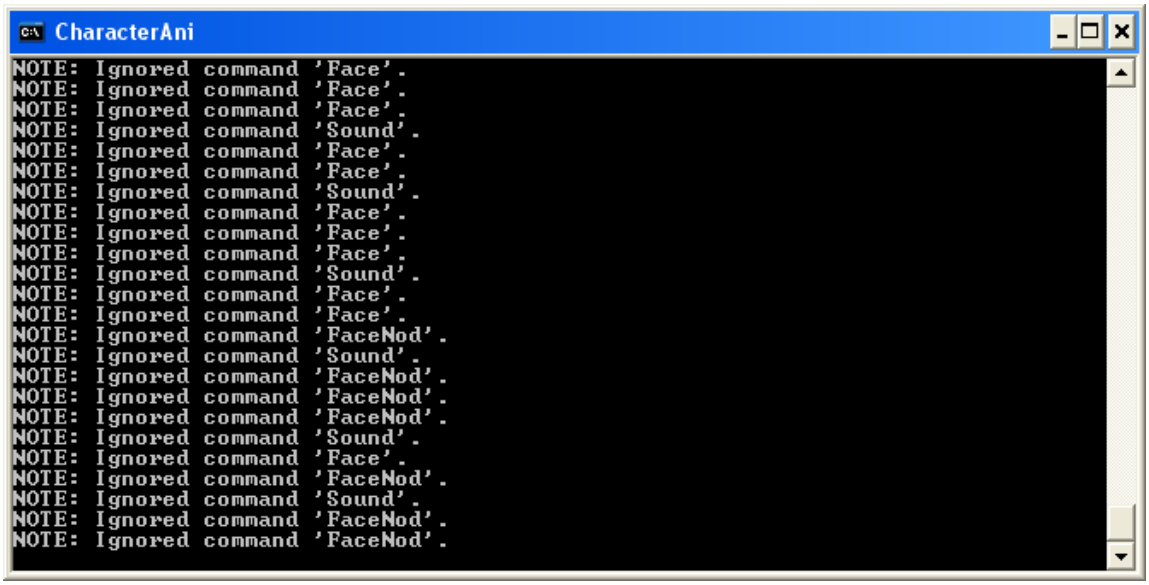


Figure 7-9 Character command window (Character ignoring)

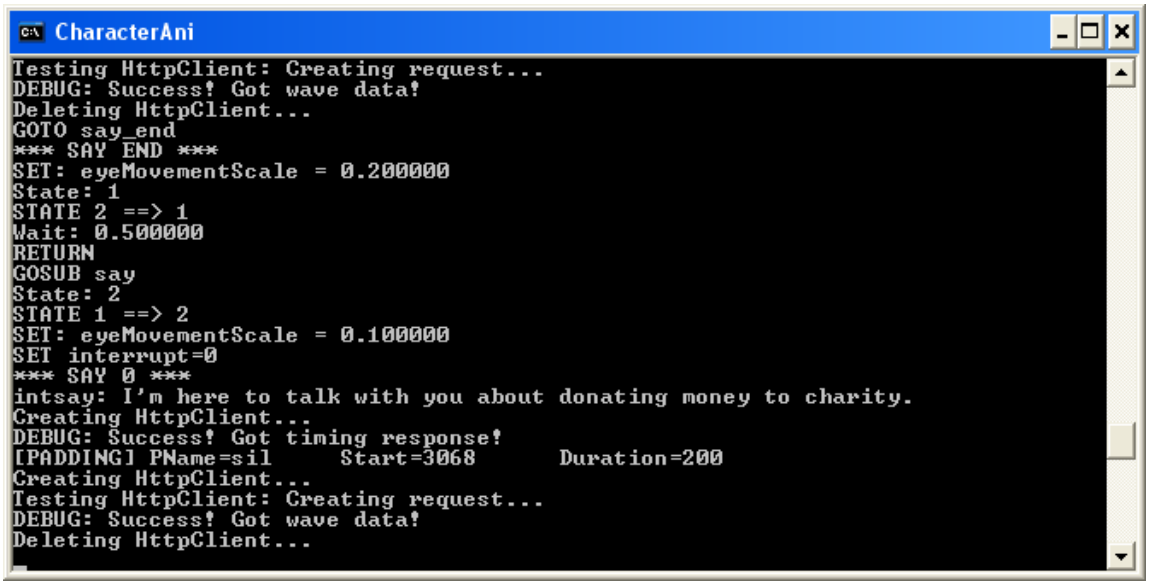


Figure 7-10 Character command window (Character not reacting)

7.6.1. Character rendering

The character is rendered in full 3D graphics using standard OpenGL (Khronos Group, 2008), with the positions and orientations of the skeleton determined by the animation

framework (cal3d) discussed in section 7.6.2. The character model and textures were originally created in 3ds max (Autodesk, 2008) for previous work (Lexicle.com, 2005). The character model used was a cartoon-styled 'mad professor' model, called in this thesis Alfie – see Figure 7-11. This was chosen partly due to its availability (available for use within the school of Computing Science at Newcastle University), but more importantly because the cartoon styling lowers subjects' expectations of the character and bypasses the 'uncanny valley' effect (Mori, 1970) that seemed to be present during development of the real ECA, after the synthetic ECA studies were complete, when a more realistic character was used. This effect was not evaluated empirically but was based on observations made during development. In depth discussion of computer graphics and various techniques therein are beyond the scope of this thesis. The aim of the prototype was neither photo-realism, nor to work at the cutting edge of computer graphics – merely using standard computer graphics techniques to create a 3D character.

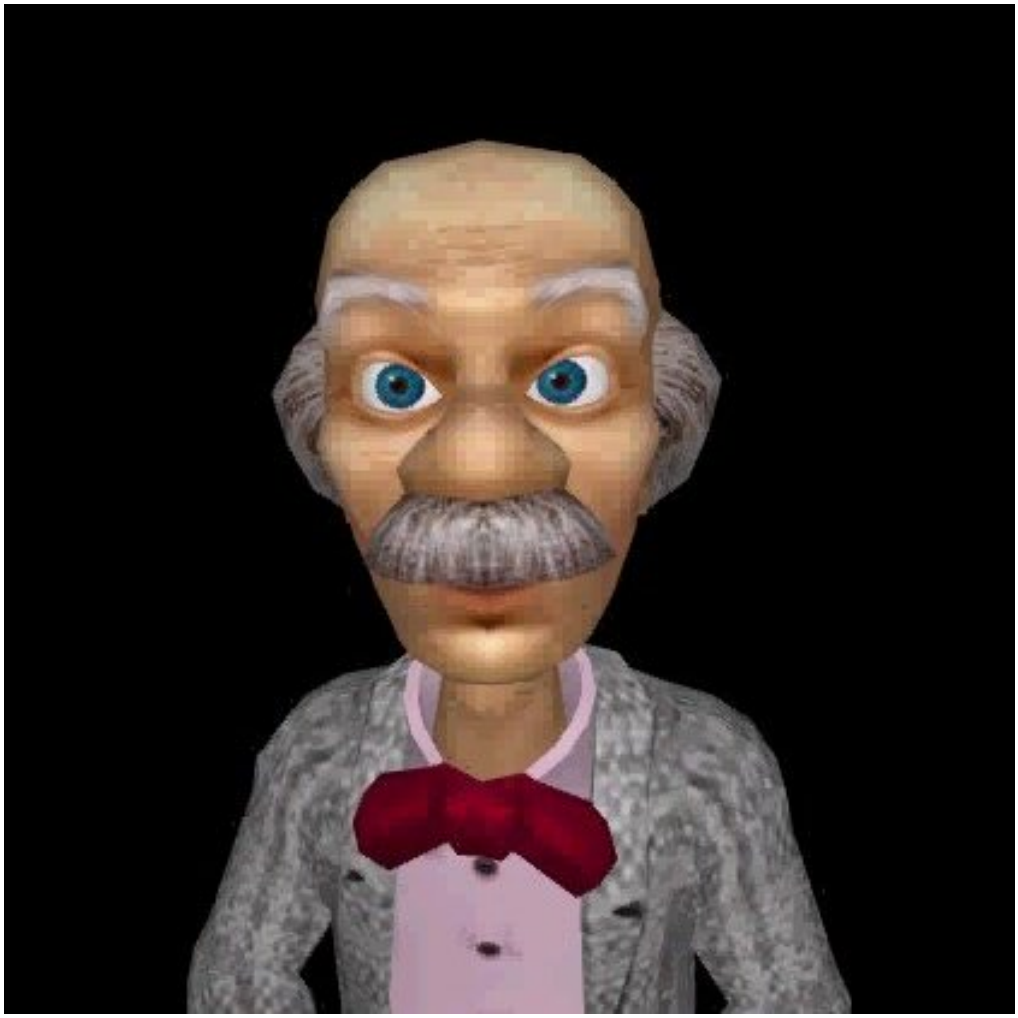


Figure 7-11 Alfie character

7.6.2. Character animation

The position and orientation of the character's skeleton were managed using the character animation library Cal3d (Laurent & Dachary, 2008). This enables playback of and blending between multiple animations. All character animations were generated off-line from pre-captured motion capture of real people during conversation. This motion capture had been done previously with real people describing cartoons for use in experiments into gesture. No motor planning was performed as this was beyond the scope of the prototype; playback and blending of pre-existing animations of non-verbal behaviour were sufficient for prototype purposes. It can also be clear that much of real

people's motion, rather than being motor plans created at the time, is merely playback of pre-existing skilled motor plans, as first recognized by James (James, 1890; Schmidt & Lee, 2005). For example, when Alice throws a ball to Bob, she doesn't plan the action; she simply plays through a pre-existing 'throw ball' motor plan based on having acquired appropriate muscle memory. The use of motor planning could be straightforwardly integrated into the present architecture either through allowing a motor planner direct control over some or all of the skeleton, or by the motor planner creating a new animation representing a new motor plan and then playing that animation along with the presently existing animations. The non-verbal animations of the character are triggered by events from the behaviour analysis modules, and there are multiple different animations for each event in an animation library. When a behaviour analysis module triggers a non-verbal animation one of the appropriate animations is chosen at random. For example, when the character detects a nod and is in a state such that it will mimic a nod, it starts one of three different nod animations. The film-strip shown in Figure 7-12 shows an example of the character's non-verbal behaviour during an interaction. The amplitude of animations can also be controlled in cal3d, though this feature is not used. This could be used simply to make the character perform 'bigger' non-verbal behaviours in responses to 'bigger' events. For example, larger nods by the subject would create larger mimicry nods by the character.

In addition to non-verbal behaviour responses, the character also performs a variety of background movements. As with the non-verbal behaviours, these are also animations generated from motion capture of real people, but, in this case, while those people are not talking. There are two sets of these animations, those occurring while the character is attending to the subject – listening to the subject – and those occurring while the character is paying no attention to the subject – looking around the room. These two sets of animations are used according to the conversation state, and the sequence of animations within each set is random, so the character doesn't appear to cycle through the same behaviours over and over again. An example of the background animation while idle (not attending to the subject) is shown in the film-strip in Figure 7-13.

Figure 7-12 Filmstrip of Alfie character during an interaction

Figure 7-13 Filmstrip of Alfie character while idle

7.6.3. Speech synthesis

The speech component of the character is the most complex component. When the ECA is requested to speak either by the script or by an external event (from the Wizard of Oz), the text of that speech is first send to a speech server, which then returns an audio file of the speech along with the timings of both the words and more importantly the phonemes (sounds of speech) of the speech within that audio file. The given phoneme timing sequence is used to create an appropriate timing sequence of visemes (mouth shapes each corresponding to one or more phonemes). There are 28 different phonemes in the English language, mapping to 22 different visemes (Long, 2002) – some phonemes sound different but have the same mouth shape. Many visemes look similar and a reduced set is therefore used for the prototype character. The audio file is then played, while the viseme timings are using to trigger morph targets (alterations to a 3D mesh) on the character's mouth appropriately. This lip-sync is demonstrated in the film-strip in Figure 7-14

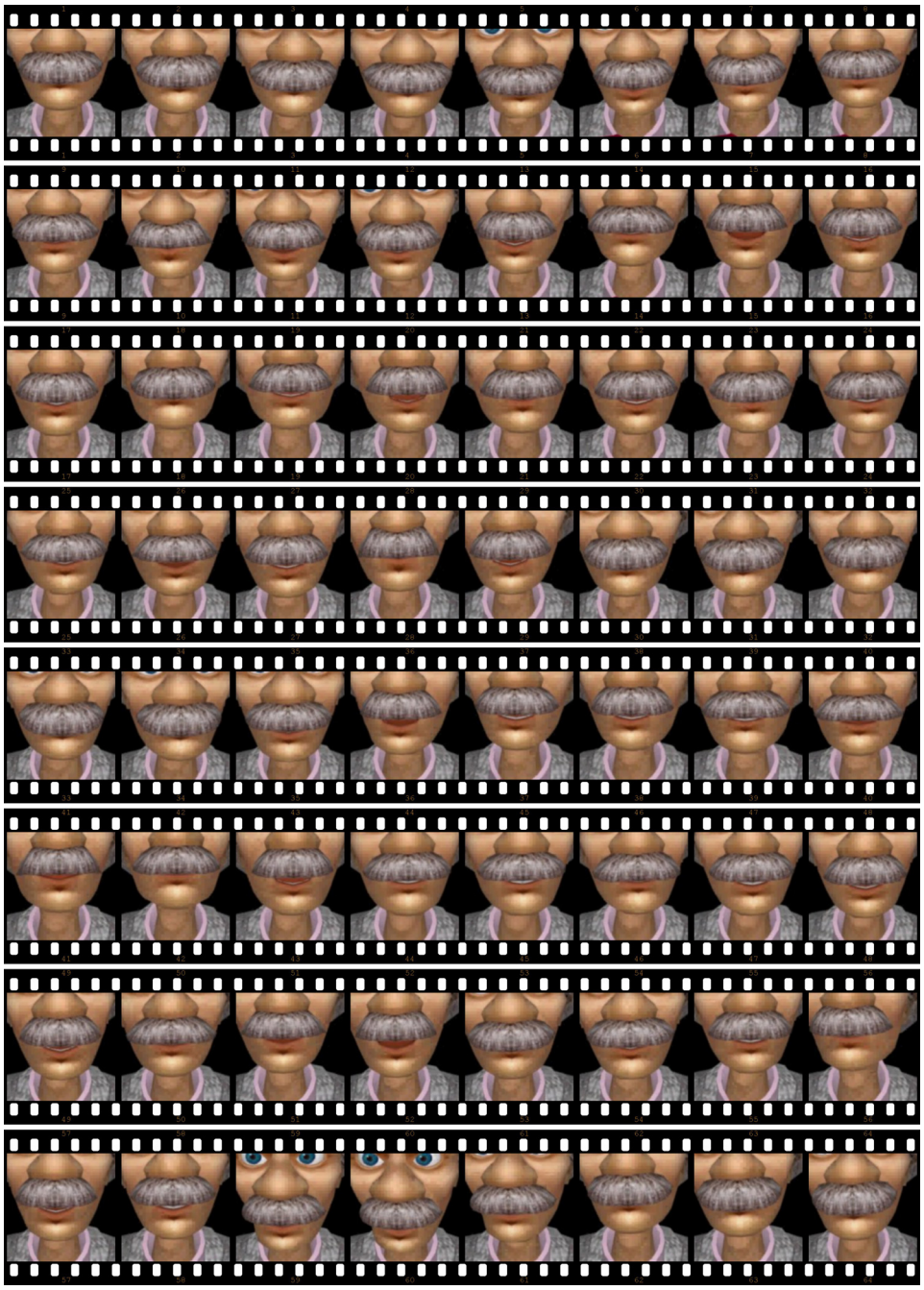


Figure 7-14 Filmstrip of Alfie character's lip movement during an interaction

The speech server is a custom wrapper around a variety of different speech engines, providing a uniform interface. The speech server can use a variety of speech engines with different APIs, such as Microsoft Speech API (Microsoft Corporation, 2008), to create speech. For the prototype a high quality voice – rVoice – from Rhetorical (Rhetorical, 2002) was used. The speech server abstracts the differences between engines so changing from the Rhetorical voice to another such as SAPI is merely a matter of requesting the speech server to use a different speech engine. Figure 7-15 shows the command window output of the speech server when as it receives and responds to speech requests.

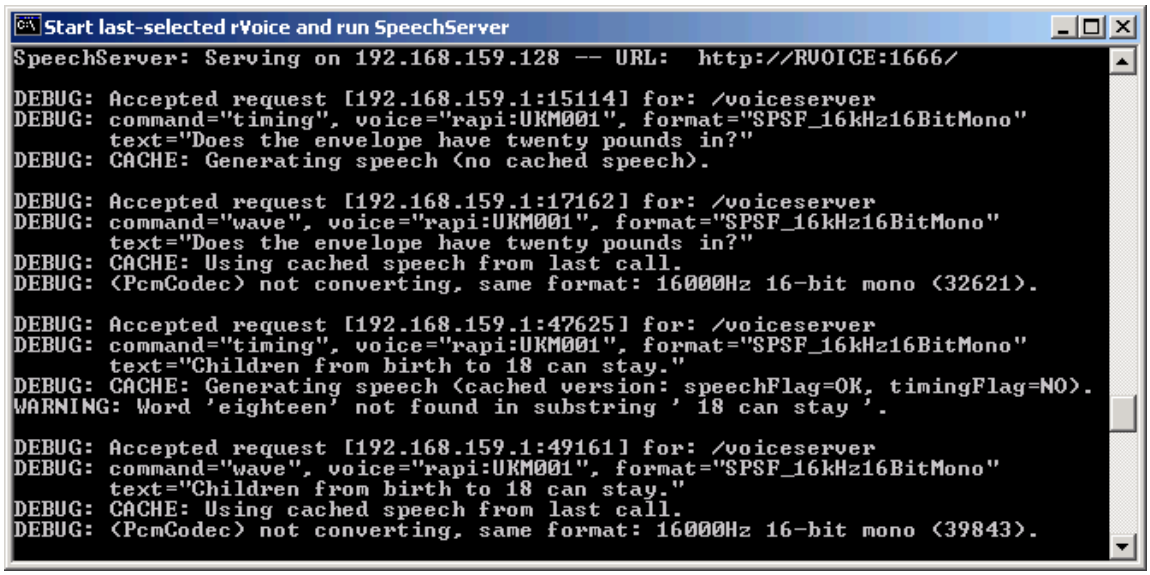


Figure 7-15 Speech server command window

Speech synthesis creates a significant computational load, which is evident both through the effect on the playing animations of the character (character's movements become jerky), and through the delay in a response from the speech server. In order to resolve this issue a caching proxy was created to cache the results of speech synthesis requests so that if the same request were made at a later date, the cached result could be returned without having to generate it all over again, thus saving the computational load and returning in a more timely manner. Within the experimental context the character says more or less the same thing to each subject (depending on how much the subject

interrupts), and therefore all the speech that will be requested can be pre-cached, so the computational load of the speech synthesis is not evident. Using a caching proxy means that if new speech is required, possibly if requested by the Wizard, then the cache proxy will pass on the request to the speech server and new speech will be generated. Figure 7-16 shows the caching proxy in action, with both cache 'hits' and 'misses'.

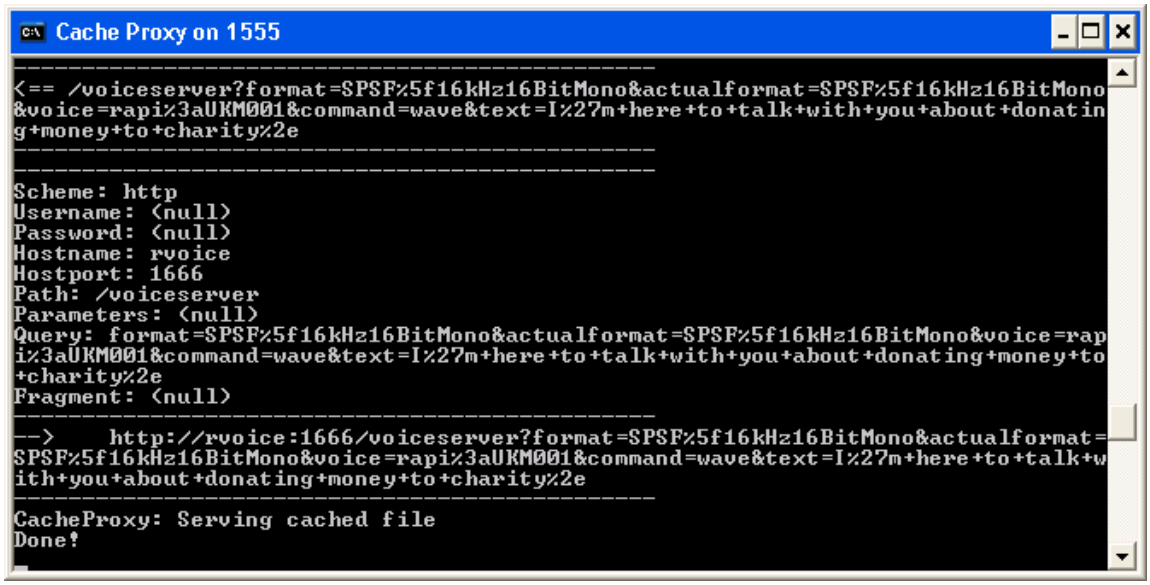


Figure 7-16 Caching proxy command window

7.6.4. Summary

The developed prototype character with streaming behaviour framework creates a fully animated 3D character with high quality speech with lip-sync. The character responds in its limited ways to non-verbal behaviour on the part of a subject, and allows interruptions and responds to affirmative utterances. The script allows the *cognitive* behaviour of the character to be changed easily, and both the *cognitive* and the non-verbal behaviours of the character can be altered independently. The streaming type architecture using UDP packets to send data allows modules to be updated, changed and reloaded easily and if desired even at run-time. In this implementation the UDP data uses only a very small proportion of the available network bandwidth. The prototype architecture is only a prototype and is not designed to be easily reusable or particularly

generic, though within its constraints it is flexible and robust. Further development of a streaming architecture for ECAs would be best pursued using a pre-existing streaming architecture. The choice of architecture is beyond the scope of this discussion, but some of the important factors are the ease of development of modules, the flexibility of the architecture, which platforms the system is required to run on, and the target audience. The chosen character appears to be mildly engaging and appears to respond to a subject's behaviour. Evaluation of the prototype, specifically for its persuasive effect, is covered in the following chapter.

8. Evaluation of behaviour-based architecture for an ECA

Empirical evaluations of ECAs provide strong evidence of their utility and other values, help validate underlying techniques used to develop those ECAs, and provide indicators for future developments. The behaviour-based architecture introduced in Chapter 6 and implemented in a prototype ECA in Chapter 7 is evaluated and discussed in this chapter. The reasons for using a direct measure of behaviour change are highlighted along with the need to evaluate a developed system to determine its efficacy, and to assist in further development. The experimental design for the prototype evaluation is given, using the same evaluation approach as the synthetic ECA studies – namely the 'giving money to charity' scenario. Full details of the experimental procedure are given, along with the procedure each subject went through. The measures taken and results obtained are discussed and non-evident differences between the two conditions on the direct measure of behaviour change are discussed. Conclusions from the prototype study are given, along with recommendations for new experimental protocols that might increase effectiveness.

The earlier study (Chapters 4 and 5) of synthetic ECAs indicated that an ECA that responded interactively to a subject's non-verbal behaviour would be more persuasion. This motivated the design of a behaviour-based architecture to enable an ECA to have those responses, and the implementation of an actual ECA system using that architecture that might be more engaging and have more social influence (as measured by persuasive impact), but the proof is of the pudding – does the interactive non-verbal behaviour in the developed ECA make the ECA more persuasive or more highly rated by subjects?

As discussed previously, most evaluations of ECAs, whether for evaluating persuasiveness or other social effects, have been based on questionnaires or structured interviews (Bailenson & Yee, 2005; Keeling et al., 2004) – measuring persuasion indirectly. As far as the researcher is aware, no studies have evaluated the persuasive effect of ECAs using a direct measure of persuasion – as defined as a difference in behaviour over a set of conditions.

The evaluation of the persuasive effect of the implemented ECA used the same approach as the persuasive effect evaluation of the synthetic ECA – i.e. to measure behaviour change (over each subject group) directly by giving each subject the opportunity to donate money from their payment to charity after an interaction with the ECA. A questionnaire was also used to elucidate the subjective views of subjects on the character and their interactions with it.

8.1 Experimental Design

The evaluation compared two conditions. Under condition 1 the ECA ignored all inputs about the subject's behaviour, so therefore could not react to the subject. Under condition 2 the ECA took cognisance of the inputs and could therefore react to the subject. The hypothesis was that under the second condition the ECA – by reacting to the subject – would be more persuasive, as measured by how much of the amount paid to each subject was given to the charity (across the whole subject group) on departure, and that subjects would rate the interactive ECA more highly on the questionnaires.

8.2 Subjects

Subjects were recruited from Newcastle University and were all post-graduate students or university staff. The condition under which the ECA was operated was determined at random by software and written to a log file. The studies were double-blind – neither the subjects nor the experimenters knew which subject belonged to which group until all the data had been recorded. Only after completion of all studies and recording all data into SPSS (SPSS Incorporated, 2006) was the log file accessed to determine which condition each subject had been exposed to.

8.3 Wizard behaviour

As in the synthetic ECA persuasion study, the character was presenting information to the subjects about a specific charity, and giving the subjects the opportunity (anonymously from the character's perspective) to donate to the charity. The character was not actively seeking to persuade the subjects, but merely presenting information about the charity.

8.4 Procedure

Each experiment consisted of a series of steps for each subject. Each step gave instructions to and for the next step, and additionally the experimenter gave subjects the full set of instructions on all steps at the start. For the duration of the each experiment, subjects were self-guided.

The first step was a Myers-Briggs (Quenk, 2000) personality type test that took the majority of the time. This was a distraction task to prevent subjects from being focused on the interaction with the character as the main important section of the study. This personality-type data was not used.

The second step (the interaction with the ECA) took place at another desk under one of the two conditions. On the desk were both a webcam and a microphone (to supply data to the analysis modules for the ECA). The character appeared life-sized on a large screen immediately across the desk and subjects could hear the character through loudspeakers. This setup is shown in Figure 8-1 Subjects were able to see the head and shoulders of the character. The ECA appeared male under all conditions.

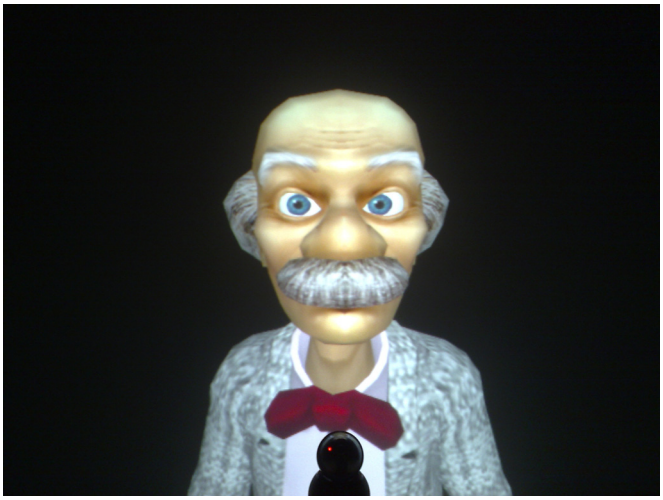


Figure 8-1 Alfie Embodied Conversational Agent in situ

It should be noted that the modules analysing the behaviour of the subjects were still active under both conditions, the only difference being whether the ECA reacted to them or not. This ensured that any difference between conditions was not due to the considerably different computational load between having the analysing modules active and inactive causing lag or other unwanted effects.

Subjects were instructed to press a button on the desk to start the interaction with the ECA (see Appendix C1). The ECA then asked some general questions about the subject (such as their name), told the subject that their payment for the study was on the desk in an envelope (£20 in the form of 8 £2 coins and 4 £1 coins), and asked them to check the money. The ECA then went on to present information about the charity. Finally, the

ECA explained that after the interaction the subject could donate some of their £20 payment for participating in the study to the charity if they chose to. The ECA then disappeared from the screen and subjects could, if they felt so disposed, donate some of their payment to the charity by placing coins in charity box on the table.

The final step of the study was, as before, a follow-up questionnaire (paper-based) consisting of a set of statements using a 5-point Likert (1932) scale ranging from -2 (strongly disagree) to +2 (strongly agree), with an opportunity to add open-ended comments. This questionnaire is given in Appendix C3.

8.5 Measures

As with the previous synthetic ECA persuasion study, the main measure was the amount of money donated to charity by each subject. Additionally, and again as with the previous persuasion study, there was a follow-up questionnaire, consisting of statements concerning the nature of the interaction and the subjects' beliefs about the ECA. For this new study, a number of questions were added to the questionnaire about how persuaded the subject felt. These questions were added because it seemed likely that the difference between the conditions of the later study would be less than between the conditions for the synthetic ECA and so were designed to detect more subtle differences – differences in subjective opinions, rather than actual behaviour.

In addition to the above measure all interactions were recorded from both points of view – the webcam and microphone footage of the subjects and the 3D character output (through screen capture) and also with audio. These recordings were for logging and post-experimental subject analyses purposes only – they are not direct metrics, though conceivably certain metrics could be calculated from them – such as the number of nods detected by the character for each subject across the conditions.

8.6 Results

A total of 47 subjects participated in the study and were neither age nor gender balanced. Access to sufficient subjects was limited, especially as the local undergraduate population was avoided due to the belief that they would (anomalously) not donate much at all to the charity. The character ignored 24 and reacted to 23 of the subjects.

For the main measure of amount of money donated to the charity, the data indicates no significant difference between the two conditions – means of £6.17 and £6.70 for ignoring and accepting inputs, respectively. A Kruskal-Wallace test of significance gives the chance of the difference between the means occurring by chance at 0.812. In other words, it is very likely that the difference is just by chance. The ECA when ignoring inputs condition has a larger variance than the ECA when accepting inputs. The distribution of donation amounts was highly non-normal across all conditions. The cross-condition data is shown visually in Figure 8-2 and Figure 8-3.

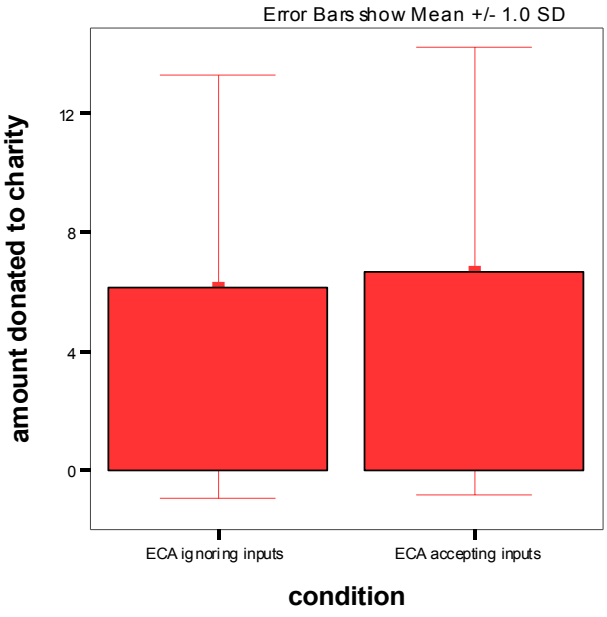


Figure 8-2 Amount donated to charity across conditions

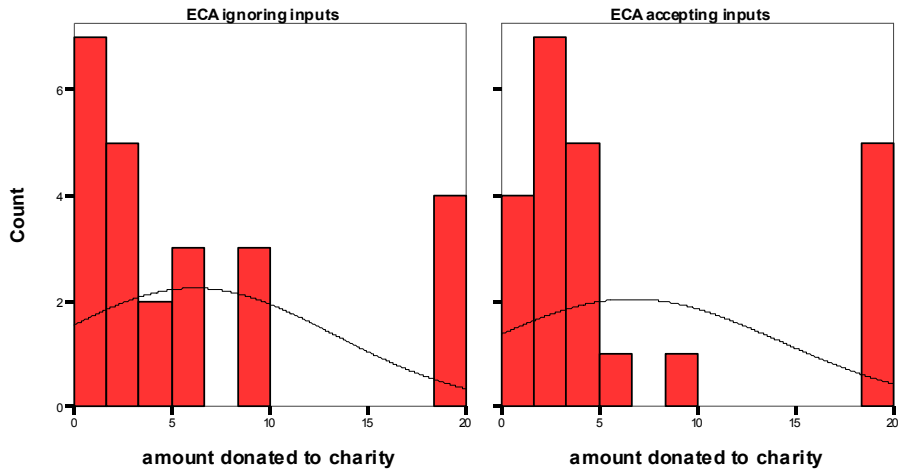


Figure 8-3 Histogram of amount donated to charity across conditions

This lack of difference was against the hypothesis, which had expected donations to be higher when the ECA accepted inputs. The 'backup' measures of the follow-up questionnaire do, in contrast, show some differences between the two conditions. Table 8-1 summarizes the statements for which the levels of agreement were significant or near significant.

	ECA ignoring		ECA accepting		
Statement	**Mean**	**Std.Dev.**	**Mean**	**Std.Dev.**	**Sig.**
I enjoyed the conversation	-0.21	1.179	0.52	0.79	**0.017**
I felt the character was well informed	1.30	0.47	1.00	0.43	**0.026**
The character could have been more persuasive	0.39	0.99	-0.17	0.72	**0.032**
The character was interesting	0.04	1.197	0.7	0.88	**0.039**
I learned something from the conversation	0.46	1.285	1.09	0.73	**0.046**
I felt in touch with the character	-0.71	0.96	-0.17	0.94	0.059

I liked the character	0.29	1.16	0.87	0.92	0.066
The character liked me	-0.63	0.824	-0.26	0.619	0.095

Table 8-1 Persuasive ECA statement agreement summary

When the ECA was accepting input subjects enjoyed the conversations more (mildly agree versus slightly disagree), felt the ECA was less well informed (agree versus slightly strongly agree), were less likely to say the character could have been more persuasive (slightly disagree versus mildly agree), found the character more interesting (mostly agree versus neither agree nor disagree) and felt they learned more (mostly don't agree or disagree versus mostly disagree). These results are summarised in the histogram pairs in Figure 8-4 to Figure 8-8 below.

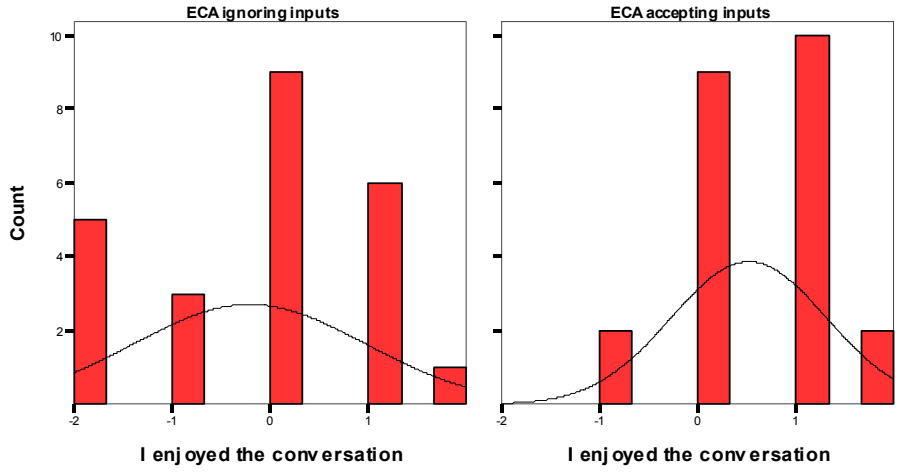


Figure 8-4 Agreement distribution – "I enjoyed the conversation"

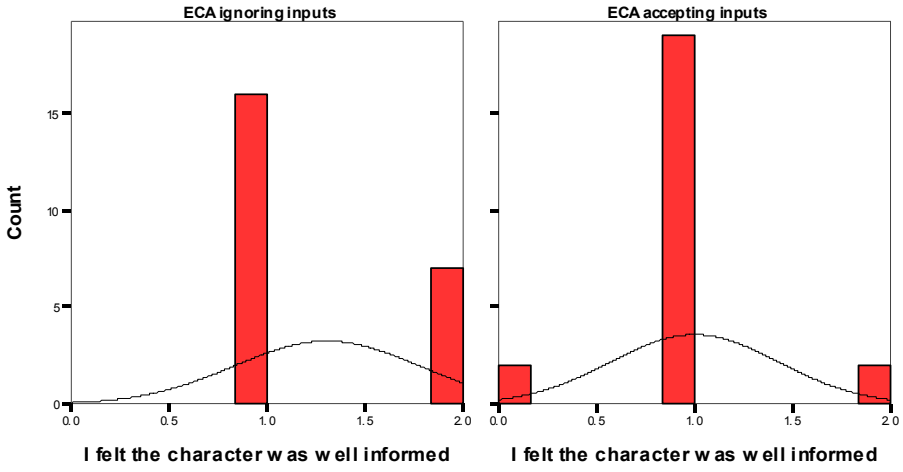


Figure 8-5 Agreement distribution – "I felt the character was well informed"

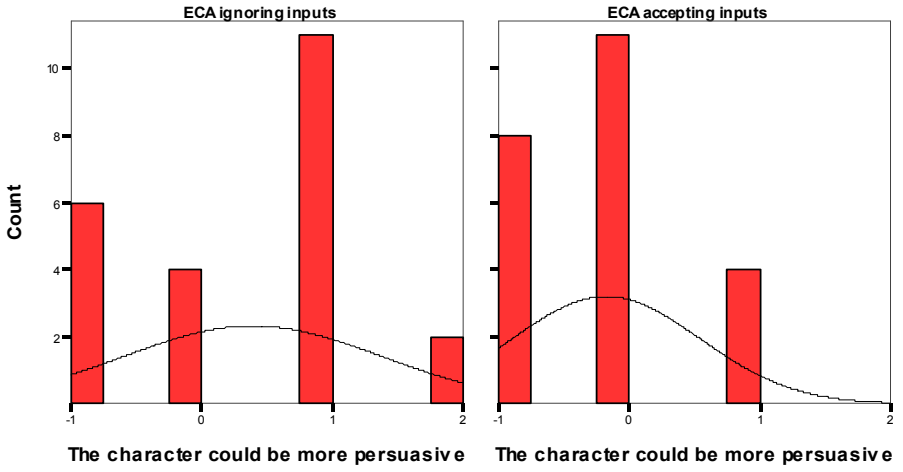


Figure 8-6 Agreement distribution – "The character could be more persuasive"

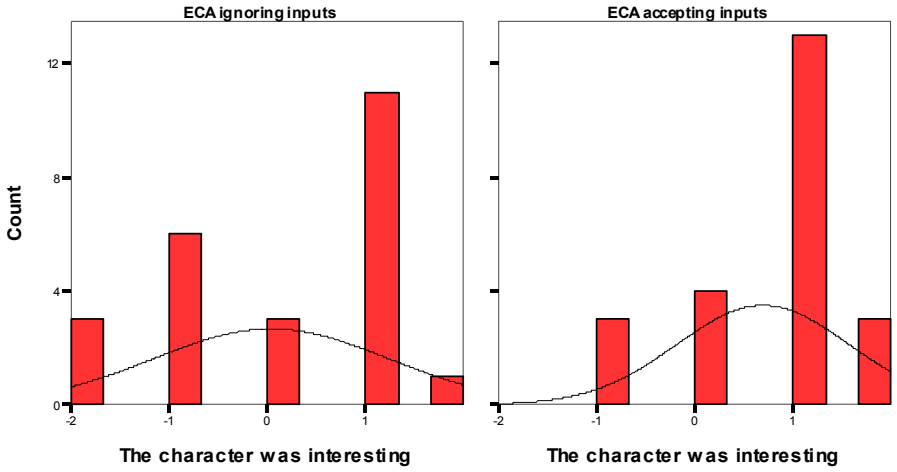


Figure 8-7 Agreement distribution – "The character was interesting"

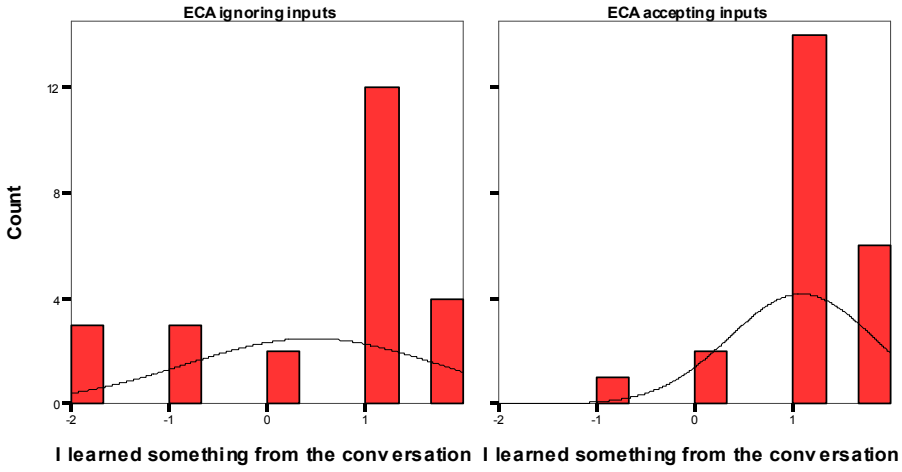


Figure 8-8 Agreement distribution – "I learned something from the conversation"

8.7 Conclusion

The direct measure of behaviour change used did not strongly indicate that an ECA reacting to a subject's non-verbal behaviour was more strongly persuasive than one which did not. However, the questionnaire results do suggest that the two conditions affected subjects differently, and so it can be concluded that with more development a non-verbally interactive ECA would likely increase levels of persuasion. This is in line with the results of the previous study with synthetic ECAs, but the questionnaire results from this study indicate that an increase in persuasiveness due to interactive non-verbal behaviour could actually occur for a real ECA in practice.

The interactive non-verbal behaviour of the ECA developed was notably rudimentary. More refined and more complex and additional behaviours and reactions could be added to the ECA along with a more sophisticated sense of conversational state. This study and the previous study of synthetic ECA combined suggest that this more advanced ECA would likely increase the persuasiveness of the ECA, towards achieving a measurable effect on actual behaviour.

8.8 Limitations of this work

The results of this study are limited to interactions within a relatively simple environment (a webcam interface) and may not generalize to more realistic or complex environments. The study does not address ECAs that may attempt to be more pro-actively persuasive, for instance by using more persuasive language or other persuasion methods.

The visual sharpness of the character could be increased, but it is not felt that this would significantly affect the outcome of this study.

It is difficult to define a ground or control group for studies with ECAs. It would have been possible, again, to use a real human (either directly or as a synthetic ECA) as a

control, or alternatively, a paper-based, audio-based, or video-based control could be used.

8.9 Observations and further work

The quantization of monies given to subjects is not believed to have affected the amounts donated, although the exact denominations may have had an effect on the actual amounts donated and on the large variations in the amounts donated. Subjects were given £20 cash in the form of £16 in £2 coins, and £4 in £1 coins. There was a non-normal distribution of donations, and donations focused on specific amounts – £0, £1, £2, £3, £5, £10, £20 – the latter three sums suggesting subjects rounded amounts to 'round' numbers. This was true also in the previous study and the non-normal distribution makes statistical analysis more complex, though differences between conditions were still found in both studies. An alternative method of directly measuring persuasion might avoid this situation – i.e. a technique that does not require people to choose a discrete amount, as people seem biased towards 'round' numbers. There was also a clear ceiling effect, with many subjects donating the full £20, as well as a ground effect with a significant proportion of subjects giving the minimum £0.

9. Conclusions and Discussion

This concluding chapter of the thesis overviews the work presented on non-verbal behaviour in humans, and the social influence of ECAs, along with the various empirical studies run to elucidate and demonstrate aspects of non-verbal behaviour and its value to ECAs. Overall conclusions are given along with recommendations for future studies and development.

This thesis focused on the extent to which non-verbal behaviour in ECAs can affect the actions or behaviour of real people, which aspects of non-verbal behaviour may be important in creating a persuasive effect, and how these aspects could be used to aid the development of ECAs. Throughout the thesis attention was given to how ECAs can be evaluated in objective empirical studies, for social influence effects or otherwise. Based on the fact that that non-verbal behaviour is natural and highly important in interactions between people, and that as people treat ECAs like real people it was expected that non-verbal behaviour would also be important for interactions between humans and ECAs. Specifically, non-verbal behaviour on the part of the ECA that responds to the non-verbal behaviour of the human interactant would be important. The concept of synthetic ECAs was introduced as a paradigm in order to investigate and evaluate potential social influence of ECAs – how much social influence ECAs may eventually have.

Under this paradigm a synthetic ECA was designed and implemented. It was demonstrated that people reacted to this synthetic ECA as if it was a real ECA, even though the synthetic ECA's behaviour, both verbal and non-verbal, was far advanced on the present state of the art. The validated synthetic ECA was then used to empirically evaluate the 'persuasive potential' of ECAs using a direct measure of behaviour change. The synthetic ECA appeared no less persuasive than a real human in the same scenario, so it was suggested that ECAs have the potential to have as much social influence as real people. It was also found that when a synthetic ECA could not see the subject it was interacting with the level of persuasion was significantly lower that when it could. This suggested that it was important that an ECA should react to the non-verbal behaviour of its interactant. It was clear from the non-verbal behaviour literature that it is important

that these reactions occur in a timely manner just as they do in real human-human interactions.

This result along with a perspective on the historical development of robot control systems motivated the suggestion of using behaviour-based hybrid architecture for ECAs, enabling both fast interactive low-level behaviours along with slower high-level 'cognitive' behaviours. It was proposed that implementing this hybrid architecture using a modern streaming architecture approach would be appropriate, and a prototype ECA was developed with this in mind, to determine whether this approach was effective from the perspectives of both effective non-verbal behaviour (in this case, affecting persuasion) and effective software development. The prototype ECA was evaluated using the same methodology and direct measure of behaviour change as in the synthetic ECA studies.

Development of the ECA using a behaviour-based architecture using a streaming approach was straightforward. Each module could be designed, implemented, tested, and debugged independently. This suggests that using behaviour-based architectures with a streaming approach would scale well to the development of more sophisticated ECAs. The networked aspect of the design also means that the approach can easily scale well with more and more computationally expensive modules.

Evaluation of the prototype ECA showed little difference between the two conditions of the ECA reacting to and ignoring the subjects' behaviour using the direct measure of behaviour change (how much money was donated to the charity). However, questionnaire results showed a significant favour towards the reacting ECA. It was suggested that with additional behavioural modules, more sophisticated conversation state, and further refinement of the present modules this favour would increase sufficiently to cause an effect that could be measured directly. The evaluation also showed that the ECA worked in a technical sense – people engaged with the character under both conditions and consistently reported that enjoyed the conversation, learned things from the conversation, etc.

Overall, ECAs will be capable of persuading people and exerting social influence and that for these purposes, and presumably more widely, it is important for the non-verbal behaviour of ECAs to respond interactively to the non-verbal behaviour of their human interactants. Also using a streaming architecture/hybrid architecture approach for the development of ECAs would be an effective way forward to enable this non-verbal interactivity.

One of these days your fridge will try to persuade you into having a glass of orange juice instead of another beer!

9.1 Further discussion

As discussed in Chapter 2, non-verbal behaviour is extremely complex and is only just becoming understood in an empirical way. Most knowledge and literature in the non-verbal behaviour area is descriptive, lacking generative or computational models. There are many theories about where various aspects of non-verbal behaviour come from, what they are depend on, and what various non-verbal behaviours mean, but these theories are difficult to test in practice. Neuro-imaging technologies are becoming a powerful tool in various areas of psychology and neuroscience and show strong promise of assisting in developing stronger theories and generative models of non-verbal behaviour.

The model of non-verbal behaviour which an ECA has internally is not required to be realistic or to be based upon how human brains work. The requirement is only that the non-verbal behaviour that an ECA produces is effective, realistic or convincing. ECAs have only recently started using non-verbal behaviour, and the evaluations of these ECAs have been limited. Stronger evaluations and innovative evaluation methodologies

will help to establish that the non-verbal behaviour that these ECA produce is effective and that their models are appropriate to producing effective non-verbal behaviour in the subject. Of course, if ECAs are developed using models similar to theoretical ones about real humans, then the evaluation of these ECAs does, to some extent, validate the underlying theoretical models of real people. In that way, the development of ECAs and the consequent evaluation thereof may provide new information and knowledge for the psychological and other communities which provided much of the original knowledge for the development of the human aspects of ECAs.

Development of these ECAs with sophisticated non-verbal behaviour will enable further experiments in psychology and psycholinguistics (as well as other ares) that are not possible or not easily possible without such technology. In addition to being driven from the gesture generation system, ECAs could also just play back data captured from real human subjects. Furthermore, these ECAs could play back that data in an altered form. The movements could, for example, be amplified, making the gestures bigger and facial expressions more obvious.

Real humans cannot produce gestures in a controlled manner, and find it exceedingly difficult to produce 'incorrect' gestures. A simple example for Westerners is to try shaking your head while saying yes, or nodding your head while saying no. With thought and practice this is possible but the cultural training is very difficult to overcome. In some cultures, the meanings of head nodding and shaking are reversed from Western assumptions but the principle remains the same. A whole variety of experiments not possible with real people could be performed to determine what aspects of gestures are important to understanding and to the underlying psychology

A more complex example of behaviours that real people find difficult or impossible to perform incorrectly is that of beat emphasis. For example, when a person is describing a dog, a very big dog, and wants to emphasise the bigness, a beat gesture is made by the hand 'stroking' down on the word, and is performed on the word *big* in the phrase 'it was a *big* dog'. The duration of the beat exactly matches the duration of the word *big*

and is synchronised with it. It is hard, almost impossible, for real people to say that same sentence with the word *big* extended without also extending the duration of the associated beat gesture. This can be mastered with practice, but it is very unnatural. To make an ECA system perform gestures in this unnatural way would be easy, as it would also be to alter the playback of real human data to this unnatural form. Systematic investigation into these alterations could help in the discovery of what is important in gesture, and what things trigger people to think that something is wrong with the interaction. These can be extended to other areas of non-verbal behaviour beyond gesture and further still.

The idea of using persuasion as an evaluation metric for ECAs and more specifically as an objective and empirical measure for evaluation was introduced in Chapter 4. Persuasion is only one of many possible ways to evaluate the social influence of an ECA and social influence is only one of many aspects which are worthy of evaluation. It is not suggested that persuasion is the best or the only metric that could be devised, merely that is an example of an objective; empirical measure and that it could be used for other ECAs.

Evaluation of ECAs is difficult and highly context-sensitive because ECAs try to replicate at least some aspects of human behaviour and evaluation of humans is difficult and highly context-sensitive. There are many different ways of evaluating humans, for many different purposes, and a single evaluation strategy would be highly inappropriate. While evaluation is a difficult problem, within certain contexts evaluation strategies can be developed to aid in the development of effective ECAs and evaluation should not be shied away from. The value an ECA adds to an institution (business, website, game, educational establishment, etc.) is the ultimate important factor, but this cannot usually be measured directly, so some evaluation strategy aligned with the aims/needs of the institution is required. Furthermore, evaluation strategies are important in aiding the development of ECAs to provide indicators that an ECA of a sufficient quality and also to provide indicators on ways in which an ECA could be improved.

The development of ECAs is a complex task and as discussed in Chapter 3 most groups working with ECAs tend to develop their own 'in-house' ECA. This is an inordinate repetition of hard work and means that ECAs are developing relatively slowly. Using a streaming type architecture could help encourage and enable researchers to share their developments and therefore to focus on increasing the behavioural variability of various ECAs rather than being hampered by repetitive development. It should be noted that streaming architectures are only one way of encouraging this sharing of development resources. The general modularisation of ECAs and integration with various open source software packages also provides these same advantages, and in fact, both could occur together. Graphics engines such as Crystal Space (Crystal Space Team, 2008), Delta3D (Delta3D, 2008), Irrlicht (Irrlicht, 2008), Ogre3D (Ogre3D, 2008), and Panda3D (Carnegie Mellon University, 2008b) provide strong character animation facilities and perform rendering themselves, but importantly from the ECA perspective they do not support real-time lip-sync. If lip-sync, such as that based on the lip-sync component developed for the prototype ECA of this research, were added to any of these engines, ECA developers could focus more strongly on the behavioural capabilities of their ECAs.

The significant developments over recent years of various XML mark up languages for ECAs suggests that researchers are trying to build bridges so that the deliberative parts of ECAs can be shared, and co-developed more effectively. Parallel development of openly available frameworks and content (character models, animations, etc.) for the character animation side would support this collaborative effort well.

The observations previously on the quantisation of donations could be addressed in a variety of ways – a separate donation measure could be used by, for example, having the character inform subjects the longer they crank a handle the more money will be donated to charity, although it would be important to make the handle action quite tough so they would stop eventually. Alternatively, subjects could be invited to play a game with the character, where continuing to play the game continues to donate money – such as at the FreeRice website (FreeRice, 2008) – see Figure 9-1.

Figure 9-1 FreeRice website (FreeRice, 2008)

This latter form of interaction could be an installed longitudinal study inviting passers-by to play. The form of interaction would also be significantly more interactive and with strong scope for both verbal and non-verbal behaviour of an ECA, especially in response to an interactant. The behaviour (or presence) of an ECA could be controlled and varied through software, and with little support needed from experimenters considerable longitudinal data could be collected. This scenario could also be easily replicated by other institutions, so 'between-character' comparisons could be made. It would also provide a strong and simple control case – the simple site with a touch screen. The role of the ECA within this type of scenario would also be better defined, and what reactions and behaviours an ECA should have would therefore also be easier to define. Furthermore, the conversational state between the ECA and a subject would be more complex and the variation in behaviour between states would be more varied. Comparisons of various different attributes of an ECA (2D versus 3D, male versus female, gender matched to subject or not, age matched to subject or not, clothing style, etc.) could be made, as well as comparisons between ECAs and other forms of persuasion. For example, real video samples could be used instead of an ECA as the domain is restricted enough that sufficient video could be generated.

Finally, integrating an ECA into a website, such as the FreeRice one at Figure 9-1, would provide a good test bed for evaluating various ECAs and various persuasive strategies, with large numbers of subjects and at almost no cost, while also providing exposure and possibly positive regard to an institution that presented the website. This approach of using a charitably donating website (or similar) and using ECAs to attempt to encourage subjects to donate more money is suggested as an appropriate methodology for further investigations into the persuasiveness of ECAs based on the experiences described within this thesis of investigating this persuasiveness.

References

Abbattista, F., Lops, P., Semeraro, G., Andersen, V., & Andersen, H. H. K. (2002). *Evaluating Virtual Agents for E-Commerce*. Paper presented at the First International Joint Conference on AAMAS. Bologna, Italy.

Advanced Realtime Tracking GmbH. (2008). Arttrack2.

Allen, K. (2003). Are Pets a Healthy Pleasure? The Influence of Pets on Blood Pressure. *Current Directions in Psychological Science, 12*(6), 236–223.

Ambady, N., LaPlante, D., Nguyen, T., Rosenthal, R., Chaumeton, N., & Levinson, W. (2002). Surgeons' Tone of Voice: A Clue to Malpractice History. *Surgery, 132*(1), 5-9.

Anderson, P. A. (1985). Nonverbal Immediacy in Interpersonal Communication. In Siegman, A. W. & Feldstein, S. (Eds.), *Multichannel Integrations of Nonverbal Behavior*. Lawrence Erlbaum Associates.

André, E., Rist, T., & Müller, J. (1998). Webpersona: A Life-Like Presentation Agent for the World-Wide Web. *Knowledge-Based Systems, 11*(1), 25-36.

Arkin, R. C. (1998). *Behavior-Based Robotics*. Mit Press. Cambridge Mass.

Autodesk. (2008). 3ds Max,

Aylett, M. P., & Pidcock, C. J. (2007). The Cerevoice Characterful Speech Synthesiser Sdk. In *Intelligent Virtual Agents* (Vol. 4722, pp. 413-414). Springer. Berlin / Heidelberg.

Badler, N. (1997). *Real-Time Virtual Humans.* Paper presented at the 5th Pacific Graphics conference, Seoul, Korea.

Bailenson, J. N., & Yee, N. (2005). Digital Chameleons. *Psychological Science, 16*(10), 814-819.

Ball, J. E., & Breese, J. (1998). *Emotion and Personality in a Conversational Character.* Paper presented at the WECC '98: The First Workshop on Embodied Conversational Characters.

Bauer, P. (1991). Multiple Testing in Clinical Trials. *Statistics in Medicine, 10*, 871-890.

Bavelas, J. (1996). Debunking Body Language [video]. University of Victoria.

Baylor, A. L. (2006). *Interface Agents as Social Models: The Impact of Appearance on Females' Attitude toward Engineering.* Paper presented at the CHI 2006 conference, Montreal, Canada.

Beattie, G. (2003). *Visible Thought : The New Psychology of Body Language*. Routledge. London ; New York.

Bell Labs. (1997). Background: Bell Labs Text-to-Speech Synthesis: Then and Now. http://www.bell-labs.com/news/1997/march/5/2.html (Retrieved 09/01/2008)

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press. Princeton.

Bente, G., Krämer, N. C., Petersen, A., & de Ruiter, J. P. (2001). Computer Animated Movement and Person Perception. (Methodological Advances in Nonverbal Behavior Research). *Journal of Nonverbal Behavior, 25*(3), 151-166.

Bente, G., Krämer, N. C., Trogemann, G., Piesk, J., & Fischer, O. (2001). Conversing with Electronic Devices: An Integrated Approach Towards the Generation and Evaluation of Nonverbal Behavior in Face-to-Face Like Interface Agents. In Heuer, A. & Kirste, T. (Eds.), *Intelligent Interactive Assistance and Mobile Multi-Media Computing. Proceedings of the Imc2000* (pp. 67-76). Neuer Hochschulschriftenverlag. Rostock.

Bernsen, N. O., & Dybkjær, L. (2004). *Evaluation of Spoken Multimodal Conversation.* Paper presented at the 6th international conference on Multimodal interfaces conference, State College, PA.

Bers, J. (1996). A Body Model Server for Human Motion Capture and Representation. *Presence: Teleoperators and virtual environments, 5*(3), 381-392.

Bickmore, T. (2003). *Relational Agents: Effecting Change through Human-Computer Relationships.* MIT. Cambridge, MA.

Bickmore, T., Caruso, L., Clough-Gorr, K., & Heeren, T. (2005). 'It's Just Like You Talk to a Friend' – Relational Agents for Older Adults. *Interacting with Computers, 17*(6), 711-735.

Birdwhistell, R. L. (1971). *Kinesics and Context : Essays on Body-Motion Communication*. Penguin Press. London.

Blascovich, J. (2002). *A Theoretical Model of Social Influence for Increasing the Utility of Collaborative Virtual Environments.* Paper presented at the Collaborative virtual environments conference, Bonn, Germany.

Bonasso, R. P. (1991). *Integrating Reaction Plans and Layered Competences through Synchronous Control.* Paper presented at the International Joint Conference on Artificial Intelligence (IJCAI).

Brooks, R. A. (1986). A Robust Layered Control System for a Mobile Robot. *Robotics and Automation, IEEE Journal of, 2*(1), 14-23.

Brooks, R. A. (1987). *Planning Is Just a Way of Avoiding Figuring out What to Do Next*. Massachusetts Institute of Techonology.

Brooks, R. A. (1991). Intelligence without Representation. *Artificial Intelligence, 47*, 139-159.

Bryson, J. J. (2003). The Behavior-Oriented Design of Modular Agent Intelligence. In Müller, J. P. (Ed.), *Agent Technologies, Infrastructures, Tools, and Applications for E-Services: Node 2002 Agent-Related Workshops, Erfurt, Germany, October 7-10, 2002. Revised Papers* (Vol. 2592, pp. 61-76). Springer Berlin / Heidelberg. Erfurt, Germany.

Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., et al. (2004). *Toward Real-Time Multimodal Processing: Eyesweb 4.0*. Paper presented at the AISB 2004 Convention: Motion, Emotion and Cognition. Leeds, UK.

Cannon, W. B. (1929). *Bodily Changes in Pain, Hunger, Fear and Rage; an Account of Recent Researches into the Function of Emotional Excitement* (2nd ed.). Appleton. New York, London,.

Carnegie Mellon University. (2008a). Cmu Sphinx, http://cmusphinx.sourceforge.net/html/cmusphinx.php

Carnegie Mellon University. (2008b). Panda3d, http://panda3d.org/

Cassell, J. (2000). *Embodied Conversational Agents*. MIT Press. Cambridge, Mass.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., et al. (1999). *Embodiment in Conversational Interfaces: Rea.* Paper presented at the CHI'99.

Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2001). *Beat: The Behavior Expression Animation Toolkit.* Paper presented at the 28th annual conference on Computer graphics and interactive techniques conference, Los Angeles, CA.

Cellerier, A. (2005). Vlc Media Player, http://www.videolan.org/

Cerney, M. M. (2005). *From Gesture Recognition to Functional Motion Analysis: Quantitative Techniques for the Application and Evaluation of Human Motion.* Iowa State University. Ames.

Chartrand, T. L., & Bargh, J. A. (1999). The Chameleon Effect: The Perception-Behavior Link and Social Interaction. *Journal of personality and social psychology, 77*(6), 893-910.

Chau, M., & Betke, M. (2005). *Real Time Eye Tracking and Blink Detection with Usb Cameras* (No. 2005-12). Boston University. Boston, MA

Chen, D., & Haviland-Jones, J. (2000). Human Olfactory Communication of Emotion. *Perceptual and motor skills, 91*(3 part 1), 771-781.

Chomsky, N. (1957). *Syntactic Structures*. Mouton Publishers. The Hague.

Churchill, W. S. (1946). Iron Curtain Speech. Westminster College, Fulton, MO.

Connell, J. (1991). *Sss: A Hybrid Architecture Applied to Robot Navigation.* Paper presented at the IEEE conference on Robotics and Automation (ICRA).

Crystal Space Team. (2008). Crystal Space, http://www.crystalspace3d.org

Cycling74. Max/Msp, http://www.cycling74.com

DeCarlo, D., Stone, M., Revilla, C., & Jennifer J. Venditti. (2004). Specifying and Animating Facial Signals for Discourse in Embodied Conversational Agents. *Computer Animation and Virtual Worlds, 15*(1), 27-38.

Dehn, D. M., & Mulken, S. v. (2000). The Impact of Animated Interface Agents: A Review of Empirical Research. *International Journal of Human-Computer Studies, 52*(1), 1-22.

Delta3D. (2008). Delta3d, http://www.delta3d.org/

DePaulo, B. M., & Friedman, H. S. (1998). Nonverbal Communication. In Gilbert, D. T., Fiske, S. T. & Lindzey, G. (Eds.), *The Handbook of Social Psychology* (4th ed., pp. 2 v.). McGraw-Hill. Boston, New York.

Derlega, V. (1995). Health, Health Care, and Nonverbal Behavior: An Issue Overview. *Journal of Nonverbal Behavior, 19*(4), 189-190.

Duncan, S., & Fiske, D. W. (1977). *Face-to-Face Interaction : Research, Methods, and Theory*. Erlbaum. Hillsdale, NJ.

Efran, J. S. (1968). Looking for Approval: Effects on Visual Behaviour of Approbation from Person Differing in Importance. *Journal of Personality and Social Psychology, 10*, 21-25.

Efron, D. (1941). *Gesture and Environment*. King's Crown Press. New York.

Ehrlichman, H., & Weinberger, A. (1978). Lateral Eye Movements and Hemispheric Asymmetry: A Critical Review. *Psychological Bulletin, 85*, 1080-1101.

Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica, 1*, 49-98.

Ekman, P., & Friesen, W. V. (1969). A Tool for the Analysis of Motion Picture Film or Video Tape. *American Psychologist, 24*, 240-243.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the Face*. Prentice-Hall. Englewood Cliffs, NJ.

Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971). Facial Affect Scoring Technique: A First Validity Study. *Semiotica, 3*, 37-58.

Essa, I. A. (1995). *Analysis, Interpretation and Synthesis of Facial Expressions.* MIT. Cambridge, MA.

Fikes, R. E., & Nilsson, N. J. (1971). Strips: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence, 2*, 189-208.

Fischer, J., & Bartz, D. (2005). *Real-Time Cartoon-Like Stylization of Ar Video Streams on the Gpu* (No. WSI-2005-18). Wilhelm Schickard Institute for Computer Science, University of Tübingen, Germany.

Fischer, M., & Zwaan, R. A. (2008). *Grounding Cognition in Perception and Action : A Special Issue of the Quarterly Journal of Experimental Psychology*. Psychology. Hove.

Fleming, V., Thorpe, R., & Vidor, K. (Writer) (1939). The Wizard of Oz. In LeRoy, M. (Producer). USA: Warner Bros. Inc.

Fogg, B. J. (1998). *Persuasive Computers: Perspectives and Research Directions.* Paper presented at the CHI 98 conference, Los Angeles, CA.

Freedesktop.org. (2007). Gstreamer. http://gstreamer.freedesktop.org/)

Freedman, N., & Hoffman, S. P. (1967). Kinetic Behavior in Altered Clinical States: Approach to Objective Analysis of Motor Behavior During Clinical Interviews. *Perceptual and Motor Skills, 24*(2), 527-539.

FreeRice. (2008). Freerice. http://freerice.com/ (Retrieved 2008-07/27)

Freleng, F. (Writer) (1950). Canary Row. In Selzer, E. (Producer), *Merrie Melodies*. USA: Warner Brothers.

Frick, R. W. (1985). Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin, 97*, 412–429.

Gat, E. (1991). *Reliable Goal-Directed Reactive Control for Real-World Autonomous Mobile Robots.* Virginia Polytechnic Institute and State University. Blacksburg, VA.

Gat, E. (1998). On Three-Layer Architectures In Kortenkamp, D., Bonnasso, R. P. & Murphy, R. (Eds.), *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems* (pp. 195-210). MIT Press. Cambridge, MA.

Gibbs, R. W. (1999). *Intentions in the Experience of Meaning*. Cambridge University Press. Cambridge ; New York.

Gladwell, M. (2005). *Blink : The Power of Thinking without Thinking* (1st international mass market paperback ed.). Little Brown and Co. New York.

Grammer, K., Filova, V., & Fieder, M. (1997). The Communication Paradox and a Possible Solution: Toward a Radical Empiricism. In A. Schmitt, K. A., K. Grammer & K. & Schafer (Eds.), *New Aspects of Human Ethology* (pp. 91-120). Plenum. New York.

Greatbatch, D., & Clark, T. (2005). *Management Speak : Why We Listen to What Management Gurus Tell Us*. Routledge. London.

Hall, E. T. (1966). *The Hidden Dimension* ([1st]. ed.). Doubleday. Garden City, NY.

Hall, E. T. (1973). *The Silent Language*. Anchor Press/Doubleday. Garden City, N.Y.,.

Hall, E. T. (1974). Proxemics. In Weitz, S. (Ed.), *Nonverbal Communication* (pp. 205-229). Oxford University Press. New York.

Harper, R. G., Wiens, A. N., & Matarazzo, J. D. (1978). *Nonverbal Communication : The State of the Art*. Wiley. New York.

Hayes-Roth, B., van Gent, R., & Huber, D. (1996). Acting in Character. In Trappl, R. & Petta, P. (Eds.), *Creating Personalities for Synthetic Actors* (pp. 92-112). Springer-Verlag. Berlin.

Hearn, G. (1957). Leadership and the Spatial Factor in Small Groups. *Journal of Abnormal and Social Psychology, 54*, 269-272.

Heslin, R. (1974). *Steps toward a Taxonomy of Touching.* Paper presented at the Midwestern Psychological Association. Chicago.

Hess, E. H., & Goodwin, E. (1973). The Present State of Pupilometrics. In Jannise, M. P. (Ed.), *Pupillary Dynamics and Behavior*. Plenum Press. New York.

Hinde, R. A. (1972). *Non-Verbal Communication*. Cambridge University Press. Cambridge.

Hungerford, M. W. (1878). *Molly Bawn*.

Intel Corporation. (2005). Opencv, http://www.intel.com/technology/computing/opencv/index.htm

Irrlicht. (2008). Irrlicht, http://irrlicht.sourceforge.net

Isbister, K., & Doyle, P. (2004). The Blind Men and the Elephant Revisted: Evaluating Interdisciplinary Eca Research. In Ruttkay, Z. & Pelachaud, C. (Eds.), *From Brows to Trust* (Vol. 7). Kluwer Academic. Dordrecht.

Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. (2000). *Helper Agent: Designing an Assistant for Human-Human Interaction in a Virtual Meeting Space.* Paper presented at the Human Factors in Computing Systems (CHI 2000) conference, Hague, Netherlands.

James, W. (1890). *The Principles of Psychology*. H. Holt and company. New York,.

Jones, S. E., & Yarbrough, A. E. (1985). A Naturalistic Study of the Meanings of Touch. *Communication Monographs, 52*(1), 19-56.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall. Upper Saddle River, NJ.

Katafiasz, M. (2006). Multipurpose Multimedia Processing with Gstreamer. http://www-128.ibm.com/developerworks/aix/library/au-gstreamer.html?ca=dgr-lnxw07GStreamer (Retrieved 12/12/2007)

Katagiri, S. (2000). *Handbook of Neural Networks for Speech Processing*. Artech House. Boston.

Katie, K. (1997). The Body Language of Proxemics. http://members.aol.com/katydidit/bodylang.htm (Retrieved 20/05/2005)

Keeling, K., Beatty, S., McGoldrick, P., & Macaulay, L. (2004). *Face Value? Customer Views of Appropriate Formats for Embodied Conversational Agents (Ecas) in Online Retailing.* Paper presented at the 37th Hawaii International Conference on System Sciences conference, Hilton Waikoloa Village, Island of Hawaii (Big Island).

Kendon, A. (1967). Some Functions of Gaze-Direction in Social Interaction. *Acta Psychologica, 26*, 22-63.

Kendon, A. (1972). Some Relationships between Body Motion and Speech. An Analysis of an Example. In Siegman, A. P., B (Ed.), *Studies in Dyadic Communication* (pp. 177-210). Pergamon Press. Elmsford, NY.

Key, M. R. (1980). *Relationship of Verbal and Non-Verbal Communication*. Mouton.

Khronos Group. (2008). Opengl, http://www.opengl.org/

Kimball, S., & Mattis, P. (2006). Gnu Image Manipulation Program (Version 2.210), http://www.gimp.org/

http://cvs.gnome.org/viewcvs/gimp/plug-ins/common/cartoon.c?view=markup

Kleinke, C., & Pohlen, P. (1971). Affective and Emotional Responses as a Function of Other Person's Gaze and Cooperativeness in a Two-Person Game. *Journal of Personality and Social Psychology, 17*, 308-313.

Knapp, M. L., & Daly, J. A. (2002). *Handbook of Interpersonal Communication* (3rd ed.). SAGE Publications. Thousand Oaks, CA.

Kopp, S., Tepper, P., & Cassell, J. (2004). *Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output.* Paper presented at the Proceedings of the 6th international conference on Multimodal interfaces conference, State College, PA.

Kranstedt, A., Kopp, S., & Wachsmuth, I. (2002, 2002). *Murml: A Multimodal Utterance Representation Markup Language for Conversational Agents.* Paper presented at the Proceedings of ECA's: Let's specify and evaluate them! Workshop held in conjunction with AAMAS 2002 conference, Bologna.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior, 27*, 145-162.

Laurent, D., & Dachary, L. (2008). Cal3d, http://home.gna.org/cal3d/

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. Cambridge; New York.

Lexicle.com. (2005). Alex, http://www.lexicle.com/

Li, D., & Parkhurst, D. J. (2006). *Open-Source Software for Real-Time Visible-Spectrum Eye Tracking.* Paper presented at the COGAIN conference.

Li, D., Winfield, D., & Parkhurst, D. J. (2005). *Starburst: A Hybrid Algorithm for Video-Based Eye Tracking Combining Feature-Based and Model-Based Approaches.* Paper presented at the IEEE Vision for Human-Computer Interaction Workshop at CVPR.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology, 140*, 1-55.

Liu, Y., Guo, D., Sun, J., & Wei, Z. (2003). *A Real-Time Control Model for Intelligent Virtual Human*.

Loehr, D. P. (2004). *Gesture and Intonation.* Unpublished Doctoral Thesis, Georgetown University. Washington DC.

Long, B. (2002). Speech Synthesis & Speech Recognition Using Sapi 5.1. http://www.blong.com/Conferences/DCon2002/Speech/SAPI51/SAPI51.htm#Animation (Retrieved 04/09/2008)

MacDorman, K. F., Minato, T., Shimada, M., Itakura, S., Cowley, S. J., & Ishiguro, H. (2005). *Assessing Human Likeness by Eye Contact in an Android Testbed.* Paper presented at the the XXVII Annual Meeting of the Cognitive Science Society. conference, Stresa.

Maes, P., & Brooks, R. A. (1990). *Learning to Coordinate Behaviors.* Paper presented at the National Conference on Artificial Intelligence (AAAI).

Massaro, D. W., & Stork, D. G. (1998). Speech Recognition and Sensory Integration. *American Scientist, 86*(3).

McCafferty, S. G. (1998). Nonverbal Expression and L2 Private Speech. *Applied Linguistics, 19*(1), 73-96.

McCarthy, J., Minsky, L., Rochester, N., & Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artifiicial Intelligence*.

McNeill, D. (1992). *Hand and Mind : What Gestures Reveal About Thought*. University of Chicago Press. Chicago, IL.

McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press. Chicago, IL.

McNeill Lab. (2003). Communicative Gesture. http://web.archive.org/web/20041028223219/http://mcneilllab.uchicago.edu/topics/comm.html (Retrieved 25/5/5, 2005)

McNeill Lab. (2006). Center for Gesture and Speech Research. http://mcneilllab.uchicago.edu/)

Mehrabian, A. (1971). *Silent Messages*. Wadsworth. Belmont, CA.

Meso. Vvvv: http://vvvv.meso.net/ http://vvvv.meso.net/

Microsoft Corporation. (2007). Microsoft Directshow 9.0. http://msdn2.microsoft.com/en-gb/library/ms783323.aspx (Retrieved 05/03/2007, 2007)

Microsoft Corporation. (2008). Microsoft Speech Api: Microsoft Corporation, http://www.microsoft.com/speech/speech2007/default.mspx

Minato, T., Shimada, M., Itakura, S., Kang, L., & Ishiguro, H. (2005). *Does Gaze Reveal the Human Likeness of an Android?* Paper presented at the 4th International Conference on Development and Learning.

Minsky, M. L. (1986). *The Society of Mind*. Simon and Schuster. New York.

Minsky, M. L. (2006). *The Emotion Machine : Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster. New York.

Moravec, H. P. (1989). Sensor Fusion in Certainty Grids for Mobile Robots. In *Sensor Devices and Systems for Robotics* (pp. 253-276). Springer-Verlag New York, Inc.

Morency, L.-P. (2006). Watson: Head Tracking and Gesture Recognition Library (Version 2.1a). Cambridge, MA: MIT,

Morgan, J. F. (2007). P Value Fetishism and Use of the Bonferroni Adjustment. *Evidence-based mental health, 10*(2), 34-35.

Mori, M. (1970). Bukimi No Tani [the Uncanny Valley]. *Energy, 7*, 33-35.

Mullennix, I. W., Stern, S. E., Wilson, S. J., & Dyson, C.-l. (2003). Social Perception of Male and Female Computer Synthesized Speech. *Computers in Human Behavior, 19*(4), 407-424.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science (in press), 15*(2), 133-137.

Nespoulous, J.-L., Perron, P., & Lecours, A. R. (1986). *The Biological Foundations of Gestures : Motor and Semiotic Aspects*. Erlbaum. Hillsdale, NJ.

Nichols, K., & Champness, B. (1971). Eye Gaze and the Gsr. *Journal of Experimental Social Psychology, 7*(6), 623-626.

Noot, H., & Ruttkay, Z. (2003). *The Gestyle Language.* Paper presented at the International Workshop on Gesture and Sign Language based Human-Computer Interaction conference, Genova.

Ogre3D. (2008). Ogre3d, http://www.ogre3d.org/

Organic Motion. (2007). *Organic Motion Unveils First Major Breakthrough in Motion Capture Technology in More Than 20 Years.*

Perneger, T. V. (1998). What's Wrong with Bonferroni Adjustments. *British Medical Journal, 316*(7139), 1236-1238.

Previc, F. H., Declerck, C., & de Brabander, B. (2005). Why Your "Head Is in the Clouds" During Thinking: The Relationship between Cognition and Upper Space. *Acta Psychologica, 118*(1-2), 7-24.

Puckette, M. (1996). *Puredata.* Paper presented at the International Computer Music Conference conference, San Francisco, CA.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press. Cambridge, Mass.

Quenk, N. L. (2000). *Essentials of Myers-Briggs Type Indicator Assessment*. Wiley. New York ; Chichester.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, 72*(2), 257-286.

Rabiner, L. R., & Juang, B. H. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 4-15.

Raine, A. (1991). Are Lateral Eye-Movements a Valid Index of Fucntional Hemispheric Asymmerties? *British Journal of Psychology, 82*, 129-135.

Reeves, B., & Nass, C. I. (1996). *The Media Equation : How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press. New York.

Remland, M. S., & Jones, T. S. (1995). Interpersonal Distance, Body Orientation, and Touch: Effects of Culture, Gender, and Age. *Journal of Social Psychology, 135*(3), 281-297.

Reynolds, C. W. (1987). Flocks, Herds, and Schools: A Distributed Behavioral Model. *SIGGRAPH (Computer Graphics), 21*(4), 25-34.

Rhetorical. (2002). Rvoice (Version 2002 (no longer trading)),

Rhoads, K. (1997). Working Psychology – Everyday Influence. http://www.workingpsychology.com/evryinfl.html (Retrieved 17/09/2007)

Rickenberg, R., & Reeves, B. (2000). *The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces* Paper presented at the SIGCHI conference on Human factors in computing systems conference, The Hague, The Netherlands.

Riggio, R. E., & Feldman, R. S. (2005). *Applications of Nonverbal Communication*. L. Erlbaum Associates. Mahwah, N.J.

Rozak, M. (1996). Talk to Your Computer and Have It Answer Back with the Microsoft Speech Api. *Microsoft Systems Journal, 11*(1).

Ruttkay, Z., & Pelachaud, C. (Eds.). (2004). *From Brows to Trust* (Vol. 7). Kluwer Academic Publishers. Dordrecht.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A Simplest Systematics for the Organisation of Turn-Taking for Conversation. *Language, 50*, 696-735.

Sauerland, U. (2007). *Presupposition and Implicature in Compositional Semantics*. PalgraveMacmillan. Basingstoke u.a.

Schmidt, R. A., & Lee, T. D. (2005). *Motor Control and Learning : A Behavioral Emphasis* (4th ed.). Human Kinetics. Champaign, IL.

Screaming Bee. (2006). Morphvox: Screaming Bee LLC, http://www.screamingbee.com/product/MorphVOX.aspx

Selfridge, G., & Neisser, U. (1960). Pattern Recognition by Machine. *Scientific American, 203*(3), 60-68.

SPSS Incorporated. (2006). Spss (Version 14): SPSS Incorporated, http://www.spss.com/

Steels, L. (1990). Towards a Theory of Emergent Functionality. In Meyer, J.-A. & Wilson, S. W. (Eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior, Sab'90* (pp. 451-461). The MIT Press. Cambridge, MA.

Stern, S., Mullennix, J., Dyson, C., & Wilson, S. (1999). The Persuasiveness of Synthetic Speech Versus Human Speech. *Human Factors, 41*(4), 588-595.

Stone, B., & Lester, J. (1996). *Dynamically Sequencing an Animated Pedagogical Agent.* Paper presented at the Thirteenth National Conference on Artificial Intelligence conference, Portland, OR.

Tepperman, J., Traum, D., & Narayanan, S. (2006). *Yeah Right: Sarcasm Recognition for Spoken Dialogue Systems.* Paper presented at the InterSpeech ICSLP conference, Pittsburgh, PA.

Thorisson, K. R. (1996). *Communicative Humanoids. A Computational Model of Psychosocial Dialogue Skills.* MIT. Cambridge, MA.

Tinbergen, N. (1952). *The Study of Instinct*. Clarendon Press. Oxford Eng.

Tobii Technology AB. (2006a). Clearview (Version 2.5.1): Tobii Technology AB, http://www.tobii.se

Tobii Technology AB. (2006b). Tobii X50 Eye-Tracker. http://www.tobii.se/dFX50.html, 2006)

Trager, G. L. (1958). Paralanguage: A First Approximation. *Studies in Linguistics, 13*(1-2), 1-12.

Trappl, R., & Petta, P. (1997). *Creating Personalities for Synthetic Actors : Towards Autonomous Personality Agents*. Springer. Berlin ; New York.

Troika Tronix. (2008). Isadora (Version 1.2). Brooklyn: Troika Tronix, http://www.troikatronix.com/isadora.html

Valve Corporation. (2004). Half-Life 2. Bellevue, Washington: Valve Corporation, http://half-life2.com/

Vicon. (2005). Http://Www.Vicon.Com.

Wang, J., Xu, Y., Shum, H.-Y., & Cohen, M. F. (2004). *Video Tooning.* Paper presented at the SIGGRAPH conference, Los Angeles, CA.

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.

Weisz, J., & Adam, G. (1993). Hemispheric Preference and Lateral Eye Movements Evoked by Bilateral Visual Stimuli. *Neuropsychologia, 31*(12), 1299-1306.

Weizenbaum, J. (1966). Eliza – a Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM, 9*(1), 36-45.

Wilbur, M. P., & Roberts-Wilbur, J. (1985). Lateral Eye-Movement Responses to Visual Stimuli. *Perceptual and motor skills, 61*, 167-177.

Wilson, M., & Wilson, T. P. (2005). An Oscillator Model of the Timing of Turn-Taking. *Psychonomic bullentin and review, 12*(6), 957-968.

Winograd, T. (1968). Shrdlu, http://hci.stanford.edu/~winograd/shrdlu/

Yngve, V. (1970). *On Getting a Word in Edgewise.* Paper presented at the Papers from the 6th regional meeting.

Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). *Can a Virtual Cat Persuade You?: The Role of Gender and Realism in Speaker Persuasiveness.* Paper presented at the Human Factors in computing systems (SIGCHI) conference, Montréal, Québec.

Zanna, M. P. (1996). *Advances in Experimental Social Psychology. Vol. 28*. Academic Press. San Diego, Calif.

Appendix A1 – Synthetic ECA verification Wizard script

S: *Hello*

W1: Hi there, my name is …. What's your name?

S: *response*

W2: I'd love to know about your house. Could you describe it for me? How many rooms there are? Who do you live with? Where is your house?

S: *response*

W3: Thanks. Do you like living there? Where would you prefer to live?

S: *response*

W4: Ahhh. Ok. On a different note: if you were given a million pounds today, what would you do with it and why?

S: *response*

W5: That's interesting. Unfortunately, I don't have a million pounds for you. Maybe we should talk about something else. I really like going on holidays, especially in winter. What's been your best holiday ever? Where did you go?

S: *response*

W6: Oh cool. I haven't been there before. I guess I'll put it on my list of places to go. I could really do with a holiday right now – I've been working so much. Maybe I'll just have to survive with a good night out. Any suggestions?

S: *response*

W7: That sounds good. My main hope is that the weather is sunny tomorrow so I can get outside for some fresh air. Any chance you've seen the weather forecast?

S: *response*

W8: Well, I'm not much of a believer in weather forecasts anyways. You'd have thought that now in the 21st century they could do a bit better. Maybe I'll just move somewhere that has nicer weather all the time. Spain? Ecuador? What do you think?

S: response

W9: Thanks, that's really helpful. Some day maybe it'll happen. Well, I've gotta go. It's been so nice chatting with you. Perhaps we can do it again some time?

S: possible response

W0: See you later then. Bye.

Appendix A2 – Synthetic ECA verification distraction task

How many pool balls are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many people are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many geese are there?

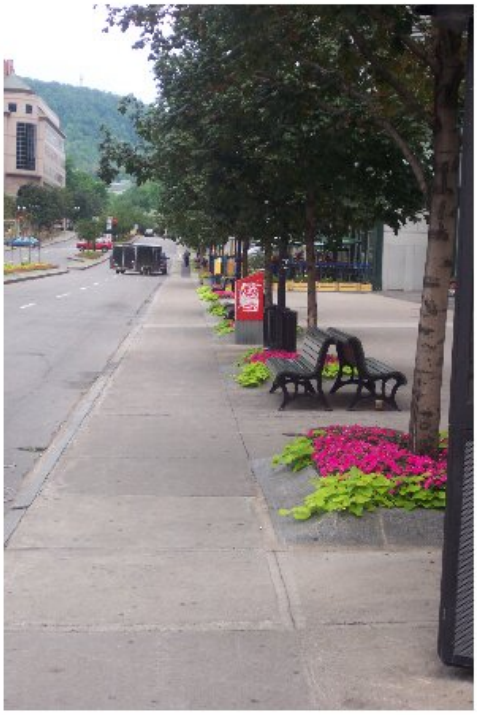
1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many pizza boxes are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many cars are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER



How many trees are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many burgers are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many people are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many balls are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many socks are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many tulips are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many trees are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many people are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many skiers are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many bright dots are there?

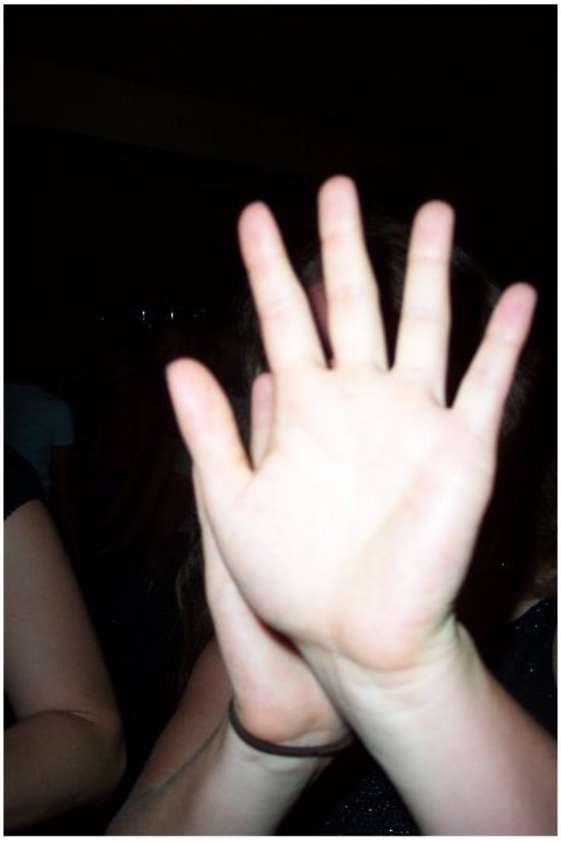
1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many motorbikes are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

How many pigs are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER



How many fingers are there?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 OTHER

Thank you

Please continue the conversation

Appendix A3 – Synthetic ECA verification questionnaire

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I enjoyed the conversation					
I learned something from the conversation					
The conversation was boring					
The conversation was difficult					
The conversation was engaging					
The conversation was interactive					
I would like to talk more with the character					
It was difficult to talk with the character					
The character led the conversation					
The conversation was natural					
I liked the character					
The character was interesting					
The character looked good					
The character looked at me					
The character was intelligent					
The character behaved realistically					
The character showed emotions					
The character was friendly					
The character was male					
I felt the character was confident					
The character was consistent					
The character listened to me					
The character showed facial expressions					
The character used the whole body during conversation					
The character's movement and speech were well coordinated					
The character understood me					
The character liked me					
The character was aware of me					
I felt threatened by the character					
I trust the character					
I felt in touch with the character					
The character made me anxious					
The character was interested in me					

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The character's body was human					
The character's body was computer generated					
The character's speech was human					
The character's speech was computer generated					
The character was a human					
The character was computer generated					

Please add any further comments you have about the character or the conversation below:

Appendix B1 – Synthetic ECA subject instructions

Instructions

Please put on the headphones and adjust the microphone to be in front of your mouth

To start the conversation with the character turn on the screen using the button labelled "start stop"

Then say "**Hello**"

Appendix B2 – Synthetic ECA character scripted sections

Conversation start (wait for them to speak):

S: Hello

W: Hi, my name is …….., what's yours?

S: response

W: Hi name. I'm here to talk with you about donating money to charity. To your right, on the desk is an envelope with your payment for taking part in this study. Could you just open it and check it has Ten Pounds in?

S: response

W: Great. I'm speaking on behalf of St Oswald's Hospice – specifically, the Children's service. Have you heard of it?

…

Closing:

W: Well, thanks for listening. If you would like to donate today please feel free to do so in the red box to your right, but first please turn off the screen and take off the headphones. Then you are free to go. The exit button is to the right of the door. Bye for now…

Appendix B3 – Synthetic ECA character information section

Welcome to St. Oswald's Hospice

St Oswald's opened to its first patient in 1986 to provide palliative care in the North East. That service has grown and expanded to meet the needs of the patients and families in the area and now we are one of the leading specialist centres in the country.

Children's Services

St Oswald's provides a specialist short break service to children with progressive, life shortening conditions.

We offer a 24-hour, 7-days a week service, supported by a team of skilled staff who can meet the complex health, emotional and social needs of the children and their families.

Our 'home from home' environment offers families a choice. They can either stay together every time, or own their own, safe in the knowledge that he or she will be cared for by our specialist team.

Children from birth to 18 are able to stay on our unit.

Children's Care Team

Our Children's care team includes nurses, physiotherapists, nursery nurses, health care assistants and volunteers.

Other members of our team include a chaplain, housekeepers, cooks, maintenance and admin staff.

St Oswald's medical team provides day-to-day cover. Out of hours medical cover is provided by a GP on call service.

We have access to a paediatric consultant but should we need advice, we will ask a child's own consultant.

However, should a child become acutely unwell while staying with us, we contact the emergency services.

How we do it

We are an independent, self-financing voluntary organisation. We are a registered charity and rely on voluntary giving, to ensure our essential services. We make no charge for our services, ensuring Hospice care is available to everyone.

The annual running costs for our adult services are approximately £4.3 million. We receive less than 30% of this sum from local Health Authorities. The remaining 70% of our funding comes through charitable giving.

Our Board of Trustees, led by Chairman, Tony Jameson, are responsible for managing the Hospice.

Everybody involved with St Oswald's – trustees, management, staff and volunteers alike – strive to abide by our Hospice Philosophy, which defines the values of the organisation for patients, families, carers and all those involved in its work.

The Story So Far

The Vision:

St Oswald's Hospice was founded in the early 1970's by Dorothy Jameson, a local lady who felt that North East people, facing terminal illness, ought to receive the same type of care and support offered by St Christopher's Hospice in London, where her daughter was working.

So, she set about talking to friends and members of the local church, as well as groups within the business, legal and medical professions – spreading the idea of a local hospice, encouraging them to get involved, share the vision and ensure her plans came to fruition.

Into Action:

Dorothy then organised a ten-man committee, responsible for finding a suitable site, appointing an architect and registering as a charity and limited company. In 1982, the committee launched an appeal to raise £2 million to build and run a local hospice. North East people gave their whole-hearted support to the project and we opened our doors in July 1986.

Although Dorothy sadly died over ten years ago, her legacy lives on through her son, Tony, who is Vice Chairman for St Oswald's.

Continuing Support:

While, there have been many changes since we opened in 1986 – most notably the addition of a purpose-built Day Services wing in 1997 and the opening of our Coleman Education Centre, a year later – the Hospice continues to be very well supported by local individuals, companies and organisations.

Such a ground swell of support has enabled us to make a further addition – a children's service, which opened in June 2003.

What we do

St Oswald's is a registered charity and provides hospice care to local adults and children.

Our adult service has gained a local, national and international reputation for our Specialist Palliative Care provision and through our Education Department, have pioneered significant advances in our field.

Within our Children's Service, we offer specialist short breaks to North East children with life shortening conditions.

We provide specialist care for children and support and advice for parents, within a relaxed home-from-home environment.

We make no charge for any of our services, ensuring hospice care is available to everyone.

Where we are now

As our running costs rise by over £1m to over £4m per year, never has it been more important for us to secure ongoing, regular giving to sustain our vital services to local people.

We rely on charitable funding, yet with the continued help of Jiggy, our Fundraising Mascot, we're hopeful everyone in the North East will continue to do their bit for St Oswald's.

There are lots of ways you can support St Oswald's.

Appendix B4 – Synthetic ECA character questionnaire

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I enjoyed the conversation					
I learned something from the conversation					
The conversation was boring					
The conversation was difficult					
The conversation was engaging					
The conversation was interactive					
I would like to talk more with the character					
It was difficult to talk with the character					
The character led the conversation					
The conversation was natural					
I liked the character					
The character was interesting					
The character looked good					
The character looked at me					
The character was intelligent					
The character behaved realistically					
The character showed emotions					
The character was friendly					
The character was male					
I felt the character was confident					
The character was consistent					
The character listened to me					
The character showed facial expressions					
The character used the whole body during conversation					
The character's movement and speech were well coordinated					
The character understood me					
The character liked me					
The character was aware of me					
I felt threatened by the character					
I trust the character					
I felt in touch with the character					
The character made me anxious					
The character was interested in me					

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The character was a human					
The character was computer generated					

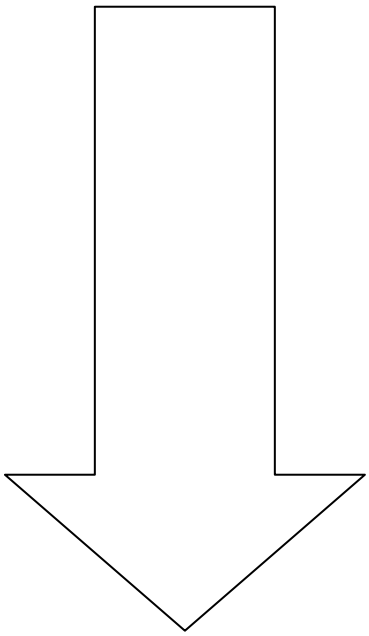
Please add any further comments you have about the character or the conversation below:

Appendix C1 – ECA subject instructions

Instructions

Please sit down and make yourself comfortable

To start the conversation with the character press the button below



Appendix C2 – ECA script

```
# Character Animation Script

### Initialization
:init

log $timestamp "SOFTWARE-RESTARTED"

set precache 0


display 0
clearcolor #000000

panx 0
pany -62.5
panz 90

anglex -69.6
angley 170.6
anglez 0.0

distance 359.67



#goto debug_voice
goto release_voice

:debug_voice
speechUrl "http://localhost:1666/voiceserver?"
speechVoice "sapi%3AMicrosoft+Sam"
speechPrefix ""
goto voice_end

:release_voice
speechUrl "http://localhost:1555/voiceserver?"
speechVoice "rapi%3AUKM001"
speechPrefix ""
#speechPrefix "\r(-1)+"
goto voice_end

#speechVoice "sapi%3AMicrosoft+Sam"
#speechVoice "sapi%3AMicrosoft+Mike"
#speechVoice "sapi%3ACepstral+Millie"
#speechVoice "sapi%3ACepstral+Lawrence"
#speechVoice "sapi%3ArVoice+UKM001+-+male"
```

```
:voice_end

speechPad 0.200

speechEnergyThreshold 300

nodAnimSet 3
nodDelayIn 0.2
nodDelayOut 0.2
nodWeight 0.8

stateDelay 0.4
stateWeight 0.6

# Script delays (just initial values - overwritten later
anyway after input choice)
set sectionWaitTime 1.5
set speakWaitTime 0.5

# (TIMES -- SOME ARE NOT AS OBVIOUS AS THEY LOOK!
THOROUGHLY TEST CHANGES, ESPECIALLY AFFIRMATION vs
INTERRUPT)
interruptWait 0.8
interruptAffirmWait 0.3
soundTimeout 0.8
minSoundAffirm 0.0
minSoundInterrupt 0.5

delay 0.2

state 0


goto first_start



:inject
state 2
eyeMovementScale 0.1
sayrepeat $inject1
eyeMovementScale 0.2
state 1
wait 0.5
return
```

```
:say

if $precache = 1 goto say_precache

#set recovery "say"
state 2
eyeMovementScale 0.1

set interrupt 0
echo "*** SAY 0 ***"
if $say_argc < 1 goto say_end
if $say_argc > 1 intsay $say1
if $say_argc = 1 sayrepeat $say1
if $interrupt = 0 goto say_end

set interrupt 0
echo "*** SAY 1 ***"
if $say_argc < 2 goto say_end
if $say_argc > 2 intsay $say2
if $say_argc = 2 sayrepeat $say2
if $interrupt = 0 goto say_end

set interrupt 0
echo "*** SAY 2 ***"
if $say_argc < 3 goto say_end
if $say_argc > 3 intsay $say3
if $say_argc = 3 sayrepeat $say3
if $interrupt = 0 goto say_end

set interrupt 0
if $say_argc < 4 goto say_end
if $say_argc > 4 intsay $say4
if $say_argc = 4 sayrepeat $say4
if $interrupt = 0 goto say_end

set interrupt 0
if $say_argc < 5 goto say_end
if $say_argc > 5 intsay $say5
if $say_argc = 5 sayrepeat $say5
if $interrupt = 0 goto say_end

:say_end
echo "*** SAY END ***"
```

```
eyeMovementScale 0.2
state 1
wait $speakWaitTime
return


:say_precache
delay 0.1
if $say_argc >= 1 say $say1
delay 0.1
if $say_argc >= 2 say $say2
delay 0.1
if $say_argc >= 3 say $say3
delay 0.1
if $say_argc >= 4 say $say4
delay 0.1
if $say_argc >= 5 say $say5
delay 0.1
return



:section
state 1
wait $sectionWaitTime
return



### Any special first-start code here
:first_start

goto start



### Script code
:start

# Setup - disable inputs while idle
eyeMovementScale 0.7
state 0
display 0
inputs 0
delay 0.2
```

```
# (show character)
display 1

# (wait for key-press)
delay

# (determine input condition and log the choice)
balanced_random inputs
inputs $inputs
if $inputs = 0 log $timestamp $inputs "START-INPUTS-
IGNORED"
if $inputs = 1 log $timestamp $inputs "START-INPUTS-
ACCEPTED"

# Set script delays based on input choice
if $inputs = 0 set sectionWaitTime 1.1
if $inputs = 1 set sectionWaitTime 1.5
if $inputs = 0 set speakWaitTime 0.5
if $inputs = 1 set speakWaitTime 0.5


# (introduction)
call say "Hi, my name is Alfie what's yours?" "Sorry, what
was your name?"

# (wait for response)
state 1
delay 0.2
if $inputs = 0 wait 4
if $inputs = 1 wait 4

state 2
call say "Hi there." "Hi"
call say "I'm here to talk with you about donating money to
charity." "I'm going to talk about donating money to
charity."
call say "To your right, on the desk is an envelope with
your payment for taking part in this study. Could you just
open it and check it has Twenty Pounds in?" "Does the
envelope have twenty pounds in?"

# (wait for response)
state 1
delay 1
if $inputs = 0 wait 6
if $inputs = 1 wait 6
```

```
call say "Great." "Ok."

### Introduction
call say "I'm speaking on behalf of St Oswald's Hospice - specifically, the Children's service." "I'm talking about the Children's service at St. Oswald's Hospice."
call say "St Oswald's opened to its first patient in 1986 to provide palliative care in the North East." "It opened in 1986."
call say "That service has grown and expanded to meet the needs of the patients and families in the area and now we are one of the leading specialist centres in the country." "The service is now a leading specialist centre."
call section

#goto quick

### Children's Services
call say "St Oswald's provides a specialist short break service to children with progressive, life shortening conditions." "St Oswald's provides services to children with progressive, life shortening conditions." "St Oswald's helps children with life shortening conditions."
call say "St Oswald's offers a 24-hour, 7-days a week service, supported by a team of skilled staff who can meet the complex health, emotional and social needs of the children and their families." "St Oswald's offers service 24 7, with a team of skilled staff who can meet the needs of the children and their families." "St Oswald's offers 24 7 services that help children."
call say "The 'home from home' environment offers families a choice."
call say "They can either stay together every time, or own their own, safe in the knowledge that he or she will be cared for by a specialist team." "Children can stay with their family or on their own."
call say "Children from birth to 18 are able to stay on the unit." "Children from birth to 18 can stay."
call section

### The Story So Far
call say "St Oswald's Hospice was founded in the early 1970's by Dorothy Jameson, a local lady who felt that North East people, facing terminal illness, ought to receive the same type of care and support offered by St Christopher's Hospice in London, where her daughter was working." "St
```

```
Oswald's Hospice was founded by Dorothy Jameson in the
early 70's to provide similar care and support as St
Christopher's Hospice in London." "St Oswald's was founded
in the early 70's by Dorothy Jameson."
call say "So, she set about talking to friends and members
of the local church, as well as groups within the business,
legal and medical professions - spreading the idea of a
local hospice, encouraging them to get involved, share the
vision and ensure her plans came to fruition. North East
people gave their whole-hearted support to the project and
the doors were opened in July 1986." "She set about talking
to friends and many local people, encouraging them to get
involved. North East people gave whole-hearted support and
doors opened in July 1986." "Dorothy set about encouraging
friends, and many local people in the North East to get
involved, which they did, whole-heartedly and St Oswald's
open in 1986."
call section

### Continuing Support
call say "There have been many changes since St Oswald's
opened in 1986 - most notably the addition of a purpose-
built Day Services wing in 1997 and the opening of our
Coleman Education Centre, a year later - the Hospice
continues to be very well supported by local individuals,
companies and organisations." "There have been many changes
since St Oswald's opened, but the Hospice continues to be
very well supported by local individuals, companies, and
organisations."
call say "Such a ground swell of support has enabled us to
make a further addition - a children's service, which
opened in June 2003." "All this grand swell support enabled
the opening of a children's service in June 2003."
call section

### Children's Care Team
call say "The Children's care team includes nurses,
physiotherapists, nursery nurses, health care assistants
and volunteers." "The Children's care team includes many
different staff."
call say "Other members of our team include a chaplain,
housekeepers, cooks, maintenance and admin staff." "and
also a chaplain, housekeepers, cooks, maintenance and admin
staff."
call say "St Oswald's medical team provides day-to-day
cover. Out of hours medical cover is provided by a GP on
```

```
call service." "If needed out of hours medical cover is provided by an on call GP."
call say "However, should a child become acutely unwell while staying with at St Oswald's, the the emergency services will be contacted."
call section

### How we do it
call say "St Oswald's is an independent, self-financing voluntary organisation and a registered charity and relies on voluntary giving, to ensure essential services. No charge is made for services, ensuring Hospice care is available to everyone." "St Oswald's is independent and self-financing and as a registered charity relies on voluntary giving. St Oswald's doesn't not charge for services so care is available to everyone."
call say "The annual running costs for the adult services are approximately 4.3 million pounds. Less than 30 percent of this sum is from local Health Authorities. The remaining 70 percent of funding comes through charitable giving." "St Oswald's adult services costs about 4.3 million pounds each year. 70 percent of this funding with through charitable giving."
call section

### Where we are now
call say "Running costs rise by over 1 million pounds, and so never has it been more important for St Oswald's to secure ongoing, regular giving to sustain our vital services to local people." "Running costs rise by over 1 million pounds. It has never been more important to secure ongoing, regular giving."
call say "St Oswald's relies on charitable funding, yet with the continued help of Jiggy, their Fundraising Mascot, they're hopeful everyone in the North East will continue to do their bit for St Oswald's." "St Oswald's hope that with the continued help of Jiggy, their fundraising mascot the North East will continue to support them."
call say "There are lots of ways you can support St Oswald's." "You can support St Oswald's in many ways."
call section

### End
:quick
call say "Well, thanks for listening." "Thanks for listening."
```

```
wait 1
call say "If you would like to donate today, please feel
free to do so in the red box to your right." "If you wish
to donate there is a donation box on your right, on the
table."
call say "The exit button is to the right of the door."
"Press the exit button right of the door to exit."
wait 1
call say "Bye for now and thank you." "Bye bye. Thank you."
"Bye. Thanks." "Thanks. Take care."
wait 3

# (hide character)
display 0
log $timestamp "END-SCRIPT"

# (wait for key-press)
delay

#goto start


:end
```

Appendix C3 – ECA character questionnaire

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I enjoyed the conversation					
I learned something from the conversation					
The conversation was boring					
The conversation was difficult					
The conversation was engaging					
The conversation was interactive					
I would like to talk more with the character					
It was difficult to talk with the character					
The character led the conversation					
The conversation was natural					
I liked the character					
The character was interesting					
The character looked good					
The character looked at me					
The character was intelligent					
The character behaved realistically					
The character showed emotions					
The character was friendly					
The character was male					
I felt the character was confident					
The character was consistent					
The character listened to me					
The character showed facial expressions					
The character used the whole body during conversation					
The character's movement and speech were well coordinated					
The character understood me					
The character liked me					
The character was aware of me					
I felt threatened by the character					
I trust the character					
I felt in touch with the character					
The character made me anxious					
The character was interested in me					

For each of the statements below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The character talked about giving money to charity					
The character wanted me to give money to charity					
I know what charity the character was talking about					
It was clear that the character was not from the charity					
I felt pressure to donate money					
I wanted to donate money					
I want to know more about the charity					
I liked the charity					
The charity was a worthy cause					
The charity needs money to keep running					
I felt influenced by the character					
The character didn't affect how much money I gave					
I thought the character was manipulative					
I felt the character was well informed					
The character made me feel giving money would be good					
The character could have been more persuasive					
The character felt the charity was worthy					
The character made me feel guilty					
The character was a human					
The character was computer generated					

Please add any further comments you have about the character or the conversation below: