Designing and Implementing a Platform for Collecting Multi-Modal Data of Human-Robot Interaction

Brian Vaughan, Jing Guang Han, Emer Gilmartin, Nick Campbell

Speech Communication Laboratory
Centre for Language and Communication Studies
Trinity College Dublin, Ireland
7-9 South Leinster Street
Dublin 2, Ireland
E-mail: [bvaughan; hanf; gilmare; nick] @tcd.ie

Abstract: This paper details a method of collecting video and audio recordings of people interacting with a simple robot interlocutor. The interaction is recorded via a number of cameras and microphones mounted on and around the robot. The system utilised a number of technologies to engage with interlocutors including OpenCV, Python, and Max MSP. Interactions over a three month period were collected at The Science Gallery in Trinity College Dublin. Visitors to the gallery freely engaged with the robot, with interactions on their behalf being spontaneous and non-scripted. The robot dialogue was a set pattern of utterances to engage interlocutors in a simple conversation. A large number of audio and video recordings were collected over a three month period.

Keywords: human-robot interaction; multi-modal data collection; audio interface; platform-robot; WOZ, face detection

1 Introduction

Spoken dialogue systems have been applied to different fields including mobile communications, internet, games, talking terminals, handicap aids, etc. However, these interactive systems are mainly task-oriented and restricted to certain fields of expertise. Moreover, interactive speech technology is increasingly being incorporated into a number of ubiquitous consumer devices. While this is a welcome development, the combination of speech recognition and speech synthesis to create functional and

responsive interfaces is hampered by the unidirectionality of current synthesiser systems. Ideally synthesisers should have feedback and monitoring mechanisms to take account of the response and attitudes of the listener and adjust its output accordingly.

Work on Multiparty Interaction [1, 2] has demonstrated that a camera can be useful in processing human spoken interaction; therefore we developed a robot platform, providing the eyes and ears of the synthesiser, that is capable of observing the interlocutor throughout a conversation, as per [3, 4]. Furthermore, the text-to-speech aspect of most synthesisers assumes that there is a form of one-to-one mapping between text and speech which enables the processing of spoken language. While this one-to-one mapping between text and speech may be sufficient for the processing of linguistic aspects of human communication, it is not adequate to model the social information exchanged in many everyday human social interactions. In the last decade, considerable research has gone into making dialogues systems and robots more sociable. Examples are Semaine [5], Greta [6] or Max [7]. They incorporate modules for speech recognition, parsers (key-word-spotting) and speech synthesis. Despite their high complexity, however, these systems have not yet mastered social interaction.

Spoken interaction not only involves an exchange of propositional content but also the expression of affect, emotions, attitudes, and intentions of the speakers [8]. These are conveyed by different cues, which include backchannels, prosodic features and also non-speech sounds (clicks, sighs) as well as non-verbal information such as facial expressions, (co-speech) gestures and postures. It can therefore be assumed that a better understanding of these different levels investigated from a multi-modal perspective will enable spoken dialogue systems to reach a higher level of efficiency and to provide a more attractive user interface to a wide range of interactional technologies. We aimed to collect a number of audio and video recordings to better understand these issues and to inform the further development of an interactive synthesiser platform that utilises the visual modality to inform its output.

2 Objectives

In this paper we present details of the development of a system to capture human-robot interaction and its implementation in a real-world setting where naive interlocutors were able to engage with a conversational robot in real time. We present details of how the robot-platform was designed in order to enable multi-modal data collection of human-robot interaction. Using this platform we aimed to understand how the timing of simple responses can be used to engage and maintain human interlocutors in a conversational interaction.

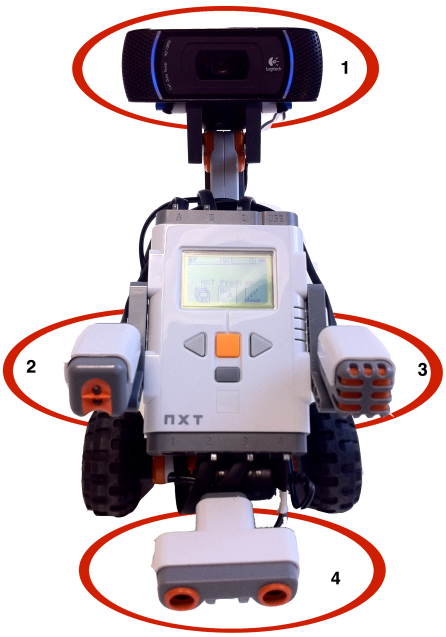

Figure 1
The ”Herme” Robot used in the experiment. (1) HD web-cam with built in speaker. (2) Light sensor, (3) Microphone and (4) distance centre

3 The Platform-Robot

3.1 Using the Lego Mindstorms NXT System

The robot was built using LEGO Mindstorms NXT technology [9], programmed in Python [10], and served as a mobile sensing platform for the camera. One computer controlled the robot, performed face tracking, and displayed the robots eye view, in real time, to participants, while another computer initiated a conversation and monitored participants reactions in order to step through the pre-determined sequence of utterances. The Mindstorms NXT system offers a versatile and configurable data acquisition platform: the Lego components can be configured to mount and control video cameras and microphones. Moreover, the NXT microcomputer brick can control a number of different motors and sensors, sending and receiving information to and from each. Using the Python programming language, these can be used to trigger responses based on the certain input criteria, e.g. the distance sensor was used to automatically move the robot rapidly backwards if someone tried to pick it up or interfere with its movement. The Mindstorms system used was the 9797

education NXT base set, which comprises a number of Lego building pieces, three servo motors, a light sensor, a sound sensor and an ultrasonic sensor as well as the NXT microcomputer brick. From this base kit, a moveable Lego robot was created that contained, along with the three NXT sensors, a HD USB video camera, a high quality microphone and a small loudspeaker.

3.2 NXT-Python: the Mind of the Robot

NXT-Python (PNXT) is a third party python based programming interface that was designed for the Lego Mindstorms NXT robot. It allows full access to the controls of the Lego robots leg, arm and head motors and the set of sensors, including a sound sensor, light sensor, ultrasonic sensor, etc. Communication between the robot and NXT Python was possible via bluetooth and USB. The Bluetooth option meant that no physical cable was required, but in preliminary tests it proved to be far less accurate than the USB connection. Precision of movement in response to the visual input and external control was necessary, so the USB connection was chosen as the primary method of communication.

3.3 Face Detection: the Robot's Eyes

Several face detection algorithms exist for determining the location and size of faces in an image. OpenCV was used for the face detection and image processing. Previous work [11] has shown that this software, with certain tweaks and modifications, is capable of correctly finding faces at more than 98% accuracy in a range of lighting environments. To lend more naturalness to the robots interaction, the face detection system was integrated with the wheel motors so that the robot turned a certain degree to face the interlocutor before beginning to talk. This was designed to give the impression that it was watching the speaker when talking. Gaze has been reported in the literature to be highly relevant in social interaction [12, 13, 14] and has been found to be important for modulating and directing attention [15]. The robot was therefore engineered to turn to face its interlocutor, and place them at the centre of the visual field of the camera. This face-detection program was run as a pre-process, sending output to the NXT Python program via the standard i/o stream of the computers operating system, over the USB connection.

4 Designing a Platform-Robot to Collect Robot-Human Conversations

Using a platform-robot to collect audio and visual recordings of robot-human interaction requires first addressing the complex issue of how to design a system which would enable a semi-natural, or natural-seeming interaction between a robot and

Figure 2
A robot's-eye-view of a conversational interaction

human interlocutors. The challenge was two-fold: designing a system (i) which is both simple in nature, and cost effective, while not equipped with speech recognition components, but (ii) which gives the illusion of intelligence. Our system does not emulate intelligence but uses common phatic and conversational devices to try to engage a human in a conversation without resorting to any complex information processing, using only utterance length information and synchrony of movement timing as well as body-movement (or head-movement) information to make inferences about the human participants dialogue acts.

5 Holding a Conversation

We tested a Max/Msp [16] based automatic conversational system during the early stages of the exhibition. An application was written in Max/Msp that used sound intensity thresholds to determine when a participant had started and stopped talking. This system worked well in some cases but due to the constant fluctuation of the background noise in the gallery, and the need to have more control over the timing of responses, this was replaced with a semi-automatic system and a Wizard of Oz setup, programmed in Python, to engage with participants. We used our findings from the Wizard of Oz setup to inform the timing of utterances in the semi-automated system.

5.1 The Dialogue

A fixed dialogue sequence was used to engage participants (see Appendix). The utterances in the dialogue were designed to be broad and applicable to a wide number of potential interlocutor responses. We hit upon an effective dialogue sequence very early on in the research and gained much insight regarding the perpetuation of a conversation through monitoring peoples responses to the fixed dialogue and altering the timing of responses accordingly. We have gathered a lot of data and this will serve as the basis for visual processing for automating monitoring of reactions in discourse.

5.2 Making Contact

Initial contact was made when a person entered the field of view of the robot, using open CV [17]. Once a face was detected, the robot orientated itself so the detected face was placed at the centre of the field of view. If there was more than one face, the largest face was deemed to be the closest and the focus of the robots attention. Simultaneously a simple greeting pattern was used: Hello? Hi followed by a pause and a repeat of hello and a final hi. We found this to be a simple and effective strategy for drawing passer-bys into a conversation.

5.3 Talking to People

Following the initial hello contact, the main body of the interaction was carried out by stepping through the fixed dialogue (See appendix). The process of stepping through the robots utterances was predetermined but the timing was variable; understanding and developing a conversation timing model was the main focus of the research. The environment of the Science Gallery was very noisy, so no speech recognition was incorporated into the robot. When it was too noisy to hear what an interlocutor was saying, the video feed was used to determine when someone was speaking and when they had stopped. The system was essentially a polite listener that, while not always replying with the most appropriate response, did reply with an appropriately timed response. For example, the question Do you like the exhibition? was followed after a short gap with really, and then after another short gap by why? and then a longer listening gap until the next phase of the dialogue was begun. By maintainingg control of the interaction, we were able to substitute polite listening for any form of understanding of what was said by the interlocutors. The pertinent aspect of the processing here was in the timing of the various utterance sequences. This was more successful when carried out by a human as part of a WOZ interaction, than by the automatic systems we tried that used visual, audio, and motion-detecting sensors. Interlocutors were encouraged to continue the conversation through the use of interjections. Once such interjection, I like your hair, almost always elicited a laugh and was very successful in keeping interlocutors in-

terested. Similarly, Do you know any jokes? usually elicited a negative response, to which the robot laughed, but the subsequent tell me a knock-knock joke was in almost all cases dutifully complied with, as was the polite listening to the robots own joke in turn. The conversation was ended by the robot thanking the interlocutor and asking them to sign a consent form and read aloud the unique identifier number on the form so that it was recorded, both on video and audio. In this way recordings could be matched to with the relevant consent form.

6 Collecting Audio and Visual Recordings

The interactions with the robot were recorded from a number of different angles. EvoCam [18] software was used to capture and record input from several video cameras in one video file. This simplified the management of the recordings as each interaction resulted in one video file containing synched audio and video from several different angles. Audio data were also collected using directional Sennheiser shotgun microphones mounted on top of a large tv screen that showed the robots eye view alongside a short animation that explained the basic concept of the robot and invited people to interact with it.

Two Sennheiser MKH60 P-48 shot-gun microphones were mounted at the top of the main screen, along with a Logitech C-910 HD webcam that provided a top- down overview of the interaction. On the platform itself were two robots, one engaged in interaction with the visitors while the other recorded the interactions using a Logitech HD Webcam to ensure that they were recorded from a more inclusive angle. Microphones on the webcams provided a close-up source of sound that could be used in conjunction with that of the shotgun microphones. An Apple i-Sight camera was mounted at the corner of the display to provide a wide overall view of the scene. A movement sensor was added during the latter half of the exhibition to trigger the onset and off-set of the conversations as an additional control sensor. Despite being able to monitor the vocal and gestural behaviour of interlocutors, we were not able to detect a switch of interlocutors if one walked away just as another came into the field of view.

There were three computers running inside the platform, out of sight: two Mac-Minis for the robot and a large Unix machine for the data collection and storage as well as providing the skype interface for the wizard in our lab who was able to control only the timing of each utterance. The default APPLE synthesiser was used as the voice of the robot, using the voice of Princess modified acoustically by a Roland Sound Canvas UA-100 to shift the formants and pitch upwards rendering the voice smaller and more appropriate to the device. The machines ran continuously, streaming all data to disk while the gallery was open. In all, we collected 433 signed consent forms and 1.5 terabytes of recordings from more than a thousand conversations. All recordings are securely stored but for legal reasons only those clearly including the consent form id number will be included in the final corpus.

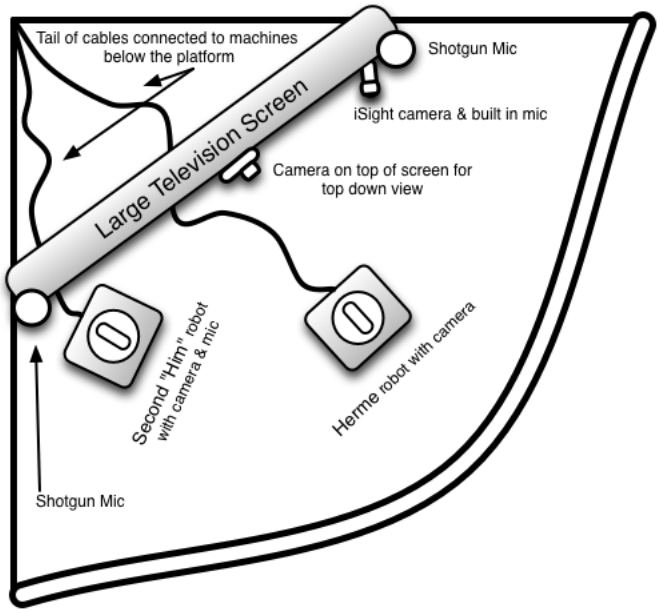

Figure 3
Schematic diagram of the experimental setup in the Science Gallery

Conclusions

This paper has presented a method for collecting natural human machine interaction in a noisy public environment without any external constraints on the participants. The system employed high definition video and multiple audio streams of these interactions. The corpus obtained in this way forms a valuable resource for the study of machine mediated discourse interaction. More than a third of respondents successfully completed the dialogue and signed consent forms for the extended use of the data. The rest, who left before signing a consent form, provide a valuable base for comparison by which we can improve detection of dialogue success. Several organisational, economical, ethical and legal issues were addressed. Specifically, a low-cost solution for the collection of massive amounts of real-world data. Participants were not paid and walked in off the street voluntarily. All age groups and social classes were included. The experiment and data collection method was overseen and cleared by the ethics committee of the University. We are now collating all the recordings, where consent was given, into a multimodal corpus of human-robot interaction that will then be annotated and analysed to further improve our understanding of machine mediated dialogues and HCI in general.

Acknowledgements

This work was undertaken as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631. We thank the Digital Hub and the Science Gallery for this opportunity to present this research in the Human+ exhibition from April to June 2011. Thanks to John Dalton for his help with programming the Max/Msp conversation module. Thanks to Dr. Celine De Looze, Catha Oertel and Ciaran Dougherty for their work as wizards.

Appendix

```
-hello? hi... hello . .
hi
- my name is hermee - herm e - hermee
whats your name?
how old are you?
- really
I'm nearly seven weeks old
- do you have an i d number
i need an i d number to talk to you
i d numbers are on your right
thank you
- are you from dublin?
- really
I'm from the Speech Communication Lab herein TCD-tell me about
you...
- really?
oh
- tell me something else
oh
really
- why are you here today?
really?
why
d'- do you like the exhibition
really
why?
i like your hair
- do you know any good jokes?
tell me a funny joke
ha ha haha ha
tell me a knock knock joke
who's there
who?
who
```

```
ha ha haha ha
- I know a joke
what's yellow and goes through walls
a ghost banana
ha ha hehe he.
ho hoho ho ho
- thanks for your help
goodbye, see you later
goodbye
```

References

[1] SSPNET: The Social Signal Processing site is host- ing FreeTalk (http://www.sspnet.eu/)

[2] AMI: The Augmented Multi-party Interaction project (http://www.amiproject.org)

[3] Chapple, Eliot,D.,(1939)"Quantitative analysis of the interaction of individuals", pp.58-67 in Proceedings of the National Academy of Science U S A. February; 25(2).

[4] Kendon, Adam, (1990) Conducting Interaction: Patterns of Behaviour in Focused Encounters. Cam- bridge: Cambridge University Press.

[5] Semaine, SAL, Sensitive Artificial Listener "http://www.semaine-project.eu/".

[6] Greta, Embodied Convcersational Agent "http://perso.telecom-paristech.fr/ pelachau/Greta/".

[7] E. de Sevin and E. Bevacqua and S. Chandra Pammi, C. Pelachaud, M. Schröder and B. Schuller "A Multimodal Listener Behaviour Driven by Audio Input", International Workshop on Interacting with ECAs as Virtual Characters, 2010.

[8] N. Campbell, "Getting to the heart of the matter; speech as the expression of affect; rather than just text or language", Language Resources and Evaluation,Vol.39, N.1, 109-118, 2005.

[9] LEGO MINDSTORMS NXT an intelligent mi- crocomputer brick http://mindstorms.lego.com/en- us/Default.aspx

[10] Python N. NXT Python Homepage. 2011.

[11] D. Douxchamps and N. Campbell, "Robust real time face tracking for the analysis of human behaviour", Proceedings of the 4th international conference on Machine learning for multimodal interaction, Springer-Verlag, 2007, pp. 1-10.

[12] A. Kendon, “Some functions of gaze-direction in social interaction”, Acta Psychologica, vol. 26, pp. 2263, 1967.

[13] M. Argyleand, J.Graham, “Thecentral Europeexperiment: Looking at persons and looking at objects”, Environmental Psychology & Nonverbal Behavior, vol. 1, no. 1, pp. 6–16, 1967.

[14] S. Duncan, “Some signals and rules for taking speaking turns in conversations”, Journal of Personality and Social Psychology, vol. 23, pp. 283–292, 1972.

[15] R. Vertegaal, R. Slagter, G. van Der Veer, and A. Nijholt, “Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes”, SIGCHI Conference on Human Factors in Computing Systems, 2001, p. 308.

[16] Max/MSP http://cycling74.com/products/maxmspjitter/.

[17] OpenCV (Open Source Computer Vision Library) http://opencv.willowgarage.com/

[18] EvoCam, http://www.evological.com/evocam.html.