Proceedings of the Institute of Acoustics

A SPEECH DATABASE COLLECTION ARCHITECTURE

G P Walker

British Telecom Research Laboratories, Martlesham Heath, Suffolk, U.K.

1. INTRODUCTION

British Telecom Research & Technology has been developing speech recognition algorithms for products and services for some time. Corpora of both isolated and connected speech have been collected to aid this development and many techniques have been tried to ensure that the collected speech is as natural as possible and that it has been gathered with efficiency.

The aims in designing a collection system were as follows: the collection must be automated on a PC and the dialogue easily altered. Since both "clean" and telephone quality speech have been required in the past, such a system would need to have the flexibility to use different interface cards. Prompting would need to be verbal or visual. The data would need to be copied to an optical disc for transferral. The desired flexibility of such a system suggested that, once the hardware configuration had been defined, the software should remain static whilst allowing dialogues to be modified easily.

The resulting architecture is a robust data collection system, capable of the above, but easy to drive. This paper will describe the architecture and give examples of its use. It will also discuss the talker selection, vocabulary, dialogue and the psychology behind dialogue design.

2. THE DATA COLLECTION SUITE

In order to formalise and automate the collection of speech databases, we have designed and developed a suite of modules and libraries which can be linked together to form a collection program capable of recording speech under various recording methods using various prompting methods.

Collections made prior to the inception of this architecture were made with ad hoc programs where the dialogue was embedded into the code and so any change to the dialogue necessitated a change to the code and subsequent re-compilation. The development of the new architecture was driven by the goal of a dialogue-independent program, removing the need to write code for each collection. The flexibility of choice of prompt methods and collection devices was also built in.

The modules are written in such a way that other collection devices and prompt methods may be added at a later date, making the system expandable. Each includes strong error handling routines to avoid program hangs or crashes. The program allows easy entry of parameters which affect the recordings (e.g. duration and timeout) via parameter-entry screens and information screens, each with context-dependent help information.

The dialogue system is a novel approach to speech data collection. It allows a very quick change to be made to a dialogue, using a simple text editor. (This is described in more detail in section 5.)

Having one collection suite for all collections ensures that the file format, including header content, remains compatible across all collections.

The target machine was an IBM-PC AT or compatible with enough local disc space to record at least one talker's recording session. Transferral of the data from the machine was facilitated with the use of optical discs. (The backup is also under the control of the program and can be automated to run at a particular time of day, if the equipment is connected for long periods. Data is removed from the local disc if the transferral is successful.)

3. COLLECTION DEVICES

Two PC interface cards have been used for data collection; one for high quality wide-band speech and one for telephone-band speech. Each card has a related software library for interfacing to MS-DOS via the "C" programming language. Modules for each of the cards have been written for the architecture so that they are interchangeable depending on the card being used. To ensure complete compatibility, if a particular card cannot perform certain functions the modules associated with that function are empty.

When the collection device has been decided, the various relevant modules and libraries are linked with the main collection code.

3.1. Data Translation DT2823 I/O Card

This card is one in a family of high speed analogue and digital I/O cards[1]. For the purposes of this paper, only those details relevant to the architecture will be given. The DT2823 is a PC expansion card capable of up to 130 kHz sampling rate at 16 bits resolution. It also has two digital i/o ports of 8 bits. Data Translation's ATLAB subroutine library is used to control the card functions.

Wide-band speech is normally sampled at 20 kHz on a single channel, whilst the digital input port is used to control the program. When the system is used under a semi-automatic mode, a person monitoring the collection can determine whether a phrase or word needs to be repeated. This is done via the digital port, with a hand-held device, avoiding any necessity to use the keyboard.

Speech is written to disc at DMA speeds to enable continuous performance.

3.2. British Telecom Speech Card

The British Telecom Speech Card (BTSC) is a multi-purpose PC expansion card designed to allow advanced interactive speech services to be carried over the telephone network. It has the following features:

- Full telephone line interface facilities including automatic call initiation, auto-answering and tone recognition
- Speaker-independent recognition of up to 50 words in real time

- Speaker-dependent recognition of up to 400 words in real time
- Waveform encoded speech storage and retrieval at 16 or 64 kBits/sec
- Buffered speech i/o lines under software control, to monitor a telephone line or to record and play back messages outside the telephone environment.

The BTSC, like the DT2823, comes with a subroutine library to enable high-level program control.

4. PROMPTING METHODS

We have used two types of prompt methods: visual prompts, where the word or phrase is displayed on a terminal screen (VDU), and verbal prompts. Verbal prompts take two forms:- the talker may be requested to say a particular word such as

Please say "ONE"
Please say "NINE" etc.

or may be requested to refer to a sheet of paper (dialogue sheet)

Please say the first sequence on your sheet
And the next sequence etc.

Which form of prompting is to be used depends on the type of collection (i.e. what kind of speech quality) and how it is to be run (i.e. monitored or automatic.) Generally for wide-band high-quality speech recordings we would use head-mounted microphones in an acoustic booth with a visual display prompt. Telephone recordings tend to use verbal prompts.

The first verbal method shown above avoids the necessity for a dialogue sheet but there is a high possibility of introducing mimicry. The second method encourages the talkers to speak at their own rates and in their own accents but, having the words in front of the speaker allows the speaker to 'talk ahead' and denies the possibility of randomisation of word presentation and recording. 'Talking ahead' is a problem discussed in the next section.

Once again, software modules to drive the chosen type of prompting method are linked with the relevant libraries.

5. THE DIALOGUE SYSTEM AND DIALOGUE DESIGN

The method of altering the dialogue without altering the code is the greatest strength of this architecture. Once the hardware configuration has been defined, the program does not need to be changed from one collection to another. This is particularly important during the development stage of a collection when the interaction effects of dialogue prompts and talker utterances may cause problems. (See section 5.2)

5.1. The dialogue system

Our experience shows that small changes are often made to the vocabulary content after the initial pilot collection, either through interaction effects or omissions. Since the dialogue is expressed in a simple text file which can be edited with any text editor, altering the dialogue is a very simple matter.

For example, the dialogue file for a verbal prompt dialogue may take the form:

```
1 PF01
3 PF02 0001 3.0 1.0
3 PF04 0002 3.0 1.0
3 PF02 0012 3.0 1.0
5 PF10
```

In each line, the first number indicates what sort of prompt is required. 1 means prompt only (e.g. play a welcome message.) 3 means prompt and then record. 5 means prompt only but also indicates that line-noise should be recorded before the end of the session.

The next entry (e.g. PF01) is the name of a speech file to be used as the prompt. If a 'prompt and record' line is encountered, the next entry (e.g. 0001) is the number of the file in which to record the talker's utterance. This would result in a filename W0001.DAT being created for instance. The remaining fields indicate the speech onset and recording time and trailing noise (post-speech) time respectively.

Any type 3 entries, between type 1 and 5 entries, may be randomised (under program control) if necessary. Other categories have been defined, but are not listed here, which cope with all other possible recording variations. The above example gives a general idea. In a visual prompt dialogue, the filenames are replaced by the phrase text to be displayed.

5.2. Dialogue design and tuning

The first consideration in getting anyone to take part in a collection should be the recording session time. Sessions can be expected to last longer in an acoustic booth than over a telephone call, even when the call is free. So, much larger vocabularies can be recorded per session in acoustic booths. Also, it is much harder to keep the talker's attention during telephone collections and this is where dialogue design is important.

Generally, with acoustic booth collections, it is a good idea to group the words and phrases into categories - for instance, all digits should be recorded in the same section, groups of digits in another, phrases in another and so on. However, the opposite seems to be the case with telephone collections. The talker remains much more attentive if the groups are much smaller or even completely random.

Tuning the dialogue to overcome any undesirable effects is part of the development process, especially when dialogue sheets are used. Such effects may be caused by page turns (building pauses into the dialogue file overcomes this) or the talker losing his/her place (overcome by ensuring that prompts are sufficiently different to their near-neighbours.) A particularly difficult problem to overcome is that of 'talking ahead', i.e. the talker speaking before the prompt has finished. With speech detection, the problem can be overcome by asking the talker to repeat the last phrase or word. Without it, however, prompts need to be carefully recorded and endpointed to remove any trailing silence. If all else fails, part of the end of the prompt may have to be removed! This does not have to sound as bad as it seems.

For instance, if the prompt was

Please say the next sequence...
or And the next...

then part of the "-ence" or even all of the t sound in "next" can be removed without the talker realising, especially if the recording is being made over a band-limited system like the telephone network.

6. DEFINING THE POPULATION AND VOCABULARY

(Whilst this does not have a direct influence on the collection architecture, it is important to consider the implications on dialogue design. It also helps to show the flexibility of the system.)

Defining the population and vocabulary is largely a matter for the customer, that is, the recognition research team. Indeed the team is responsible for the definition of the vocabulary, its order of presentation and the talker population.

6.1. The Talker Population

The larger the corpus, the more varied should be the population. Consideration should be given to the ranges of age, accent and sex. Market research companies are excellent starting places for getting the right population. If the population requirement is small, colleagues, their relatives and friends may be persuaded into donating their speech to science! This can be very useful if the same talkers are required to add words to an existing corpus.

6.2. The Vocabulary

As one would expect, British Telecom Research and Technology has concentrated on isolated and connected digit recognition research. Some algorithm development requires multiple repetitions of each word or phrase and the dialogue can be arranged to cater for this. In the case of large vocabularies, each session may only include part of the entire vocabulary. Different dialogue files can be used and switched within collections if required.

7. CASE STUDIES

Two large databases have recently been recorded with the described architecture. One of them, concentrating on isolated words, includes the speech of more than 10,000 telephone calls. The other, for a study in connected digit and spelt words, includes more than 600 calls.

The first, called the "Next" database after the afore-mentioned problem with prompt endings, had a 70 word/phrase vocabulary and ran over 3 pages of a dialogue sheet. The sheet was sent to the talker, together with a letter explaining the collection and a telephone number to call. In order to collect as many talkers in as short a time as possible, eight PCs were used, each connected to a telephone line. The same collection program, configuration and dialogue were installed on each PC. The final dialogue was reached after asking various colleagues to comment on the format and content of the vocabulary. The ability to change the dialogue immediately was a great help.

The telephone exchange which the lines came from was able to direct the incoming calls to any available machine so that the amount of "line engaged" calls was kept to a minimum.

This database is, we believe, the largest speech database in the world, based on the number of talkers. The 10,000+ calls were taken from 45 different linguistic regional areas around England, Wales, Scotland and Northern Ireland in a successful attempt to cover a large number of accents. A wide age range across both sexes is represented. This database has been used to enable improvements in recognition accuracy for telephone network services. It takes up approximately 12 Gbytes of storage space.

The second, called "Trinity" because of the digit triples being recorded, had a much larger vocabulary, split into three sessions. Each talker was only required to complete one session, however, resulting in a telephone call of about 4 minutes. Three telephone numbers were given, one for each dialogue, one machine per line. Although this meant occasional "line engaged" problems, the market research company were instructed to release a certain number of dialogue sheets per week, to avoid the problem, together with a letter as above.

The size of the vocabulary was due to the combinations of all connected digit occurrences. All possible combinations of each digit preceded or succeeded by every other digit were required, though not by every talker. This database is being used within a research project on connected digit recognition. It takes up approximately 730 Mbytes of storage space.

In both cases, the same program was used but with different vocabularies. Since automatic backup is an option included in the suite, the equipment was left running (apart from one thunder storm that interrupted the power!) for the duration of the collections.

8. CONCLUDING REMARKS

The speech data collection architecture.described in this paper is a robust yet flexible tool for use in large or small data collection exercises. It has proved its worth in the two collections described. Such a system takes away the recoding required by some collection systems, yet is capable of being used for telephone quality and high quality speech.

9. REFERENCES

[1] Data Translation Inc., '1990 Data Acquisition Handbook'