MANA for the Ageing

David M W Powers, Martin H Luerssen, Trent W Lewis, Richard E Leibbrandt,
Marissa Milne, John Pashalis and Kenneth Treharne
AI Lab, School of Computer Science, Engineering and Mathematics,
Flinders University, South Australia
David.Powers@flinders.edu.au

Abstract

We present a family of Embodied Conversational Agents (ECAs) using Talking Head technology, along with a program of associated research and user trials. Whilst antecedents of our current ECAs include "chatbots" desgined to pass the Turing Test (TT) or win a Loebner Prize (LP), our current agents are task-oriented Teaching Agents and Social Companions. The current focus for our research includes the role of emotion, expression and gesture in our agents/companions, the explicit teaching of such social skills as recognizing and displaying appropriate expressions/gestures, and the integration of template/database-based dialogue managers with more conversational TT/LP systems as well as with audio-visual speech/gesture recognition/synthesis technologies.

1 Introduction

Embodied Conversational Agents (ECAs) are animated or robotic agents that engage users in real-time dialogue. As a development of the Chatterbot TT/LP system, they address a fundamental criticism of the Turing Test (TT) as incarnated in the Loebner Prize (LP), viz. the lack of understanding of the world, the lack of understanding people, the lack of personality (Harnad,1992; Shapiro,1992). This has in fact been acknowledge by Loebner who has insisted that more than "pen pal" conversation is necessary to win his $100K prize and Gold medal, and arranged design of a multimodal test [3]. At a technological level ECAs are a showcase for a large variety of language and human interface technologies including speech and face recognition and synthesis, speech understanding and generation, and dialogue management. However, at a deeper level they are a platform for exploring affect – the effect of multimodal features, including in particular expression and gesture on the human user.

Our aim is not to pass the Turing Test, although perhaps some descendant of our system will eventually do so. Rather our focus is to provide an effective agent for specific tasks where the limitations of current conversational companions, or dialog technologies, serve to match rather than conflict with the application constraints. Whereas limiting the topic was seen as a trick and a cheat in the Loebner Prize, our aim is to demonstrate and develop useful technologies and we are not interested in philosophical debates about intelligence. For these naturally constrained applications human level grammatical and syntactic understanding is not required, and the simple ELIZA-like approach of template matching is perfectly adequate as a first step (Weizenbaum, 1966).

Our initial Talking Head was based around the Stelarc Prosthetic Head[1] which combines multiple off-the-shelf components: keyboard input to a chatbot (*AliceBot*[2]) is linked to speech synthesis (*IBM ViaVoice*[3]) and 3D face rendering (*Eyematic*[4]). More recently we have adopted Head X[5] which is capable of generating a continuous, synchronized, optionally subtitled audiovisual speech stream in many different languages, with the ability to switch and modify voices and morph different faces at the same time as interacting with the user. The system is designed to be able to use different speech and face technologies, and we in general use Microsoft's SAPI[6] for speech recognition and generation plus the FaceGen face generation technology[7].

[1] http://www.stelarc.va.com.au/prosthetichead/
[2] http://www.alicebot.org/about.html
[3] http://www.ibm.com/software/pervasive/viavoice.html
[4] http://google.about.com/od/n/g/nevenvisiondef.htm
[5] http://csem.flinders.edu.au/research/programs/th/
[6] http://msdn.microsoft.com/speech
[7] http://www.facegen.com

Proceedings of the 2010 Workshop on Companionable Dialogue Systems, ACL 2010, pages 7–12,
Uppsala, Sweden, 15 July 2010.

2 Teaching ECA Applications

We have been predominantly exploring the application of our Talking Head as a virtual tutor of various subject areas. Initially our focus was language teaching/learning, but more recently demand for assistance with social teaching and assistant/companion applications has redirected our efforts.

The Talking Head has been extended for teaching and environmental/social interaction purposes with intelligent software that integrates inputs from various input sources such as cameras, microphones, touch sensors, and the like. A situational model is constructed that represents the physical environment in which encounters with the user take place. A teaching application can monitor a student's spoken utterances using both audio and video, can try to identify the student's facial expressions, and can make reference to physical objects in the surroundings (including specially-devised teaching 'props').

In addition to spoken utterances (the principal mode of output used in these applications), the Head may make use of audiovisual content presented on additional computer monitors and provide non-linguistic output that involves other sensory modalities, e.g. by making use of haptic devices. The multimodal capabilities of our ECA Teaching Agent are particularly valuable as they allow tutor and student to ground their interaction in a shared physical and social environment. Another invaluable aspect of our ECA for language teaching is the ability to model a student speaking the target language with a correct accent and authentic facial expression and gestures, with their own face and voice.

It is important in teaching, and in particular in language teaching, not to give the student any examples of incorrect or poor grammar, accent, etc. In a classroom context, students are held back and given poor example by other students, as well as by teachers who are not native speakers. Seeing or hearing their own incorrect written or spoken examples is immensely counterproductive. A good language teacher will reflect back, with appropriate degree of inflectional and gestural approbation, what they have said in corrected form. Having a close-up face as well as a voice to emulate allows unconscious recognition of the cultural and linguistic characteristics that are part of language, including the way of holding the mouth that affects even the way a person pauses or pronounces a neutral vowel sound, as well as the whole vowel system. With languages that have new consonants or vowels, or different variants that are treated as allophonic in their first language, seeing how those sounds are made can be very important to achieving an authentic accent. Body language, hand gesture, volume and tone, are all parts of this that are beyond the competence of current speech recognition and synthesis. This ability for our ECA to control vocal and gestural 'accent' is thus a primary focus of our research.

One specific application of the Language Teaching Agent is for teaching children with a partial or complete hearing impairment to speak and lipread, where the face rather than the voice is their primary cue. A related one is for teaching corresponding speaking and signing skills to their families. A third is for teaching literacy to indigenous children who have reasonable verbal competence in English (in our case) as a national language, as well as their tribal language and often a trade language as a first and second language.

Preliminary trials with comprehension testing found that appropriate facial expressions could enhance performance by a full grade point (Related-reference, 2008). However, it also identified that inappropriate expressions could negate this advantage – in particular it seemed that in one case the ECA was seen as laughing at rather than laughing with the subject matter. This has required us to modify our emotion model to include humour with both positive and negative affect. Moreover the emotional markup was performed by hand by one of the authors. We are currently engaged in a complex sequence of staged trials to develop appropriate ways of eliciting the desired AV expressions, getting multiple people to markup the texts, getting multiple subjects to classify and evaluate both real and head expressions, prior to undertaking a more comprehensive range of evaluations with the newly developed texts and markups, as well as a human head baseline. Currently there is very little in the way of audiovisual (as opposed to single image only) corpora of spontaneous or acted emotions and expressions.

Figure 1. Example of FaceGen morphing: female to male. Morphing is also used to provide speech gestures/visemes, emotion gestures/expressions, as well as explicit gestures like winks.

2.1 Social Tutors for Children

Once we started working with organizations that provided assistance to those with various disabilities and disadvantages, a major common factor emerged: the social problems that go with the disability or with looking different, or even just being from a different social or cultural background. Social skills tutoring of children with autism, hearing impairment and other disorders looks to be a promising application of our ECA Teaching Agent, which can accurately model facial expressions, and whose appearance and interactions can be customized to meet learners' needs. Initially we have focused on children with Autism Spectrum Disorders and our initial trials are in this ASD community.

Individuals with autism typically lack the skills needed to participate successfully in everyday social interactions, particularly reading non-verbal cues. Additionally, sufferers often feel more comfortable learning through technology than with other people, who may be judgmental or unpredictable.

Two lesson sequences reflecting common difficulties for children with autism were developed, the first on basic conversation skills and the second on managing bullying. There was a 54% average improvement from pre- to post-testing for the managing bullying module and a 32% average improvement for the conversation skills module, showing clearly that learning can take place through this method (Related-Reference, 2009).

3 Independent Living for the Ageing

The Memory, Appointment and Navigation Assistant (MANA) system is a broad project to assist elderly people, and those suffering from dementia or other ailments, with independent living in the privacy of their own home and the dignity of an ongoing personal life style.

3.1 MANA Calendar

The initial MANA Calendar application utilizes Head X to provide a talking head companion with an interface to Google Calendar, allowing doctors/carers to enter appointments/events that are provided to patients by the Head on a flexible reminder schedule. Eventually, it will provide localized assistance on how to get to the appointment based on public timetables, trip-planners and previous visits, but currently this information is supplied by carers.

The initial Calendar application of the MANA system was developed in 2009 based on preliminary input from an Alzheimer's Association for deployment in the homes of Alzheimer's sufferers. A preliminary exploration of potential faces and voices was conducted using a focus group approach organized through the NGO. For this preliminary stage we developed a dozen representative face/voice/script combinations and had representatives of the community select (individually and anonymously) their preferred face and voice. In associated discussion, it was apparent that a major influence was how authoritative the ECA appeared, and this was influenced by both face and voice (as well as the accent as their were only a couple of high quality voices available for each of the different accents). Some comments indicated that the person was too young or not serious enough, while positive comments were along the lines of that's matron, or an orderly, or that's someone authoritative – I'd do what they told me. At a later stage, if we have funds for a comprehensive study, it would be interesting to examine this formally, but for now we believe our "experts" and have developed our trial around the two most popular and authoritative male and female faces and voices. As a final stage, we dynamically combined and altered their preferred faces to achieve those characteristics preferred by the group.

Figure 2. Four MANA faces selected by focus group.

These top four faces (Fig. 2) and the top four voices are those from which subjects are allowed to select the ECA for their trial. As our aim is to show the ECA in the best possible light, we aim to please and give the subject control over who it is they are inviting into their home – and they do seem to treat it as a person they are inviting.

The system comprises the following major components (Self-Reference,2010):

Web Calendar Appointment Interface: Essentially this interface works virtually identical to a standard Google calendar, where a doctor/carer can enter an appointment/event. The MANA Calendar then extracts the key aspects of the event (i.e: time, date, name, etc) and relays the information to the Calendar Manager.

Calendar Manager and Synapse Module: The central Calendar Manager converts the information into a coherent human-like message to be delivered by the Thinking Head, upon either a set reminder time or upon a person-event. As Synapse is used by system modules, intermodule communications ensure concurrent productions, e.g. the timing of voice audio and visemes (visual phonemes), appear as human-like as possible.

Thinking Head and SAPI/Mary Integration: This new Thinking Head was designed using FaceGen™ software and incorporates Mary and Nuance voices, giving greater flexibility than using the original Stelarc face and voice.

Face Detection and Motion Analysis Module: The system uses a camera which monitors the space the subject moves around in (or a part of it), and triggers upon detecting sufficient motion energy for a human body and a human face (using the algorithm of Viola & Jones (2004)). On detecting such a "person-event", the appointment message is then delivered to the subject.

Speech Recognition Trigger Module: At any time the subject can query the MANA Calendar system by uttering "MANA" and one of 3 key words "appointment" (for upcoming appointments), "date" (current date) or "time" (current time) subject to sufficiently low noise conditions. After making a timed announcement, the system enters a state in which the speech system is set to recognize several acknowledgements (like "OK").

MANA Calendar is being trialed in the homes of people with Alzheimer's disease during the first half of 2010. We require that there is at least one carer or health worker who is able to enter calendar information into Google Calendar for the primary subject. If we have a live in carer, or a spouse or relative in the carer role, we are also allowing them to enter their own appointments.

Currently we are using a multiuser Microsoft Speech Recognition system that is *not* trained to the specific user. For our (younger) voices tested pre-trial these gave pretty good results, but the system is sensitive to age and accent. We have therefore adapted the study to provide training opportunities (human and system) for those who cannot initially use the speech recognition system successfully.

In addition, we do have a back up mouse or switch arrangement that allows such a user to use the system, but we are not permitting use of this option at present. MANA Calendar is designed not to require use of either keyboard or mouse, and this is the condition that we are insisting on for our initial evaluation. MANA is meant to appear as a companion, not as a computer.

Another problem that we encountered is that the price point requested by the NGO was $1000-$1500, and for these experiments we are using a DELL Studio One which is really not quite fast enough for continuous speech. Thus if it is left on trying to follow a conversation, it ends up filling up its buffer which gives unacceptable response times. For this reason we not only require the user to say a specific keyword or name to get the attention of the system (by default, MANA), we also require the user to be looking at the ECA (Viola and Jones, 2004) before we try to interpret what they say as a command. This dramatically reduces the delays, although there is still a hiatus that is slightly longer than is comfortable (about two seconds rather than the desired one second). This problem does not appear when run on a more powerful machine.

3.2 Mobile Living

A straightforward extension to MANA Calendar is to implement it on a mobile phone. We are currently exploring a couple of options for both technologies and platforms, the latter possibilities include the iPhone, Windows Mobile and Google Android, each of which has its *pro*'s and *con*'s.

Already MANA Calendar has options to allow the carer/healthworker to enter directions, and eventually a library of directions will be built up so that commonly visited places/recurring events, will not need reentry of directions. With the Mobile extension, MANA can also popup with reminders, make use of GPS, and let people know when to get off the bus, etc. This naturally combines in with current directions in GPS navigation systems and aids, as well as systems for keeping track of the elderly.

3.3 Teaching/Training

There are also several extensions of MANA envisaged that make use of our Teaching ECA technology, including teaching social skills, providing personalized family oriented reminders, and bridges to other technologies.

We also aim to keep the client occupied and interested in current events, interacting with family and friends, and actively stimulated and mentally engaged. The selection and implementation of these specific task-oriented activities, as well as playing games or doing exercises, is not unique but is beyond the scope of this paper and will not be reviewed. Our focus here is the naturalness and appropriateness of interaction, and exemplifying the kind of task-directed interaction which is *not* beyond the scope of current ECA technology.

3.4 Companion Robots

One of the first news items on our technology described it as "Companion Robots", picking up very quickly on this potential, notwithstanding the crude Eliza-like interactions. Interestingly this comes round full circle to the kind of ethical questions about the use of computers that were raised in the mind of her creator by those who wanted to put her to work immediately (Weizenbaum, 1976). Weizenbaum argued that we shouldn't have computerized psychiatrists who didn't really understand their patients, even if they were using the same techniques the human experts employed. And the world agreed with him! What has changed?

In terms of ECA vs Eliza technology, not much – the dialogue for HeadX is based on Alice, who whilst not much different in many ways from Eliza, at least had origins that sought to provide her with visual connection to the world. The current versions of Alice, reflect AIML code that is very similar in principle to Eliza code, and don't reflect anything of the real world except through the medium of canned dialogue.

The issue of computer control is not limited to dialogue and the issue of competence – computer controlled trains and buses and planes have been shown to be more reliable than humans under specified conditions, but still tend to be under direct supervision. Computer-guided missiles are for better or worse under an even more removed level of control. Our homes are full of gadgets, and most of us spend more time interacting with a computer and/or watching television than interacting directly with a person.

So WE will leave the ethics to society to determine what it wants. In an age where more people will be retired than working within the next twenty to forty years in most western countries, a MANA-type companion looks to be more of a necessity than a desired outcome.

Anecdotally, from our discussions with the NGOs and their staff, those who have had a district nurse or social worker visiting on a regular basis, tend to be happier with a human visitor than some technological solution. But those who do not have someone visiting regularly are more apprehensive about having a stranger in the homes telling then what to do and sapping their independence, than they are having a technology that purports to do the same things, or mediates between them and a remote visitor who does not invade the privacy of their own home.

4 Conclusion: A Competent Companion

In summary, WE see the key issue as competence, and so will conclude by outlining our approach to building the competence of MANA as a companion, rather than a calendar.

Emotion, Affect and Attitude: As discussed, one of our main lines of research at present is exploring and expanding the range of expressions and emotions, developing an AV corpus of carefully elicited spontaneous natural emotions, and cross-evaluating versus acted/programmed expressions.

AV Speech Recognition/Synthesis: Currently we can control the expression of our avatar through markup that is based on human judgements about what particular morphs of the face appear to show, and which are hand tuned to someone's

idiosyncratic idea of what a particular emotion or expression looks like – it is already reasonably effective, but as an initial step has not been properly evaluated, although our initial evaluation results have shown that at least some of the markup is effective, and that some is not (without separating out at this stage the influence of the text and the mark up). The flip side of displaying an ECA face is recognizing human faces and expressions. Similarly there is a much neglected auditory synthesis and recognition side that goes beyond phoneme and word. Our motto is "one person's noise is another person's signal" and our aim is for both speech and noise to simultaneously analyze and account for all individual differences, gender and age characteristics, emotion/affect/attitude and related human attributes, as well as explicit social and linguistic gestures and expressions, including rhythmic and tonal prosody.

Dialogue Management and Understanding: Dialogue management is a term WE don't like in the context of companiable systems – it derives from use as a database front end for ordering pizzas or taxis. It has a very limited concept of understanding related to the specific application, and Eliza or Alice type systems are perfectly capable of giving arbitrarily good results just by learning a greater range of template-response patterns. Our companionable MANA system is grounded in the home environment and is being trained to talk about and monitor and react to what is going on in the home. At the moment it is focused on body language and facial expression, and shares with the ASD system an aim to understand and react appropriately. The Alice substrate already has a reasonably comprehensive dictionary built in, but all it can do with that is define things – it can't actually productively use the knowledge. The Stelarc-Alice substrate also has at least three distinguishable personae built in – one who is male and a performance artist, one who is female and pretending to be human, and one who is neuter and surprised that you thought it should have that human characteristic. The latter two are an amalgam of hundreds of different programmer/user enhancements, whilst the Stelarc persona is the work of a single person and reflects his wry humour so that at times it does feel like you are talking to him. We are building in access to a full encyclopedia, and the ability to answer a wide variety of questions from each entry. But this also is superficial without the ability to learn and reason.

Learning and Reasoning: From a technological Artificial Intelligence perspective, our primary focus is learning. Children learn from the time they are born (actually probably more like from about three months before they are born) and their learning and play are very similar to the research and experimentation of a scientist. Piagetian Psycholinguistics, and Piaget's 20 plus books on specific aspects of child learning, development and reasoning, views learning and reasoning as developing hand in hand, with the little scientists developing new insights and deeper reasoning models, and thus enabling learning more about their world, society, culture and language. Learning to speak and understand language involves making noises and making the connection between the vocal tract/facial articulations/gestures and the heard sounds. Unsupervised learning using supervised techniques is possible using cross-modal training. Approaches from Computational Intelligence based on simple models from genetics, ant colonies and bee swarms, also provide mechanisms and analogies that help see how a system can continuously adapt and improve. Generalization and reasoning are part of this. Our ability to learn language is not independent of our ability to understand the world but an extension of it, and the constraints and nature of language are strongly influenced by the constraints and nature of the world. This also includes meta-reasoning: our reasoning about the consequences of our logic, decisions and behaviour.

References

Stevan Harnad (1992) The Turing test is not a trick: Turing indistinguishability is a scientific criterion. SIGART Bulletin 3(4) pp. 9 - 10.

David M W Powers (1998) The total Turing test and the Loebner prize, Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, ACL, pp.279-280.

M. Schröder & J. Trouvain (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology, 6, pp. 365-377.*

Stuart C Shapiro (1992) The Turing test and the economist. SIGART Bulletin 3(4) pp. 10-11.

Paul A. Viola and Michael J. Jones, 2004. Robust real-time face detection, International Journal of Computer Vision, vol. 57, pp. 137–154.

Joseph Weizenbaum (1966), ELIZA - a computer program for the study of natural language communication between man and machine, CACM 9 (1): 36–45

Joseph Weizenbaum (1976), Computer Power and Human Reason: From Judgment To Calculation, San Francisco: W. H. Freeman