The Listening Room: A Speech-based Interactive Art Installation

Alexa Wright
Centre for Art Research and Technology Education
University Of Westminster
W1T 3UW, UK
alexa@dircon.co.uk

Alun Evans, Alf Linney
Ear Institute
University College London
WC1E 6BT, UK
a.linney@ucl.ac.uk

Mike Lincoln
Centre for Speech Technology Research
University of Edinburgh
EH8 9LW, UK
mlincol1@inf.ed.ac.uk

ABSTRACT

In this paper we will present *The Listening Room*, an interactive audio installation that holds more or less meaningful conversations with up to three people at any one time. Conceived as an artwork that explores the boundaries between virtual and 'real world' experience, *The Listening Room* incorporates a number of speech technologies. This paper will give an account of the conceptual framework for *The Listening Room* and will describe the technologies employed in its realisation. To give context to the ideas behind *The Listening Room* we will describe the work with reference to two previous interactive works by the authors - *Face Value* (2000) and *Alter Ego* (2005).

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: Performing arts; Fine Art

General Terms

Performance, Experimentation, Human Factors

Keywords

Speech technology, human-machine interaction, interactive arts, spoken dialogue systems

1. INTRODUCTION

This paper reflects on some intersections of art, science and technology in the domain of human-computer interaction, with particular reference to three interactive digital artworks. Focusing on the most recent of these,*The Listening Room* (2007) (figure 1), the discussion will address the human/machine interface and the use of computers to convincingly simulate social intelligence through spoken language. *The Listening Room* installation takes the form of an intelligent room that uses speech recognition and synthesis software, a dialogue management system, microphone arrays and directional sound sources to conduct disembodied dialogues with up to three individual audience members at a time. The installation is designed to have a transparent interface and to encourage the user to attribute human sensibilities to the machine, even in the absence of any visible human features.

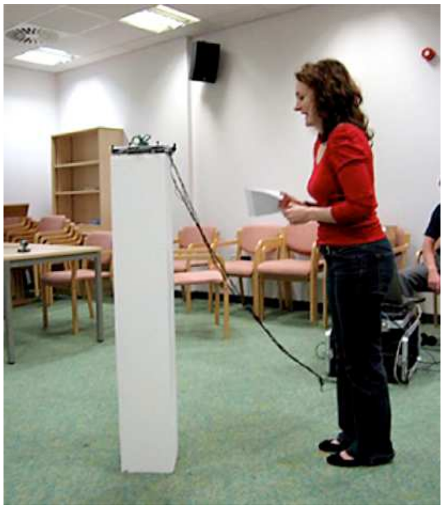

Figure 1: The Listening Room work in progress 2006

To give context to the ideas behind *The Listening Room*, we will also describe two previous interactive artworks by Alexa Wright and Alf Linney. The collaboration between Wright and Linney, which started in 1998, came out of a shared interest in the role of technology in measuring, interpreting and interacting with the human face. Research towards the two earlier works, which focus on embodied communication through the face, led us to explore the possibilities for a more immersive form of human/machine interaction. In *The Listening Room* a synthesised human presence is manifest as a disembodied voice that converses with individual audience members. Here, users are drawn into a more or less immersive social relationship with the technology, depending on their individual level of interaction. The physical presence of the user is fore-grounded in each of these three works, which progressively embrace an intricate and shifting relationship between mind, body, technology and identity. In each of these three installations the user him- or herself becomes the subject-matter of the work, and in each case the individual interacting with the

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

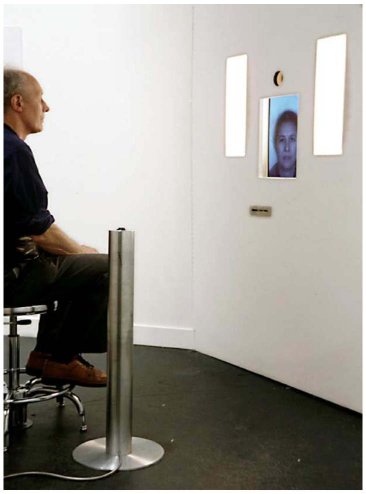

Figure 2: *Face Value* (2000) (installation shot)

work becomes a performer for other audience members. As a body of work these installations raise questions such as: "what if computers could convincingly perform human emotions?" and "can humans engage in meaningful social interactions with machines?".

Increasingly the artist seeks to create a seamless convergence of the real and the virtual, and to conceal the sophisticated technologies used to emulate human communication processes. Interestingly, however, whilst the intuitive and 'non-interventional' nature of the interface is fundamental to the user-experience, the limitations of the technologies employed also play a crucial part in the significance of this work. In the fissure between the subconscious, or intuitive fluidity of human communication and the rather more limited emotional and perceptive capabilities of the machine, a playfully self-reflexive situation is set up for the user.

2. FACE VALUE

The traditional machine model for human being, based on an idea of the brain as a computational device controlling all aspects of the self, does not go beyond a traditional Cartesian conception of mind and body as separate and separable components of the human self. However, now that cognitive neuroscience and related fields have scientifically demonstrated that emotion, rational thought and normal human function are not separable, it is interesting to note the move away from a machine-centred model for human being towards a human-centred model for the machine [10]. In the contemporary field of affective computing the desire to include machines as part of the human world is extended to attempts to give the machine social and emotional intelligence [21, 4]. The artworks described here all explore this process in varying ways.

Face Value (Figures 2 and 3), which references 19th century physiognomic principles [13, 15], relies on the computer's ability to measure the static features of the face. In the installation a computer screen displays a life-size image of an average face, derived from sixty different individuals. As the user sits in front of this screen, his or her own face appears superimposed over the average face. The interface to this work is a single button, which when pressed captures an image of the user's face. The computer then calculates differences between the dimensions of this face and those of the average. An individual 'character reading' is printed out based on these measurements.

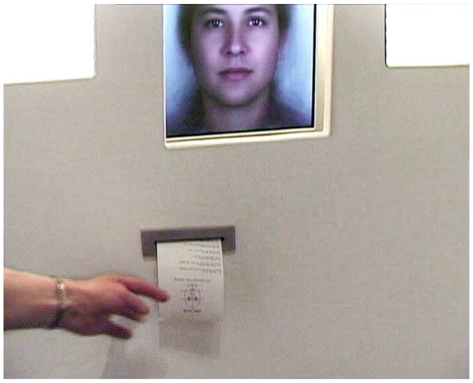

Figure 3: *Face Value* Detail

Here, the user's level of immersion in the technology is limited. Rather, the focus of the work is on the social interaction provoked amongst users as people gather to verify or deny the characteristics the system has attributed to them. Inevitably if someone smiles, or turns sideways whilst interacting with the piece they will receive a distorted reading. This fallibility of the technology has become integral to the work, which focuses its audience attention on the mechanisms we as humans use to read character from one another's appearance, rather than asserting the validity of any physiognomic system.

The human facility for almost instantaneously 'reading' socially and culturally inscribed information from another person's physical appearance suggests that human character may be linked to physical appearance, although not in the simplistic and formulaic way suggested by reductive systems such as physiognomy. The complexity of the task of holistic interpretation (i.e. scientifically measuring a person's character from his or her appearance) remains beyond even the most powerful systems of logic. Although it was designed to set up a discursive environment where the complexity of these interpretive mechanisms could be explored, *Face Value* also inadvertently points up the innate failure of the machine to contextualize and interpret information presented to it with the same intuitive flair as a human subject. *Face Value* characterizes the machine as a logician which, despite its formidable computational power, cannot avoid careless discrimination.

The most significant social exchange resulting from this work is not between the user and the computer, but between individual participants as they discuss the character attributed to them by the machine. However, despite the obvious lack of any real social intelligence in the machine, its apparent ability to assess the character of the user does offer the illusion of a meaningful social exchange with the machine, albeit on a very basic level. The queues of people waiting to use the work at public exhibitions testify to the compelling nature of a machine that enters into a dialogue with the user, particularly if the subject of this dialogue is the user him- or herself.

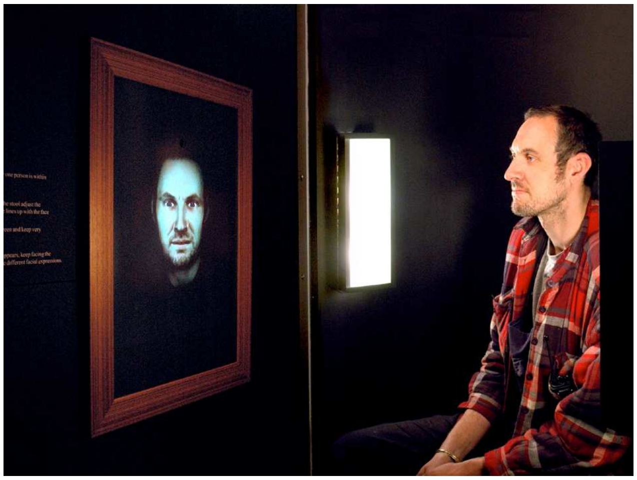

Figure 4: *Alter Ego* Installation

3. ALTER EGO (2005)

An interest in further studying and analysing the complexity of the human face and a desire to create a new work that would require measurement, analysis and re-creation of human facial expressions led to the development of ideas for a second installation, entitled *Alter Ego*. In order to render the interface transparent it was considered important that the artwork should be fully automatic. The only demand of the technology is that the user must sit still for some seconds so that the current technology for measuring and analysing the face may be fully exploited (Figure 4).

In this installation a stool with a curved black screen behind it is placed in front of what appears to be a mirror on a wall. The user is invited to sit still on the stool with a blank expression on his or her face for some seconds. The computer captures images of the face via a webcam located behind the 'mirror'. Initially, the system behaves like a mirror, reflecting the subject's face on the screen and mimicking his or her expressions. After about thirty seconds the reflection begins to react to, rather than mirror, the facial expressions of the user. For example: if the person viewed by the camera smiles, the virtual face may look surprised or angry, or may smile back. By making a range of facial expressions in front of the 'mirror', an individual audience member can interact with his or her own automatically-created avatar. The work plays on the common human experience of another self, or even several selves, that are generally unconscious, but rise to the surface in various contexts. By using a computer to create a semi-autonomous replica of the person sitting in front of it, the work invites the user to question the various facets of his or her identity.

To create both a living 'mirror' and an apparently autonomous personal image, a series of generic three dimensional (3D) polygonal facial models representing the end-point of fifteen facial expressions were constructed (Figure 5). These include small and broad smiles, surprise, anger, laughter, disgust, sadness and fear as well as more self-conscious expressions such as winking and poking out the tongue. The 3D models are warped to fit key landmark distances on a two dimensional video image of the individual face of each user. The 2D image of the user's face is then mapped onto the 3D model, which is animated by interpolation between models representing different expressions. Standard rendering techniques [11] are used to generate the images of the resulting 3D models, simulating the individual adopting different expressions. Originally the intention was to create an avatar that could perform facial expressions in the same way as the individual user. By tracking points on the face of the user and then animating these on the model in real-time the system would be able to reproduce a lop-sided smile or an idiosyncratic grimace. However, with the available resources the challenge of achieving this was too great and so the expressions performed by the existing avatar are generic. Despite this, the image of the user's face texture-mapped onto the model individualises the avatar to a surprising degree, creating an interesting tension between its likeness to the user and its awkward approximation of his or her actual facial gestures. Even the basic, symmetrical expressions the avatar is able to perform provoke a range of instinctive emotional responses from the user.

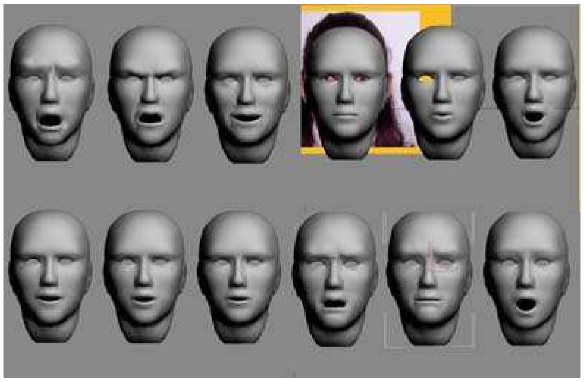

Figure 5: Morph Target Models

The question of whether it is possible for humans to engage in meaningful social interactions with machines was raised and debated in relation to the work, for example: The system

> "mimics - if only imperfectly at this stage - a real inter-subjective reaction. Although its own expressions are somewhat generic, it encourages individual, rather than generic responses from its users. The work demonstrates that machines can enter into meaningful - if limited - social exchanges with human operators. Further, it suggests that the meaningfulness of these exchanges is based, in part, on their playful and reflexive nature." [26]

The work demonstrates that, when the machine behaves in a way that can be interpreted as somewhat human, even a relatively standardised response from the avatar can elicit an emotional reaction in the user. In fact, it may be because the synthesised expressions of emotion in the artwork do not perfectly emulate 'real' human emotions that the user can become immersed in this work. The inadequacy of the technology to render a living likeness enables the user to engage imaginatively and spontaneously with his or her avatar. *The Listening Room*, which quite literally facilitates social interaction with the machine by engaging the user in spoken dialogue, was inspired by our observations of user interaction with *Alter Ego* and by a desire to experiment with another kind of interface between the 'real' and the 'virtual'. Both *Alter Ego* and *The Listening Room* belong to a new generation of digital artworks that begin to blur the boundaries between technological and 'real world' experience. By playfully encouraging the projection of personality into the machine, both works offer an illusion of meaningful social exchange.

4. THE LISTENING ROOM

4.1 Description

In *The Listening Room* installation (Figure 6) small sculptures are displayed on exhibition plinths in a traditional gallery space. People entering the space are automatically tracked using webcams positioned overhead. When someone moves past one of the sculptures the disembodied voice of 'Heather' tries to catch his or her attention by saying 'Hello', or 'Excuse me'. As an individual approaches one of the sculptures 'Heather' will then attempt to engage that person in conversation. Using keywords to interpret what is said in reply, she will try to pursue a more or less meaningful dialogue with the individual audience member. 'Heather' is able to conduct conversations at up to three different locations at any one time.

The idea for *The Listening Room* was initially inspired by watching people on the street conversing on hands free mobile phones. On seeing someone apparently talking to him- or herself in the street there is a moment of uncertainty as to how to categorise that person before the technology he or she is using becomes evident. It is this sense of uncertainty common to other human / machine interactions from intercoms to voice recognition systems - that we are interested in invoking both in participants and observers of the conversations in the installation. Although the sculptures provide a focus and talking point, the real subject of the conversations is the struggle to find common ground.

The installation is performative, in that individuals interacting with the synthesized voice become performers for other audience members. Although the physical installation of *The Listening Room* is complex, the technology is hidden and the work exists only when audience members engage in conversation with 'the voice'. An element of theatre is introduced into the work as individual users interacting with 'Heather' become performers for other audience members, who are able to hear only one side of a conversation.

4.2 Conversing with the machine

Whilst the first attempts to produce synthetic speech mechanically were made over two hundred years ago [23], the challenge both to create and to understand natural human speech is still current. Although speech synthesis technologies have progressed rapidly in recent years [5] and we are using one of the most advanced voices available, there are some occasional inconsistencies in prosody and pronunciation that belie 'Heather's' synthetic nature. Rather than posing a problem, however, this limitation of the technology is again an important aspect of the work. Whilst the appropriateness of 'Heather's' responses are sufficiently compelling that most people interacting with the work project personality and emotional content into the human/machine dialogue, the small inconsistencies in her prosody also provoke a state of uncertainty in the user.

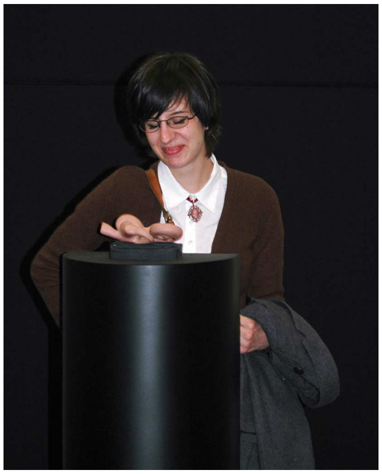

Figure 6: The Listening Room installation June 2007 (detail)

During test days we have observed that the success of an interaction between 'Heather' and a human user depends as much on the degree to which the user is prepared to invest in his or her interaction with the machine as on the ability of the machine to construct meaningful answers. For example, a shy person giving simple 'yes/no' answers will not have such a satisfactory interaction as someone who engages more conversationally with the disembodied voice. This practical observation is supported by the theories of Gordon Pask [20]. Pask suggests that language-oriented social systems are highly symbolic, with meanings implicitly agreed between human participants as a conversation progresses. He maintains that during a conversation an 'object language' is used to activate and demonstrate concepts, which are then explained and discussed using questions, answers and commands. An understanding can thus be reached through negotiation when the participants in a conversation engage in an 'I-You', rather than an 'I-It', transaction. The tree structure used to construct conversations in *The Listening Room* represents an attempt to identify and then to mimic this process. In most cases conversations between 'Heather' and the human user now flow relatively naturally, even when the system recognises only a few relevant keywords in the user's speech. This process is assisted by certain 'intelligent' details - for example, the cameras used to track an individual in the space also enable our system to determine what colour he or she is wearing and thus to comment on this. In addition, recognised keywords are repeated back to the user at some points in the conversation for example:

Heather:	You seem to be interested in the sculpture. what do you think of it?
Response:	I think it's really beautiful
Keyword:	BEAUTIFUL
Heather:	I'm glad you think its BEAUTIFUL, but do you think its art?
Response:	I think it is, yes.
Keyword:	IT IS
Heather:	IT IS, what do you mean by that?

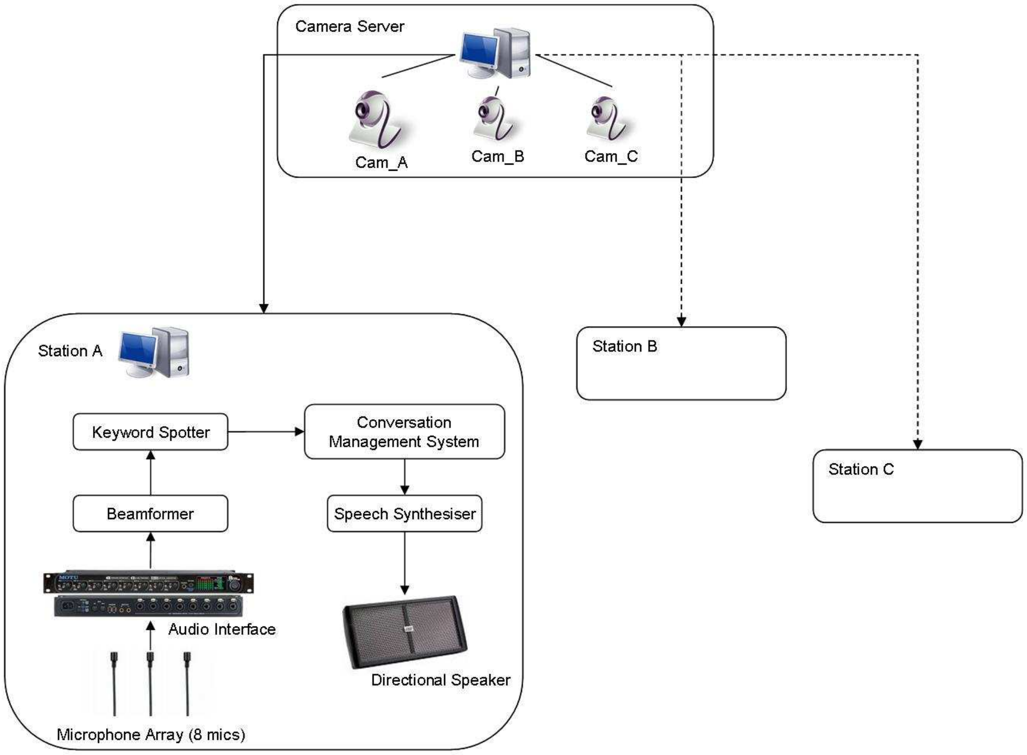

Figure 7: System Overview

The Listening Room has been developed empirically, with around ten to twelve different users involved at each test stage. In response to user feedback and to observed user interaction regular modifications have been made to the configuration of both the physical installation and the technologies used. This way of working has enabled us to respond to issues arising from 'live' user interaction that we could not have anticipated in the laboratory. It has enabled the content of the conversation tree to be developed iteratively so that, although 'Heather's' reactions are predominantly fixed, conversations appear natural.

A point of reference for users of *The Listening Room* is the Eliza Chatbot, originally designed by Joseph Weizenbaum in 1966 [24] to emulate a psychotherapist. Although Eliza recognises a number of keywords, she is mainly limited to repeating the user's questions back in a slightly altered form. Whilst some users have developed an emotional attachment to Eliza, her dependence on a conventional text-based interface renders the possibility of immersion relatively slight. Whilst working on *The Listening Room* we have observed that the sense of hearing is particularly immersive. Language is fundamental in the formation of the human subject, and the process of translating sounds into language from which meaning can be derived is very different to that of interpreting visual stimuli. With the projection of focused sound that can appear to emanate from within the user's head, the medium of spoken language enables us to render the interface more intangible than in previous works and thus to further blur the boundaries between the real and the virtual.

4.3 Artistic Context

The Listening Room has been compared to The Prosthetic Head by Australian artist Stelarc. However, the two works differ in several important ways. The Prosthetic Head incorporates a large video projection of a 3D model of the artist's head that responds to typed input from the user with facial expressions and spoken statements. This work, which relies on a conventional keyboard interface, neither aims towards transparency of interface nor claims to be immersive. Instead the artist draws our attention to the direct relationship between the operating program and the performance of the Head" [6]. Stelarc has employed an embodied conversational agent, whilst in The Listening Room the voice is deliberately disembodied. 'Heather' relies entirely on the user's imagination for her embodiment. In this sense the work is more closely aligned with Janet Cardiff's audio walks, where the user is taken on an imaginary journey through a real (physical) landscape. The voice of the artist and other binaurally recorded sounds are played back through headphones. Although Cardiff's works are not interactive, the sensed reality generated through sound and compelling narratives render them truly immersive.

4.4 Technology behind The Listening Room

4.4.1 Overview

The exhibition incorporates a number of interacting technologies to achieve its aim. These are summarised in figure 7.

A real-time video detection system tracks the movement of

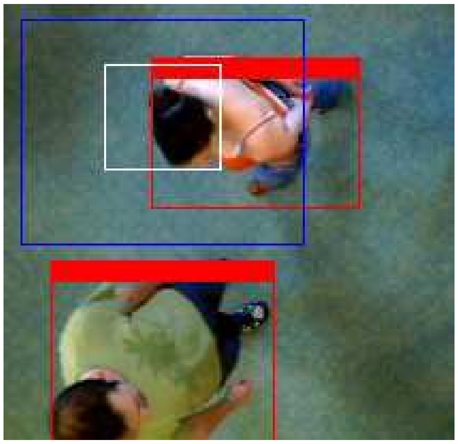

Figure 8: Video Detection System

visitors within fields of view of several cameras, thus noting the approach of a visitor and triggering the room's initial statement. The voice of the room is generated using the Cereproc Speech synthesis engine CereVoice. The responses are played to the user by means of HyperSonic Sound technology speakers meaning that only people standing close to the plinth can hear the voice. Once the room's opening comment has been made, the conversation is directed by a 'conversation tree', with the system responding to keywords uttered by the user which are recognised by a speech recognition system.

The system shares many technical similarities with more conventional 'goal orientated' spoken dialogue systems, in which the user is attempting to complete a defined task - be it book a flight or find a train time [16, 17]. The basic architecture, comprising a speech recogniser coupled via a dialogue manager to a synthesiser is typical of such systems, however *The Listening Room* is different from them in many respects which are described below.

4.4.2 Video Detection System

The video detection system uses a series of webcams placed directly above each sculpture. The movement of users entering the field of view of each camera is tracked so that the system can spontaneously initiate a conversation when a visitor approaches one of the plinths. A background mask representing the image when there are no users within the camera's field of view is subtracted from the current camera frame [25]. Any major differences between the current image and the background image are labelled as moving objects and are tracked as they move. The background mask is 'adaptive' in that it is gradually updated to ensure that small alterations in the 'empty' camera-feed (for instance changes in light levels) do not generate false positive tracking errors. The sensitivity of the tracking can be adjusted to cater for different lighting environments. The grey boxes in figure 8 show users being tracked.

Within the area encompassed by the field of view of each camera, two rectangles are defined which act as 'trigger' areas. When a visitor enters trigger area A, represented by a white box in figure 8, the installation begins a conversation targeted to that area. If the conversation has not reached a conclusion when that person steps out of trigger area B(represented by a black box in figure 8), the conversation ends prematurely. To allow users some degree of movement while involved in a conversation, Area B is significantly larger than area A.

The use of the video system extends beyond merely tracking movement and initiating conversation. It also enables the use of visual information to affect the comments made by 'Heather'. For example, the camera uses colour histograms to find the dominant colour contained within a particular tracked region. From this, the conversation management system assumes that this dominant colour is that of an item of clothing worn by the user, and can then initiate conversation by instructing 'Heather' to comment directly on the user's clothing e.g."That green colour looks nice on you. Do you often wear green?"

A goal orientated dialogue system has no requirement for video based user tracking - Users initiate the dialogue by either calling it in the case of telephone based systems or by activating it by means of a button or touch screen in the case of a kiosk based system. The system is therefore always aware when a user is present.

4.4.3 Speech Synthesis system

The speech synthesiser used to generate 'Heather's' responses is the commercially available Cerevoice system from Cereproc [2]. Cerevoice is a unit selection synthesiser. Unit selection synthesisers [14] take a large database of recorded speech and segment them into smaller units (in this case 'diphones' or sound-to-sound transitions). An index of these units is then made based on the segmentation and other acoustic properties such as pitch and duration. During synthesis the text is translated into a series of pronunciation symbols. A set of target diphones is then extracted as well as a set of target features for each diphone e.g. desired pitch. A Viterbi search is used to find the optimal set of units in the database that match the desired target features and also join well to each other based on a set of join features. The selected units are then concatenated together to form the output speech. The system generates 'Heather's' responses 'on-the-fly' which means that the conversation tree used to select the responses may be edited without having to record any new prompts that have been added. This has been particularly important given the 'incremental adaptation' approach to the design of the system which has been followed.

The Cerevoice system is tailored to preserve the variation present in the original speaker used to record the database from which the units are taken, and therefore increase the sense of character in the resulting voice. This is important in *The Listening Room* where the voice is trying to actively engage the user in conversation rather than simply provide information. The more interesting and characterful the voice, the more likely it is to achieve this aim and hold the attention of the user.

4.4.4 Audio Delivery System

This audio delivery system is very different from the handset or speakers used in previous dialogue systems and is driven by the requirement to maintain a transparent user interface. 'Heather's' responses are delivered to the user by means of Hypersonic Sound (HSS) speakers. This system modulates an ultrasound carrier with the desired speech signal and relies on the self-demodulation property of the air to render the sound audible along a narrow beam in the direction the speaker is pointing [27]. The result is a sound which seems devoid of a source and appears to originate from a point close to the listener's ear. The voice is rendered practically inaudible in other parts of the room by sound insulation in the form of acoustic foam panels placed on the wall in front of each speaker. These panels, concealed behind acoustically transparent curtains, minimise reflections of the sound beam and render the synthesised speech as directional as possible. In addition, a 'resonance speaker' is used to deliver low-level background crowd noise into the exhibition space. These speakers, when placed on any flat surface (for instance a table or window) turn that surface into a sound source which is audible throughout the room. Because of the size of the resonance surface the sound is not easily localised and this ambient noise helps to mask reflections of 'Heather's' responses in locations away from the plinth. The transparency of the interface is again maintained since the resonance device itself is small and easily concealed, and the surface is (quite literally) part of the furniture.

4.4.5 Conversation Management System

The conversation management system in *The Listening Room* takes the form of a tree structure with each node representing a prompt that 'Heather' reads and a list of zero or more keywords which the user is likely to include in their reply to that prompt. The path taken between nodes is determined in one of two ways -

- If, after the prompt is read, the system recognises one of the keywords in the users reply, then a branch associated with that keyword is followed.
- If the system does not recognise a keyword (or there are no keywords in the list) then a random branch is taken from that node.

Taking a random branch ensures that, even if the recogniser has difficulty spotting keywords for a certain user, they are unlikely to have the same conversation with 'Heather' on more than one occasion. Conversations are designed in such a way that, even when several random branches are selected during one dialogue the user usually has the impression that the system is responding to what he or she has just said. The conversation tree is designed using the Rapid Application Developer from the CSLU Speech and HCI toolkit [1], and the RAD output subsequently parsed for use in the conversation management tree. Using the toolkit allowed the tree to be visualised during construction, and rapidly adapted during the testing phases of system development.

While the implementation of the dialogue manager is similar to many other dialogue systems, the content of the tree is quite different. In a goal orientated system the user has a specific task they wish to perform and the dialogue is designed to allow them to reach that goal as quickly and accurately as possible. Checks are included to ensure information gathered is correct, and systems are engineered to make input from the user as precise and unambiguous as possible. In contrast, the user has nothing to gain from conversing with 'Heather', there is no 'correct answer' and no reason to attempt to constrain the users responses. This leads to a very different strategy for the design of the dialogue tree. The tree is designed to actively engage the user for as long as possible; open questions are the norm ("what is art" for example) and extending the length of a conversation is seen as a positive outcome.

4.4.6 Speech Input

Creating a transparent interface poses constraints on the audio capture devices for the keyword spotter and again differentiates *The Listening Room* from other dialogue systems. Where previous publicly installed systems have used microphones embedded in kiosks specifically designed to reduce the level of competing backgrond noise, those for *The Listening Room* must be concealled within the open space of the gallery. Whilst it would be possible to conceal a single microphone in the room and use the output of this for recognition, the amount of background noise and the cross talk from individuals at each of the sculptures talking to 'Heather' simultaneously, would severely limit the performance of the keyword spotter. Instead, a real-time microphone array beamforming system has been developed. A microphone array consists of a number of microphones placed at different locations within a room [3]. Signal processing techniques based on the theory of sound propagation are used to provide a single enhanced version of the signals from all the microphones based on the talker's location (a 'beamformer'). Beamforming systems have been shown to give performance in Speech recognition tasks which is far superior to using a single microphone located in the room [18, 19].

In *The Listening Room*, eight microphones are embedded in each display plinth and are captured via a 'Mark of the Unicorn 8pre' audio interface concealed within the plinth. Their output is subsequently processed by the beamformer to 'listen' in the direction of the person standing at the plinth - This direction is known, since they have already triggered the video tracking system's trigger area. The beamformer is implemented as a VST plugin [22]. VST is an industry standard architecture for creating realtime audio effects such as delays or echoes. The plugin runs inside a VST host application (in this case the host software is 'Plogue Bidule') which provides a simple interface between the plugin and the MOTU sound capture hardware.

The algorithm implemented is a superdirective beamformer - an optimal array processing algorithm which seeks to maximise the array gain while constraining the white noise gain [7, 8]. We also include a post filter that performs spectral masking of beams directed at the other plinths to effectively cancel the speech of others who may also be conversing with the system [18].

4.4.7 Speech Recogniser

The keyword spotting system is based on the ATK realtime extension to the HTK speech recognition toolkit [28]. A finite state language model comprising the keywords to be spotted in parallel with a monophone loop garbage model [9] is used for out-of-vocabulary word rejection. To improve keyword spotting performance, only the keywords associated with the current node in the conversation tree are active at any time, with the keywords being updated each time the conversation moves to a new node. The acoustic models for the recogniser are those developed for the AMI project meeting transcription system [12]. These are cross word triphone models trained on conversational telephone speech and then adapted to data recorded in a variety of meeting types. While typical conversations in the training data are very different from those we expect users to have with 'Heather', the type of speech they contain includes artifacts such as false starts, word repetitions and disfluencies. It is anticipated that users, while talking in the unfamiliar setting of an art gallery to a disembodied voice, will produce such artifacts and as such the system should perform better using such models than with those trained for a more typical dictation or read speech recognition system. The pronunciation dictionary for the system is also taken from the AMI meeting transcription system.

While we currently have no formal evaluation of the keyword spotting accuracy, recordings from the latest system evaluation day are being transcribed so that an empirical evaluation can be performed. In addition to providing quantitative performance measures, this data will also allow us to tune the parameters in the system to adjust the keyword false acceptance and false rejection ratio. Future evaluations can then explore the users experience of the system at various different ratios to identify the optimum operating point. It is interesting to note that what might be considered the optimum for such the keyword spotter in a more typical goal orientated dialogue system may not be appropriate here. Indeed it is often the 'misunderstanding' which occurs between 'Heather' and the user when the recogniser falsely recognises a word which leads to some of the more compelling conversations - for example, when the user tries to correct 'Heather's' misunderstanding, while she (with no mechanism for correction by backtracking up the conversation tree) continues the conversation regardless.

4.5 Physical Installation

As previously stated, *The Listening Room* interface is transparent, with all equipment concealed inside the plinths and in an adjoining room or specially built space (figure 9). In order to enable cables to run invisibly from the audio interface concealed in the plinths to computers in an adjoining space the floor will be carpeted. Although the physical installation of *The Listening Room* is quite complex, the user experience is simple. Visitors will initially see just three sculptures arranged on plinths in an apparently conventional gallery space.

The physical appearance of *The Listening Room* is still in development. The most recent work in progress is shown in figure 9. The sculpture shown in figure 10 will be cast in resin. The sculptures, which were initially smaller, were originally intended to encourage audience members to stand close to the plinths so that the voice of an individual interacting with the system could clearly be heard by the microphone array embedded in a particular plinth. However, the optimum distance for the beamformer is approximately 1 meter from the plinth and we have found that users often stand too close to the plinths - the larger sculptures encourage audience members to stand at the correct distance.

5. CONCLUSIONS AND FUTURE WORK

Because sound, and in particular spoken conversation, can become highly immersive this provides an ideal medium with which to explore a social interface between human and machine. A sense of immersion is achieved for the user because the interface can be rendered almost entirely transparent and because he or she is engaged in an active process of conversing with 'Heather'. We are currently

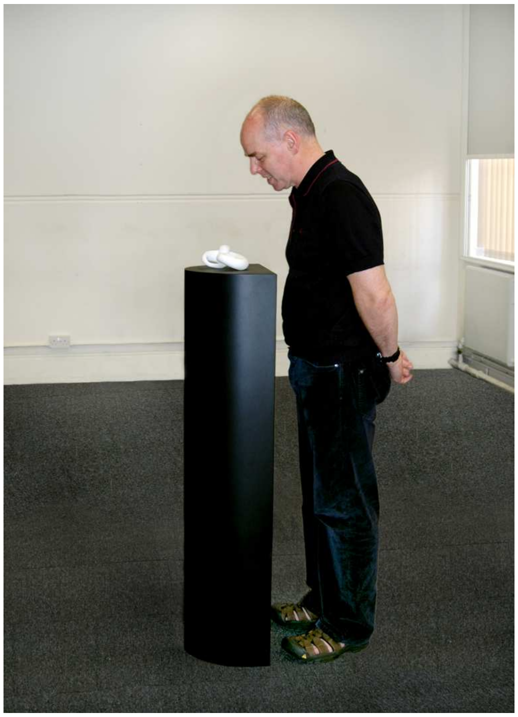

Figure 9: The Listening Room installation June 2007 (detail)

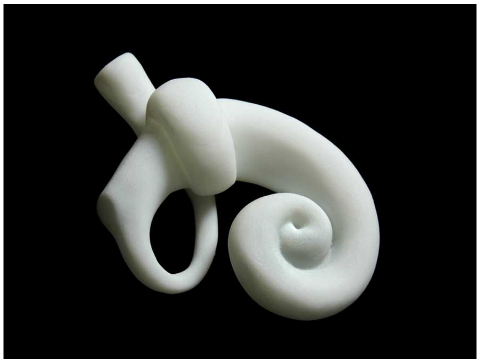

Figure 10: The Listening Room - sculpture

working to further enhance the user's sense of immersion. To this end we are researching the possibility of recognising and interpreting multiple keywords. This will enable us to construct tree nodes 'on the fly' and further reduce the predictability of each dialogue. We are also planning to introduce a memory to the system so that 'Heather can refer back to previously made statements.

To a certain degree the effectiveness of the works described here is enhanced by the limitations of the media. The minor prosodic errors which reveal 'Heather's' mechanical nature work in conjunction with her apparent social intelligence and use of language (an exclusively human trait) to provoke a interesting state of uncertainty in the user. Misrecognitions occasionally give the user the feeling that he or she is talking to a rather obtuse and distracted person who is, however, also prone to flirt with the user. It is in part this misbehaviour of the system, combined with 'Heather's' persistence in continuing the conversation regardless, which differentiate *The Listening Room* from more typical goal oriented dialogue systems which lack personality.

The aim of *The Listening Room* is quite different from that of previous spoken dialogue research because of the lack of a defined task or any measurable outcome from the user's interactions. This may lead researchers to dismiss the work as a mere toy with no relation to more 'serious' systems, however we feel that there are ways in which our work may be applied to these systems. Because our system tries to engage the user in as natural a way as possible we have observed that our users are rarely frustrated with their interactions with 'Heather' and that they are happy to continue conversations despite her apparent misunderstandings. By employing a more conversational tone, task based systems may be able to extend the length of dialogues, allowing more possibility to clarify the users input without explicit confirmation steps. The use of microphone arrays and directed audio to form the interface of the system also has important applications if multiple goal based systems are to be installed in close proximity to each other or in noisy locations.

The physical installation of the *The Listening Room* can be adapted to a variety of different spaces. It could ultimately become a guerrilla work, inhabiting an existing gallery display. With microphones embedded in plinths supporting works by other artists 'Heather's' intervention could take on a new and even more unexpected character.

6. ACKNOWLEDGMENTS

Research towards *The Listening Room* was funded by an Arts Council/AHRC Art/Science Fellowship. Production of the work is funded by a Wellcome Trust Sci/Art award. Work on *The Listening Room* at CSTR Edinburgh is supported by the European IST Programme Project AMIDA, FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

7. REFERENCES

[1] The cslu toolkit. http://cslu.cse.ogi.edu/toolkit/index.html.

[2] M. Aylett, C. Pidcock, and M. Fraser. *The CereVoice Blizzard Entry 2006: A prototype Database Unit Selection Engine.* http://festvox.org/blizzard/bc2006/cereproc_blizzard2006.pdf, 2006.

[3] M. Brandstein and D. W. (eds.). *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, 2001.

[4] C. Breazeal, A. Wang, and R. Picard. Experiments with a robotic computer, body, affect and cognition interactions. In *Proceedings of the Second International Conference on Human-Robot Interaction*, Washington DC, 2007.

[5] R. A. J. Clark, K. Richmond, and S. King. Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317–330, 2007.

[6] J. Clarke. Stelarc's prosthetic head. http://www.ctheory.net/articles.aspx?id=491, Oct. 2005.

[7] H. Cox, R. Zeskind, and I. Kooij. Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(3):393–397, June 1986.

[8] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(10):1365–1376, October 1987.

[9] H. Cuayhuitl and B. Serridge. Out-of-vocabulary word modeling and rejection for spanish keyword spotting systems. In *Proceedings of the Second Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*, pages 156–165. Springer-Verlag, 2002.

[10] A. Damasio. *Error: Emotion, Reason, and the Human Brain.* Penguin, 1994.

[11] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics: Principles and Practice (2nd Ed.).* Addison-Wesley Longman Publishing Co., Inc., 1990.

[12] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The ami system for the transcription of speech in meetings. In *Proc. ICASSP 2007*, Honolulu, Hawaii, USA., April 2007.

[13] P. Hamilton and R. Hargreaves. *The Beautiful and the Damned - The Creation of Identity in Nineteenth Century Photography.* Portrait Gallery, London, 2001.

[14] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the ICASSP 1996*, volume 1, pages 373–376, 1996.

[15] L. Jordanova. *Medicine and the Five Senses*, chapter 7. The Art and Science of Seeing in Medicine: Physiognomy 1780-1820. Cambridge, 1993.

[16] L. Lamel, S. Bennacef, J. L. Gauvain, H. Dartigues, and J. N. Temem. User evaluation of the mask kiosk. *Speech Commun.*, 38(1):131–139, 2002.

[17] M.A.Walker, R. Passonneau, and J. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proc of ACL-01*, 2001.

[18] I. McCowan, C. Marro, and L. Mauuary. Robust speech recognition using nearfield superdirective beamforming with postfiltering. In *Proc. ICASSP 2000*, volume 3, pages 1723–1726, 2000.

[19] M. Omologo, M. Matassoni, and P. Svaizer. Speech recognition with microphone arrays. In M. Brandstein

and D. Ward, editors, *Microphone Arrays*, pages 331–353. Springer, 2001.

[20] G. Pask. *Conversation, Cognition and Learning. A Cybernetic Theory and Methodology.* Elsevier, 1975.

[21] R. W. Picard. *Toward Machines with Emotional Intelligence, Chapter in The science of emotional intelligence: Knowns and unknowns.* Oxford University Press, 2007. In Press.

[22] Steinberg. *The VST SDK.* http://ygrabit.steinberg.de/ỹgrabit/ public_html/index.html.

[23] W. von Kempelen. Mechanismus der menschlichen sprache nebst beschreibung einer sprechenden maschine ("mechanism of the human speech with description of its speaking machine,"), 1791.

[24] J. Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, 1966.

[25] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[26] A. Wright, A. Linney, and E. Shinkle. Alter ego: Computer reflections of human emotions. In *Proceedings of the 6th Digital Art Conference*, Copenhagen, 2005.

[27] M. Yoneyama, J. ichiroh Fujimoto, Y. Kawamo, and S. Sasabe. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *The Journal of the Acoustical Society of America*, 73(5):1532–1536, May 1983.

[28] S. Young. *The ATK Real-Time API for HTK* http://htk.eng.cam.ac.uk/develop/atk.shtml