# Speech & Multimodal Resources
# The Herme Database of Spontaneous Multimodal Human-Robot Dialogues

*Jing Guang Han, Emer Gilmartin, Celine De Looze, Brian Vaughan, and Nick Campbell*

Speech Communication Lab
Centre for Language & Communication Studies
Trinity College Dublin
Ireland
nick@tcd.ie

## Abstract

This paper presents methodologies and tools for language resource (LR) construction. It describes a database of interactive speech collected over a three-month period at the Science Gallery in Dublin, where visitors could take part in a conversation with a robot. The system collected samples of informal, chatty dialogue – normally difficult to capture under laboratory conditions for human-human dialogue, and particularly so for human-machine interaction. The conversations were based on a script followed by the robot consisting largely of social chat with some task-based elements. The interactions were audio-visually recorded using several cameras together with microphones. As part of the conversation the participants were asked to sign a consent form giving permission to use their data for human-machine interaction research. The multimodal corpus will be made available to interested researchers and the technology developed during the three-month exhibition is being extended for use in education and assisted-living applications.

**Keywords:** Speech communication, Social interaction, Multimodal robot platform, Speech corpus

## 1. Introduction

Interactive speech technology is now being implemented in many consumer devices and can already be considered a mature technology, particularly with the convergence of (a) increased use of wireless technology allowing mobile versions of many hitherto stationary devices, (b) migration of activity from pc's to mobile devices, and (c) the advent of mass-market applications such as Siri, TellMe, and Google Mobile interfaces (Feng et al., 2011).

Much work on interactive dialogue systems has concentrated on situations where the robot collects information from the interlocutor in order to complete a task (booking a flight, paying a bill), or answering queries by searching a local database or the web. The simple conjunction of speech recognition and speech synthesis might be sufficient for such limited domain dialogues, but it is probably inadequate for the implementation of a realistic discourse-based interface for human-machine interaction (Campbell et al., 2006; Douxchamps and Campbell, 2007).

Talk between humans communicates much more than simple linguistic propositional content. Dialogue involves information sensed through different input streams; including facial expression, gaze, and gesture recognition through the visual channel, in addition to the information and cues gathered from the audio channel. For more natural interaction in dialogue systems, successful management of the various input streams is vital. The system should take on an active listening and watching role.

Spoken interaction in humans is not only task-based - chat or socially motivated dialogue is a fundamental building block of human interaction. Successful modelling of chat will contribute to the implementation of systems where a 'friendly' user-machine relationship is important, as in robot companions, educational, or assisted living applications.

## 2. Managing Conversations

The ultimate aim of the work described here is to provide a synthesiser with feedback mechanisms whereby it can immediately monitor the response of a listener and adjust its output accordingly in real-time.

Non-verbal aspects of dialogue, including facial expression, eye-gaze, and prosody have been considered as important if not more important than propositional content in terms of meaning (Beattie, 1982), while recent work on Multiparty Interaction (SSPnet, 2011; AMI, 2011), and the Freetalk Multimodal Conversation Corpus project (FREETALK, 2011)) has shown that a camera can be as useful as a microphone in processing human spoken interaction. To model this multimodality we constructed a robot platform, providing the eyes and ears of the synthesiser, that is capable of observing the interlocutor throughout a conversation as per (Chapple, 1939; Kendon, 1990).

Dialogue is built on a framework of turntaking, allowing interacting participants to achieve task-based and social goals through sequences of adjacency pairs (Sacks et al., 1974; Clark, 1997). Automatic systems have been developed to predict possible points for turntaking in task-based dialogue, based on many features including prosody and propositional content (Bull and Aylett, 1998; Raux and Eskenazi, 2008). However, we are interested in exploring the optimal timing in a non-task-based or 'chat' setting, and designed a robot system to engage in a short, friendly exchange.

The robot, Herme, was built using LEGO Mindstorms NXT technology (LEGO, 2011), programmed using the NXT Python framework (nxt-python, 2011), and served as a mobile sensing platform for the camera. It was designed to adjust its orientation according to the location of the participant in order to always look at the person, thus providing the illusion of first selecting and then paying at-

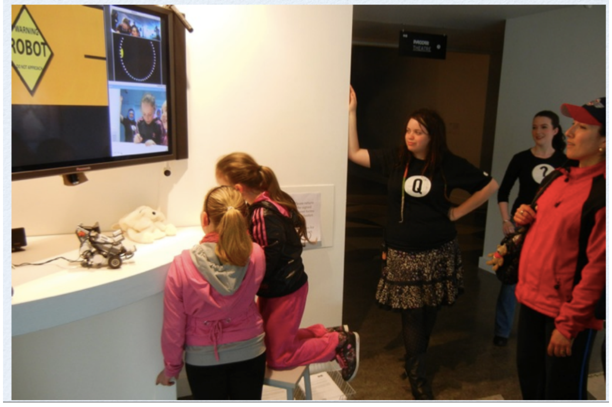

Figure 1: The LEGO NXT robot automated conversation platform, Herme, interacting with people and collecting multimodal conversational data in the Science Gallery at Trinity College Dublin

tention to one interlocutor. One computer controlled the robot, performed face tracking, sent and received data and displayed the robot's-eye view in real-time to participants, while another computer initiated a conversation and monitored participants' reactions in order to step through a predetermined sequence of utterances.

Two modalities were employed to collect data. The first used automatic sequencing, with the robot speaking-out each utterance in turn and waiting a predetermined time for the interlocutor's reply. In the second, manual sequencing was performed by a wizard who observed the interaction from a separate building using a skype connection. By studying the effect of different utterance-response timings in conjunction with different reaction types we were able to optimise the automatic sequencer, extend the time-to-failure or the success-rate of the conversations, and increase our data collection.

### 2.1. The Collection Platform

Herme was exhibited as a conversation robot on a waist-high platform in a corner of the Science Gallery in Dublin (see Fig. 1) as part of the HUMAN+ event from April 15 to June 24, 2011, (Science Gallery, 2011).

### 2.2. Making contact

As soon as a person walked into the field of view of the robot, the face recognition system triggered the start of a new conversation. The robot moved to centre the face in its field of view (using software based on OpenCV code (OpenCV, 2011)), displaying the face surrounded by a coloured circle on a large monitor above the exhibit, and simultaneously generating synthesised utterances in the pitch-shifted voice of a 'small person', saying "Hello?, Hi ....", followed by a repeat of "Hello", and then another "Hi", timed so as to maximise the probability of a response by the bystander/onlooker. Almost all visitors caught this way responded with a "Hello" or something similar before the robot emitted the second "Hi".

The art of making and keeping contact formed an essential part of the dialogue-based element of this research.

### 2.3. Talking with People

The main body of the subsequent interaction was carried out by way of the fixed-sequence dialogue shown in the appendix. The process of stepping through the utterances, and utterance groups, and waiting for the completion of the interlocutor's response was predetermined, but the timing was variable, and this formed the main point of implementation research in the development of the software. The dialogue was designed to contain an initial phase of friendly chat followed by a task - id number collection - and returning to chat and joke telling. We were lucky to hit upon an effective dialogue sequence very early on in the research and we gained much insight into the artificial maintenance of a conversation through monitoring people's responses to these utterances. The data we have gained in this way will serve as the basis for automating the monitoring of reactions in synthesised discourse.

Because of the extremely noisy environment in the Science Gallery (where 26 other high-tech exhibits were almost all emitting some form of loud and quasi-continuous noise) no attempt was made to incorporate any traditional form of speech recognition in the dialogue interface. Instead, the wizard watched the reaction of the interlocutor to each utterance, and waited "an appropriate amount of time" to make the next step through the dialogue sequence.

A successful strategy for dialogue was found in the use of short 'volleys' where the robot asked a question or made a statement followed by a related question, succeeded by a wait for the interlocutor to respond, with the robot then providing interjections of "really", "oh" or "why'?' to establish and maintain the illusion of attention and backchannelling. 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 episode was begun. By thus maintaining the initiative throughout the interaction, we were able to substitute 'polite listening' for any form of 'understanding' of the visitor's reply. Several participants commented on this dialogue architecture, with one typical comment likening it to "talking to someone at a party".The key element of processing here was in the timing of the utterance sequences. This was much better achieved by a human wizard, observing the people, than by our automatic systems using visual, audio, and motion-detecting sensors.

Participants were encouraged to stay and continue the conversation at several stages throughout the dialogue. Interjections such as "I like your hair!" surprised people, but all except one (a Muslim girl wearing a hijab perfectly concealing all of her hair) responded very positively — even the bald man who laughed back "I ain't got no 'air!". Similarly, "Do you know any good 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 (and often genuinely amused) listening to the robot's joke in turn. The robot's laugh was 'captivating' (Campbell, 2007).

### 2.4. Breaking off the conversation

Once the consent form had been signed, the id number spoken into the camera, and sufficient data on 'maintain-

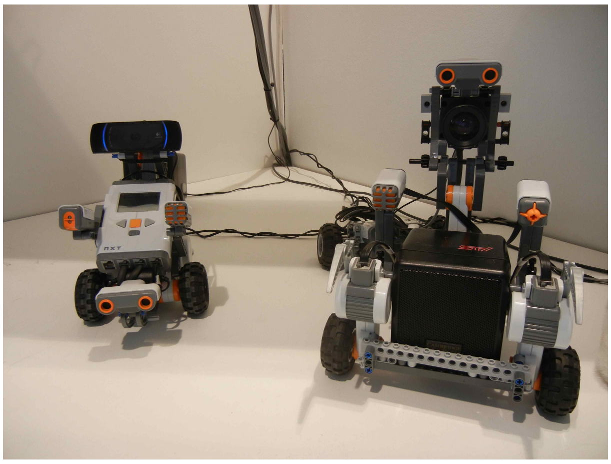

Figure 2: *Herme & Him, two robots (one female, one male) that observe people while they are talking to them. Built from LEGO, they support cameras and microphones for interactive speech synthesis*

ing contact' had been collected, the robot had just to end the conversation and get rid of the interlocutor in order to catch the next data provider. This step proved remarkably reliable as people readily take "thanks" as closure. Some might have been glad to get away at this point, but almost all engaged in the formal three-part closing sequence, similar to the three-part opening, repeating "goodbye", and "that's okay" (etc) to the robot's farewell greetings. A mark of the attraction of the interaction is the fact that several participants returned for another round, with one even coming back several days after her initial visit with a poem she had written for Herme.

## 3. Inside the Box

As shown in Figure 1, the capture environment consisted of a prominent corner booth in the Science Gallery, and was supervised by full-time gallery attendants who took care that participants understood the nature of the exhibit and that their interaction with the robot would be recorded.

There were two Sennheiser MKH60 P-48 shotgun microphones mounted at the top of the main screen (one is visible in the picture), alongside a Logitech C-910 HD webcam that provided a top-down overview of the interaction.

On the platform itself were two robots, one male and one female (see Figure 2), with the female engaging in interaction with the visitors while the male guided another Logitech HD Webcam to ensure that the interactions were recorded from a more inclusive angle. The main robot camera stayed zoomed-in to observe the face of the main interlocutor. Microphones on the webcams provided a close-up source of sound to be used in conjunction with that of the shotgun microphones. An i-Sight camera was mounted at the corner of the display to provide a wide overall view of the scene for the remote operator.

During the latter half of the exhibition we added a movement sensor to trigger onset and offset of the conversations as an additional control sensor. While we were able to monitor the vocal and gestural behaviour of the interlocutor, we were not able to automatically detect a switch of interlocutors if one walked away just as another came into the field of view.

The booth concealed three computers: two Mac-Minis for the robot and a Unix workstation which collected and stored the data and provided the skype interface for the wizard in the lab. Synthesis was made by the default APPLE synthesiser using Princess voice, shifted acoustically by a Roland Sound Canvas UA-100. The machines ran continuously, streaming all data to disk while the Gallery was open. Over the three-month period, the system crashed three times due to overheating.

In all, we collected 433 signed consent forms and 1.5 terabytes of recordings from more than a thousand conversations. The data consist of recordings from the Herme-eye camera, the Him's-eye camera, the oversight camera, the i-Sight device, and four microphones. 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.

## 4. Summary & Future Work

This work describes methodologies and tools for the extraction of data and acquisition of knowledge related to spoken interaction. It presents a novel interface for speech-based dialogue systems, for capturing natural language and multimodal/multisensorial interactions using voice activated and movement-sensitive sensors in conjunction with a speech synthesiser.

Several organisational, economical, ethical and legal issues were addressed. Specifically, we presented a low-cost solution for the collection of massive amounts of real-world data with full approval of the Faculty Research Ethics Committee of the University. Participants were not paid and walked in off the street voluntarily. All age groups and social classes were included.

The combination of entirely voluntary participation taking place in a leisure or fun setting, the varied cohort of participants and the design of the dialogue have facilitated the collection of a corpus containing a significant portion of informal chat between human and machine – a very important dialogue modality but very difficult to capture under laboratory conditions.

This resource is available for collaborative research related to LRs in interactive dialogue systems and applications in the expanding range of fields already using or currently introducing speech technology - including data management (information extraction and retrieval; audio-visual, text, and multimedia search; speech and meeting transcription), education (Computer Aided Language Learning; online and computer-aided training and education), assistive technologies and AAC, and localisation (machine and speech translation, interpreting).

We are already collaborating with researchers in several other European universities on the analysis and processing of this data, and intend to eventually provide an open, linked and shared repository, with tools and associated software.

## 5. Acknowledgements

This work is being carried out at TCD with thanks to Science Foundation Ireland (SFI Stokes Professorship Award 07/SK/I1218) and PI (FastNet project) grant 09/IN.1/I2631. We are grateful to the Long Room Hub at Trinity for inviting us to take part in the Science Gallery HUMAN+ exhibition.

## 6. References

AMI. 2011. The augmented multi-party interaction project. http://www.amiproject.org/.

G.W. Beattie. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica*, 39(1-2):93–114.

M. Bull and M. Aylett. 1998. An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *Fifth International Conference on Spoken Language Processing*.

N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps. 2006. A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow. In *Proc Language Resources Evaluation Conference*.

N. Campbell. 2007. On the use of nonverbal speech sounds in human communication. In *Proceedings of the 2007 COST Action 2102 International Conference on Verbal and Nonverbal Communication Behaviours*, pages 117–128.

E.D. Chapple. 1939. Quantitative analysis of the interaction of individuals. *Proceedings of the National Academy of Sciences of the United States of America*, 25(2):58.

H.H. Clark. 1997. *Using Language*. Cambridge University Press. Cambridge.

D. Douxchamps and N. Campbell. 2007. Robust real time face tracking for the analysis of human behaviour. In *Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction*, pages 1–10.

J. Feng, M. Johnston, and S. Bangalore. 2011. Speech and multimodal interaction in mobile search. *Signal Processing Magazine, IEEE*, 28(4):40–49.

FREETALK. 2011. Freetalk multimodal corpus of conversational speech. http://freetalk-db.sspnet.eu/.

A. Kendon. 1990. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press.

LEGO. 2011. MINDSTORMS NXT an intelligent microcomputer brick. http://mindstorms.lego.com.

nxt-python. 2011. A pure-python driver/interface/wrapper for the LEGO Mindstorms NXT robot. http://code.google.com/p/nxt-python//.

OpenCV. 2011. Open source computer vision library. http://opencv.willowgarage.com/.

A. Raux and M. Eskenazi. 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. *Proceedings of SIGdial 2008*.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Science Gallery. 2011. HUMAN+ the future of our species. http://www.sciencegallery.com/humanplus.

SSPnet. 2011. The social signal processing site is hosting freetalk. http://www.sspnet.eu/.

## 7. Appendix - the dialogue script

```
- hello?  hi . . .
hello . .
hi

- my name is hermee - h e r m e - hermee
what's 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
here in TCD - tell me about you . . .
- really?
oh

- tell me something else
oh
really

- why are you here today?
really?
why

- 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
```