An Interactive Speech System for a Robotic Exhibit

Shaun Press　　　　**Jason Chen**

shaun.press@anu.edu.au　jason.chen@anu.edu.au

Robotic Systems Lab, RSISE
College of Engineering and Computer Science
Australian National University
Canberra, Australia

Abstract

In this paper we look at providing a speech recognition and response system that performs the dual role of providing locality specific information as requested, while also engaging in general conversation when appropriate. The aim of this system is to create a softer and more entertaining interface to existing information kiosks, while maintaining their usefulness as a provider of information.

1 Introduction

Increasingly robots are beginning to appear in public places. One of the most common areas where humans

Figure 1: Questacon Robot with Active Head

and robots come in contact with each other is in tourist related areas [Schulte *et al.*, 1999]. Traditionally the purpose of such robots is as an information kiosk, where users interaction is restricted to requesting, or simply being given, specific information from the robot[Thrun *et al.*, 1999]. In this type of environment a terse robot response is often unattractive, and we propose that softening the interface to be more entertaining and fun, can be advantageous.

In this paper we report on a robot being constructed jointly by the Australian National University and Questacon. Questacon is the Science and Technology Museum located in Canberra, Australia, and its role is "To achieve its aim of promoting greater understanding and awareness of science and technology within the community, Questacon is committed to making an experience with science and technology fun, educational and interactive. Questacon has over 200 exhibits and approximately 300 000 people visit the Centre in Canberra each year." [1].

The robot being designed will be situated in the entry foyer of Questacon's exhibition centre in Canberra and will act as an exhibit to the public designed to both assist, and to demonstrate practical robotic technology. One of Questacon's requirements with regard to assisting visitors is that the interface should be able to give out Questacon specific information such as, what exhibits are interesting to see, where they are, where are the toilets, what times are the shows on, etc. However, they would also like the robot to act like an exhibit itself, and engage in an informal dialogue that is not Questacon specific. That is, the aim of the system is therefore to move seamlessly between a "hard" information as well as providing a "soft" conversational tool, without the user being made aware of the change between two. As the major purpose of the system is to provide specific museum related information, the system nonetheless will have a stronger preference for the former over the latter.

---

[1] From "What's a Questacon?" http://www.questacon.edu.au

In this paper, we report on the progress of a system that we have constructed to achieve this aim.

1.1 Related Work

The overall project draws upon previous work in the field of "Tour-Guide Robots". The RHINO project [Burgard et al., 1998] demonstrated the feasibility of placing robots inside museums, although the system lacked a "spoken dialog interface". An office based robot Jijo-2 [Matsui et al., 1999] moved this idea along, allowing office workers to interact with Jijo and to extract information relevant to the environment it operated in.

Another area of development related to this project is in the area of portable conversational agents. The InCA project [Kadous and Sammut, 2004] implemented a portable (via PDA) speech recognition and response system. Although the system utilised both a portable input/response device and a non-mobile server, the added size of a mobile robotic platform would allow the whole system to become portable.

Nonetheless the aim of our research is to extend the environment that the conversational system operates in. Most of the work cited achieves good results by restricting the domain of the conversation. We are aiming at developing a system that covers a broader conversational area, while recognising the reduced level of performance of the system overall.

2 Questacon Robotic Project

The hardware platform for this paper is the Questacon Robot (see Figure 1). The robots internal structure was originally designed by past lab member, David Austin, and was constructed by Questacon. The exterior (shown as an artistic impression in Figure 2) was designed and is currently being constructed by Questacon. The robot has a large number of sensors, including Laser Scanner, Infra-red sensors, bump sensors, active stereo vision head, microphones. For the first stage of the robots implementation, the robot will serve as a demonstration of various robot technologies. These technologies include speech recognition and response, face recognition and tracking, object detection, and localisation and mapping. For the robot to successfully demonstrate these technologies, it must engage the user in an interactive manner.

The robot will interact with users using 3 different systems. The first method is via the active head face tracking system. This system will select a face from a set of people standing in front of the cameras, and will follow the face using a combination of face recognition and skin colour tracking techniques. This interaction will last as long as the user is looking at the robot.

The second system is via a speech recognition and response system. Users will be able to speak to the robot and receive information in spoken form. This system uses a combination of grammar based recognition systems, as well as a general dictation recognition mode. Using both these system allows the robot to provide specific information upon request, as well as engaging the user via more general conversation.

The third system is a touch screen system mounted on the front of the robot. The touch screen will provide the user with various pieces of information concerning the operating state of the robot, as well as allowing the user to control which information is being presented.

The speech recognition and response system described in this paper runs on one of the three control computers inside the robot. The control computer is a standard Intel PC motherboard. Speech input and output can either be handled by the built in sound card, or via higher quality sound cards, as needed.

The system in its current state uses a standard and inexpensive headset microphone. While this is suitable for testing the system in a quiet environment, it is unsuitable for the noisy environment the robot will be eventually be operating in. As a consequence a more sophisti-
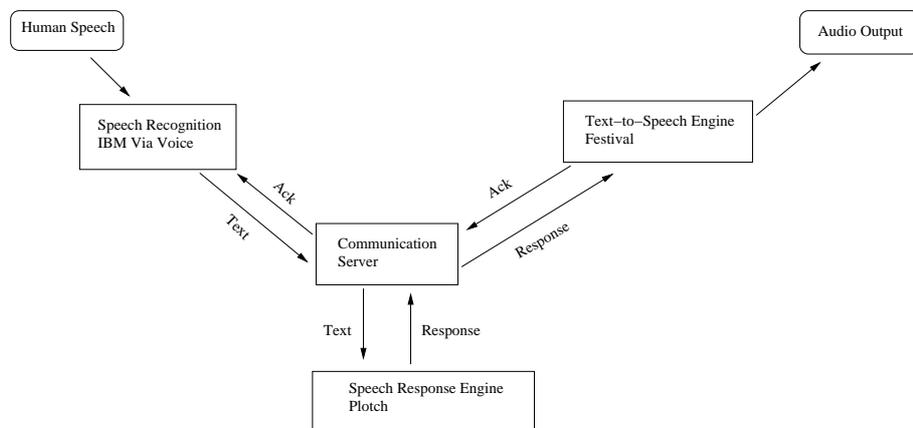

Figure 2: Questacon Robot - Artists Impression

Figure 3: Architecture of Speech System

cated microphone array is being constructed for future use.

3 Software

This section describes the software utilised in the system. Figure 3 presents the four components (each represented by a square box) and their relationship when operating. The four components are

- Speech Recognition Module
- Response Engine
- Speech Synthesis Module
- Communication Server

In overview, the operation is as follows. Audio picked up by the microphones is passed to the Speech Recognition Module which outputs text to the Response Engine. The Response Engine crafts a response and passes it on to the Speech Synthesis Module, which converts it to audible speech. The Speech Synthesis Module then advises the Speech Recognition Module that it has finished, so that the Recognition Module knows it can start to decode the human users response. At all steps, the Communication Server sits between each of the other modules and manages the communication. We note that the speech system was designed to utilise as much open-source software as possible. The speech response engine and the communication manager are both written in Python, while Festival [Taylor *et al.*, 1998] is used as the text-to-speech engine. The only piece of proprietary software in the system is the IBM Via-Voice speech recognition system. The choice of this software was motivated by the provision of a software development tool kit which allowed us to handle specific events generated by different speech recognition states.

```
Where is the cafe?
Where is the shop?
What time do you close?
Are you a robot?
What is your name?
```

Figure 4: An typical set of questions

3.1 Speech Recognition

Speech Recognition is done using the IBM Via Voice Engine. IBM Via Voice supports the two distinct modes, both of which are required for the implementation in our system. To recognise specific information requests, the "Grammar Recognition" mode is utilised. A BNF (Backus Naur Form) style grammar specifies the sentences to be recognised. An example of typical questions can be seen in figure 4 and the subsequent grammar can be seen in 5. For all other speech, general dictation mode is used. In general dictation mode, Via Voice accepts freely spoken text, dynamically decoding groups of words using Hidden Markov Model technology.

```
<kiosk> = What <whatq> | Where <whereq>
| Are <areq> .
<whatq> = time do you close | is your
name .
<whereq> = is the <location> .
<location> = cafe | shop .
<areq> = you a robot .
```

Figure 5: A BNF Grammar for questions in figure 4

The technical specifications of the Via Voice system show that phrases included in the grammar file have a higher level of recognition than expected phrases in gen-

```
robot:bot:robots:are you a machine:are you real:
6
*I am quite human
*I am not a robot
*A robot. Where?
*How dare you question my realism
```

Figure 6: An example entry from the response file

eral dictation. This is due to Via Voice's ability to attach a higher probability to words that it expects to be spoken by the user. While the dictation mode has some degree of predictive expectation (due to user training and document modelling) it does not have the same level of success as grammar file based recognition. Nonetheless, the success in recognising phrases defined in the grammar file begins to fall as the size of the grammar file increases. Based on the technical specifications of Via Voice, a maximum of 500 grammar based phrases appears to be sensible. This restriction, for example, means that it is not possible to take advantage of the higher recognition rates for grammar file based recognition for general conversation recognition.

3.2 Response Engine

The robotic speech system passes the decoded speech to a response engine. The response engine determines a suitable response based on the weighting of keywords in the speech. The response engine is based on an Eliza [Weizenbaum, 1966] style program called Splotch (author Duane Fields[2]). For the implementation of the system on the Questacon Robot, the Splotch engine was rewritten in Python to provide specific functionality required by the system. The response engine parses the speech input, scanning its dictionary file looking for keywords. An example of a dictionary entry dealing with questions of the robots identity can be seen in figure 6 When keywords are found in the input text, a weighting is attached to each word. Higher weightings are attached to nouns than prepositions, while nouns and phrases relating specifically to Questacon are giving the strongest weightings of all. The response engine supports a small variety of linguistic constructions. Apart from single word recognition, it accepts short phrases, as well as the logical construction of words (not, or, and). It also accepts phrases containing wild-cards (eg "I like to ... ") allowing the engine to construct responses using the wild-cards (eg "I have a friend who likes to ... as well").

When the dictionary search is complete a response is constructed relating to the keyword or phrase with the highest weighting. Each recognised keyword of phrase can have more than one response. The response engine will choose one of these responses at random and return the response to the system.

The response engine can also construct responses based on recognising wild card inputs. Words can be recognised as being part of a specific phrase, without the meaning of the word being significant. eg the phrase "I can ..." will be recognised even if the sentence continues with "run","jump","sneeze" etc. In constructing a response the response engine uses the phrase continuation in the reply. The constructed reply could then be "Can you teach me how to" "run","jump","sneeze" etc

3.3 Text to Speech

When the Plotch response engine has constructed a suitable phrase it is passed to the text to speech engine. For this system we use the Festival system. The Festival system accepts plain text and converts it into human sounding speech in the form of a sound file. All of the technical details of this part of the system can be found in "The architecture of the the Festival speech synthesis system" [Taylor et al., 1998].

3.4 Communication Server

The synchronisation of the software modules is a significant issue in the system, and this is undertaken by the Communication Server. The Communication Server is written in Python and accepts a network connection from each of the software modules. There is a strict ordering of message transfers, required to keep the system synchronised. (See Figure 3). The message ordering is circular in nature, and is designed to implement a state machine that controls the action of the system. The normal state for all components is where they are waiting for input. For the voice recognition system this is input via the microphone, while for the response engine and Festival TTS it is text input via the network connection to the communication module. When speech has been received and recognised, the voice recognition engine switches the microphone off [3]. The speech recognition engine then waits for a message from the communication server indicating that Festival has finished speaking, and that it is "safe" to turn the microphone back on.

4 Experiments

The system is still work in progress and we wish to show with the experiments in this section its potential. Questacon is responsible for constructing content in the system, and the following experiments were conducted with

---

[2]Written while a student at Texas A&M University. http://splotch.sourceforge.net

[3]The system needs to switch between listening mode (with the microphone switched on) and response mode (with the microphone switched off). This is to avoid the obvious problem of the spoken response by the robot being interpreted by the robot as speech from the human.

Where can I get the tickets?
(A) The tickets are available at the counter to your right
(B) Tickets are 15 dollars for adults and 10 dollars for children
(C) I prefer baseball to cricket

Figure 7: 3 possible responses to an information request

a first draft of the content they wish to include in the final system (specifically the content in the grammar file in the Speech Recognition Module and the dictionary file in the Response Engine).

To see how effective the system is, testing of the two major features is required. Firstly, the system should be able to provide accurate and specific information when requested by the user. Secondly, when the user is engaging in casual conversation the system should provide a response that satisfies the users belief that the robot is interacting with them. Specifically, in our system, we try to pick out a keyword, and then craft a response which includes that keyword. In testing both these system we are also interested demonstrating that the first outcome (providing specific information) occurs in preference to the second outcome.

All experiments were carried out using a single microphone input. The Via Voice speech recognition engine had received minimal training with regards to the testers voice. For the purpose of testing we classified the input voice as "Generic Adult Male". Future testing will involve voice models for "Generic Adult Female", "Child Male" and "Child Female".

4.1 Information Requests

The robot was asked 50 questions that requested specific information. That is, these question were entries in the grammar file. The responses were then categorised into 3 different groups: (A) Responses that answered the question asked, (B) Responses that addressed the topic of the question (i.e. keyword), without providing the specific information requested, and (C) Responses that did not provide sensible information concerning the question asked.

An example of a question from the grammar file, with its 3 possible answers from the response dictionary file, can be seen in Figure 7

Specific Information Request Results

Specific Information	On Topic	No Information
50%	32%	18%

While only 50% of the responses provided the specific information requested, 82% of the responses could be

Do you like dinosaurs?
(A) The dinosaur exhibition is in gallery 3
(B) I love dinosaurs
(C) I do not understand the question

Figure 8: 3 possible responses to a conversational question

considered as encouraging continued conversation. The cause of the remaining "No Information" responses was twofold. The majority of the response failure was due to the speech recognition software failing the interpret the question clearly. The remaining failures were caused by missing entries in the response dictionary. In both cases, the robot would respond with a "pardon", or "I don't understand".

4.2 General Conversation

Two distinct tests were carried out in measuring the general conversational abilities of the robot. The first was to simply ask 50 "free form" style questions, i.e. questions that were not in the grammar file. Again the responses (see Figure 8) were classified as in the information requests tests.

Information requests in general conversation

Specific Information	On Topic	No Information
34%	30%	36%

Once the speech recognition software moved from grammar based recognition to general dictation mode, the responses providing the specific information requested fell. This was matched with a corresponding rise in the responses that provided no or meaningless information. This change can be attributed to the drop in successful recognition of speech by Via Voice.

The second test was to see how long the conversation would last. Both grammar based and general conversational phrases were used. For testing purposes the conversation was considered to last until the robot made two successive responses that did not make sense. The allowed the user to keep the conversation going if it did not understand a single response.

The test was repeated 50 times with the following results.

Average Length	Shortest	Longest
6.5	0	15

The average number of sentences understood and responded to in a meaningful way, demonstrates that the robot has enough conversational ability to engage a human for at least 2 minutes. While this may seem brief for the moment, it both fulfils Questacon's requirement for the project, and shows that the system can exploit

improvements in speech recognition technology. The results will be strengthened further when Questacon refine the content of the response dictionary, to handle phrases that Via Voice correctly decoded, but where no response was currently found. To provide a flavour of the type conversation that occurred in this experiment, we show below one of the dialogues that occurred.

```
Hello
Hi, how can I help you?
What is your name?
My full name is Jack Questacon
How old are you?
Pardon?
How old are you?
I don't know
Are you real?
I never have been
Who made you?
Not me
Where am I?
You are where you are
Who built you?
Your mother
What about my mother?
How old is your mother?
My mother is 72
Is your mum here?
No
Why not?
Because she did not want to come
Excuse me
You are silly
Excuse me
```

Figure 9: An example conversation

5 Conclusion

This article demonstrated a working method of adding a speech response system to a robotic exhibit. We were able to craft together a number of low cost, existing technologies to create an useful and entertaining human-robot speech interface. The results in this paper show that such a system is suitable for allowing robots to engage with humans in a practical way. As improvements in the area of speech recognition and natural language understanding are found, the framework used in this paper will generate even more robust and successful systems.

References

[Burgard et al., 1998] Wolfram Burgard, Armin B. Cremers, Dieter Fox, Dirk Hahnel, Gerhard Lake-meyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. The interactive museum tour-guide robot. In AAAI/IAAI, pages 11–18, 1998.

[Kadous and Sammut, 2004] M. Kadous and C. Sammut. Inca: A mobile conversational agent, 2004.

[Matsui et al., 1999] Toshihiro Matsui, Hideki Asoh, John Fry, Youichi Motomura, Futoshi Asano, Takio Kurita, Isao Hara, and Nobuyuki Otsu. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services. In AAAI/IAAI, pages 621–627, 1999.

[Schulte et al., 1999] J. Schulte, C. Rosenberg, and S. Thrun. Spontaneous short-term interaction with mobile robots in public places, 1999.

[Taylor et al., 1998] P. Taylor, A. Black, and R. Caley. The architecture of the the festival speech synthesis system, 1998.

[Thrun et al., 1999] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, J. Schulte, and D. Schulz. Minerva: A second-generation museum tour-guide robot, 1999.

[Weizenbaum, 1966] J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine, 1966.