Conversation Piece: A Speech-based Interactive Art Installation

Alexa Wright
CARTE, University of Westminster
32-36 Wells Street
London W1T 3UW
+44 207 911 5000 x2333
alexa@dircon.co.uk

Alun Evans; Alf Linney
EAR, University College London
332 Grays Inn Road
London WC1X 8EE
+44 207 679 8926
a.linney@ucl.ac.uk

Mike Lincoln
Centre for Speech Technology
Research, University of Edinburgh.
2 Buccleuch Place
Edinburgh, EH8 9LW
+44 131 651 3175
mlincol1@inf.ed.ac.uk

ABSTRACT

In this paper we present *Conversation Piece*, an interactive audio installation that can hold conversations with up to three people at any one time. Conceived as an artwork that explores the boundaries between virtual and 'real world' experience, *Conversation Piece* incorporates a number of speech technologies. This paper describes the installation and gives an account of the technologies employed in its realization.

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 describes *Conversation Piece* (2007) (Figure 1), 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 users to attribute human sensibilities to the machine, even in the absence of any visible human features. Using computers to convincingly simulate social intelligence through spoken language, *Conversation Piece* is designed to create a seamless convergence of the real and the virtual and to raise questions such as: “what if computers could convincingly perform human emotions?” and “can humans engage in meaningful social interactions with machines?”.

A synthesised human presence is manifest in *Conversation Piece* as the disembodied voice of 'Heather', who converses with individual audience members. Whilst an intuitive and 'non-interventional' interface is fundamental to the user-experience of this work, the limitations of the technologies employed also play a crucial role in its meaning. In the fissure between the subconscious, or intuitive fluidity of human communication and the limited emotional and perceptive capabilities of the machine, a playfully self-reflexive situation is set up for the user.

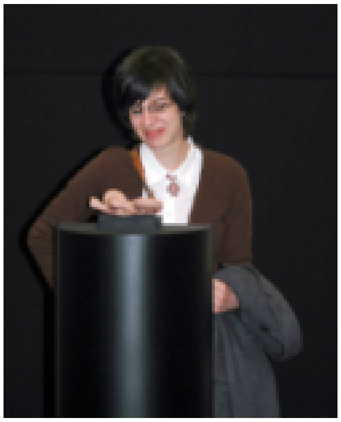

Figure 1: Conversation Piece – work in progress 2007

2. DESCRIPTION

In the *Conversation Piece* installation sculptures are displayed on exhibition plinths. 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 attempt to engage that person in conversation. Using keywords to interpret what is said in reply, she then will try to maintain a dialogue with the individual audience member. ‘Heather’ is able to conduct conversations at up to three different locations at any one time.

Rather than relying on traditional input devices to facilitate interactivity, the interface in *Conversation Piece* is rendered transparent by the use of concealed microphone arrays for speech input and focused directional speakers which ensure ‘Heather’s’ reactions can be heard only at a particular location in the space. This gives the illusion that ‘Heather’ is listening and responding to the user without any obvious physical interface, and encourages an intuitive engagement with the work. *Conversation Piece* makes use of the basic human experience of commonality through language to engender a sense of identification with the machine. This identification is

MM’07, September 23–28, 2007, Augsburg, Bavaria, Germany.
ACM 978-1-59593-701-8/07/0009.

further supported by the ability of the computer to emulate social intelligence and thus to distinguish itself from an inanimate machine.

3. CONVERSING WITH THE MACHINE

The desire to include machines as part of the human world is very current and is extended to attempts to give the machine social and emotional intelligence. *Conversation Piece* explores the idea and experience of a seamless and unencumbered convergence of the 'real' and the 'virtual'. The work playfully questions whether and in what ways it is possible for humans to have meaningful social interactions with machines. For each user this is mediated by the extent to which he or she projects personality or emotional content into the synthesized voice, and how much he or she chooses to engage with that personality.

The idea for *Conversation Piece* 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. Whilst 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 *Conversation Piece* is complex, the technology is hidden and the 'work' exists only when someone engages in conversation with 'Heather'.

Although speech synthesis technologies have progressed rapidly in recent years, there are some inconsistencies in prosody and pronunciation that belie 'Heather's' synthetic nature. Rather than posing a problem, however, this limitation of the technology is 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 inconsistencies in her prosody provoke a state of uncertainty in the user.

During test days we 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 have a less satisfactory interaction than someone who engages more conversationally. In most cases conversations between 'Heather' and the human user flow naturally, even when the system recognizes 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 also enable our system to determine what colour he or she is wearing and thus to comment on this. In addition, keywords are repeated back to the user at some points in the conversation, for example:

Heather: you seem to be interested in that sculpture - what do you think of it?

Recognised keyword: BEAUTIFUL

Heather: I'm glad you think its BEAUTIFUL, but do you think its art?

Recognised keyword: IT IS

Heather: IT IS, what do you mean by that?

4. TECHNOLOGY

The exhibition incorporates a number of interacting technologies to achieve its aim. A real-time video detection system uses an adaptive background mask to track movement within fields of view of several cameras, thus noting the approach of a visitor and triggering an initial statement. The voice is generated using the Cereproc Speech synthesis engine CereVoice, which allows the systems' spoken responses to be generated 'on the fly', without pre-recording. Responses are delivered through Hyper-Sonic Sound speakers, which use ultrasonic transducers to 'beam' sound in a particular direction, so that only people standing close to the plinth can hear the voice. Once 'Heather's' opening comment has been made the conversation is directed by a 'conversation tree'. This consists of a number of nodes representing the statements the system can make connected via a series of branches. Which branch the system takes at each node is determined by listening for certain keywords in the user's response.

One of the design criteria for the system was that the interface be completely transparent to the user. This poses constraints on the audio capture system for the keyword spotter. Such systems traditionally work best when the user wears a close talking microphone, however in this application this is neither practical, nor transparent. It would be possible to conceal a single microphone in the room, however, given the amount of background noise and the cross talk from individuals at each of the sculptures talking to the room simultaneously, keyword spotting performance would be extremely poor. Instead, we have developed a real-time microphone array beamforming system [2]. Microphones embedded in the display plinths are processed to 'listen' in the direction of the person standing at the plinth. The Keyword spotter is based on the ATK real-time extension to the HTK speech recognition system. A finite state language model comprising the keywords to be spotted in parallel with a monophone garbage model is used for OOV word rejection, with the keywords updated each time the system moves to a different node in the conversation tree.

5. ACKNOWLEDGMENTS

Research towards *Conversation Piece* was funded by an Arts Council/AHRC Art/Science Fellowship. Production of the work is funded by a Wellcome Trust Sci/Art award and collaboration with CSTR Edinburgh is supported by the EU 6th framework project AMI.

6. REFERENCES

[1] Pask, Gordon *Conversation, Cognition and Learning. A Cybernetic Theory and Methodology*. Elsevier, Oxford, 1975.

[2] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in Proc. ICASSP 2003, April 2003.