Proceedings of the Institute of Acoustics

SUBSCRIBER — A PHONETICALLY ANNOTATED TELEPHONY DATABASE

A.D. Simons (1) & K. Edwards (2)

(1) BT Laboratories, Martlesham Heath, Ipswich.
(2) Centre for Speech Technology Research, Edinburgh.

1. INTRODUCTION

This paper describes the collection and processing of the 'Subscriber' database which has been collected over the telephone network by BT Laboratories. The database consists of utterances from over 1000 talkers from throughout the British Isles who were selected as a demographically balanced sample of the adult population. Five phonetically rich sentences from each talker have been annotated to the sub-phonetic level to allow sub-word unit recognition experiments to take place.

Speaker independent speech recognition over the British telephone network using sub-word units is an area of research which has received little attention to date. A large amount of training data is needed in order to be able to model the units sufficiently well, particularly when contextual effects are modelled. If the different accents of English are to be included then the number of sub-word units needed is increased. When differences in handsets and line conditions are also considered, it can be appreciated that an extremely large amount of annotated speech data is needed in order to produce a speech recognition system which performs well in all circumstances.

The Subscriber database was collected using a fully automated procedure which meant that it was important that the dialogue was simple enough to be completed with no human intervention, otherwise calls would simply fail. Prior to the database collection it was not known how well individuals would respond to this type of task, as the dialogue was completely different to those used in previous database collections at BTL. In order to assess this, a pilot database (called the 'Bader' database) of around 100 talkers was collected.

This paper discusses the design of the dialogue, and the changes made in the light of the pilot study. Decisions made in targeting the participants are discussed and a profile of the database in terms of regions/accent groups, age ranges and sex is given. Statistics on the response rate, and the quality of the responses are also presented.

The penultimate section shows a comparison between the Subscriber database and other speech databases currently used by the speech community, and the final section gives a summary of the lessons we have learned during this exercise.

2. DATABASE COLLECTION

2.1. Collection procedure

The database was collected using a fully automated procedure as described in [3]. When the call is made the caller listens to an introduction which includes an example of what the dialogue should sound like. The caller is then asked whether they are ready to start, and are expected to answer 'yes' to this question. The collection software does not attempt to perform any speech recognition, the response given by the user is simply stored onto disk. However there are algorithms to detect whether the user spoke too soon, spoke too late, spoke at all and to stop recording once the caller has finished speaking. If a caller is asked to repeat three times due to error conditions, then the call is terminated to avoid caller frustration.

2.2. Subject Selection

The subjects were selected to give a representative sample across age, sex and region. A market research company was used to recruit the callers, as we had found this to be an effective method for subject selection in previous database collections. It was requested that the callers were local people, to enable some conclusions to be drawn about their possible accent groups. Each caller was given a detailed sheet of instructions and a dialogue sheet showing the required responses to each prompt. The calls were made on the standard Public Switched Telephone Network. Callers were reimbursed for the cost of the call prior to making the call, and on completion they were paid a small sum as a 'thankyou' for taking part. For each potential caller, their name, sex, age and location were recorded by the market researcher.

2.3. Session content

The main purpose of the database collection was to collect phonetically rich sentences for annotation to allow sub-word unit speech recognition. Due to the fact that these sentences were relatively long, it was decided to keep the call dialogue short.

Each caller was asked to give a call reference number. This allowed identification of the caller for payment, and association of the expected responses for each call. There were two questions in the dialogue, one to which the answer 'yes' was expected and one to which the answer 'no' was expected. These enabled the collection of natural rather than read yes/no responses. The caller was asked to give two telephone numbers, one of which was read from the prompt sheet, and one of which was a familiar number. Each caller read seven sentences two of which were accent diagnostic [1] and five of which were selected from the 200 'ATR' sentences designed at CSTR, these are the same sentences as those used in the 'Scribe' database collection. The calls also included five spoken and naturally spelled surnames, and a serial number.

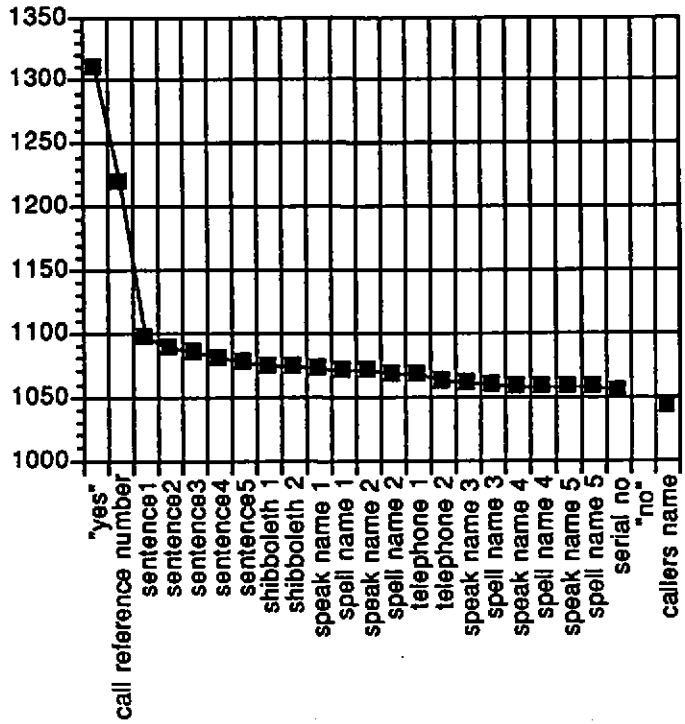
For the pilot study, there was only one accent diagnostic 'Shibboleth' sentence, which was rather long and was the first sentence required by the dialogue. It was discovered that people were having problems with the long sentence. There was also the incidence of 'telephone voice' where individuals were marked as having an Southern British Standard accent given the Shibboleth sentence, but later sentences produced by the same talker seemed to indicate otherwise. It was hypothesised that the accent changed to a more natural one for the talker as the call proceeded.

SUBSCRIBER — A PHONETICALLY ANNOTATED TELEPHONY DATABASE

Due to these observations during the Bader database collection, the accent diagnostic sentence was split into two, and was placed after the other five sentences for Subscriber.

Figure 1 shows the number of calls against the point reached in the dialogue. It can be seen that most calls are lost on the first two prompts, but once the caller gets into the main dialogue there is only a small failure rate after each prompt. Interestingly the number of callers drops off slightly more than average once their own name was requested, suggesting a reluctance to provide this information.

Figure 1. Total number of calls remaining at each point in the dialogue.



3. VALIDATION

After call receipt, the talker's name and call reference number were listened to and cross referenced against details provided by the market research company. This allowed callers who had completed a call to be reimbursed. All the utterances were then validated and a quality decision made as shown in table 1.

Table 1. Quality decision categories.

GOOD	Utterances which follow the prompt exactly and which are not significantly affected by line noise, background noise or distortion
USEABLE	Utterances which differ from the prompt or utterances which are marked by line or background noise
BAD	Utterances which are so badly articulated or distorted by noise, or which produce such a weak signal, that they are difficult to label with any degree of accuracy.

4. ACCENTS

The Subscriber database contains speech collected from around the British Isles, encompassing several accents of English. The accent decision was made based on the 'Shibboleth' sentences as described in [1], which also discusses the performance of the Shibboleth sentence in correctly determining the accent . The 9 categories used for classification are as shown in table 2.

Table 2. Accent categories.

SBS	Southern British Standard (RP)
LON	London Area
R-WEST	West of England (Rhotic)
WAL	Wales
NB-LIV	Liverpool Area
NB	North of England
R-LANCS	Lancashire (Rhotic)
R-IRISH	Ulster (Rhotic)
R-SCOTS	Scotland (Rhotic)

5. PHONETIC ANNOTATION

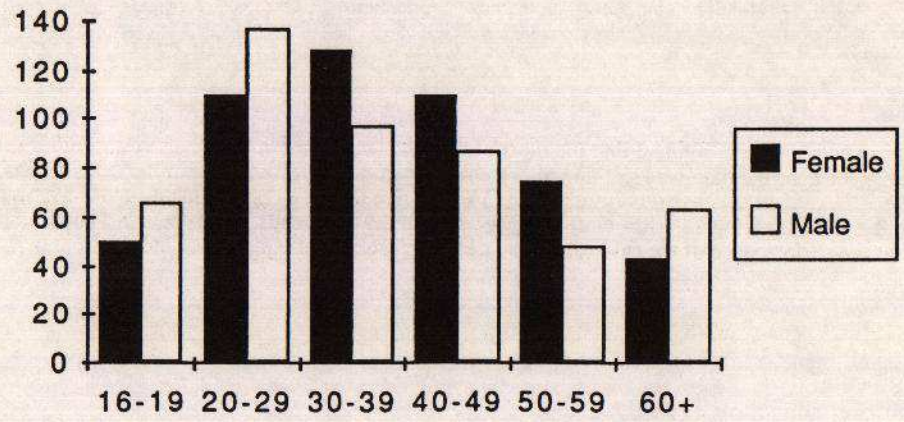
The main goal of the database collection was to segment the utterances at a phonetic level and to annotate the segments with machine readable phonetic symbols. A large inventory of symbols is needed to cover all the British accents and many of the symbols are shared between accents. A machine readable phonetic alphabet was developed at CSTR for this purpose, with a total of around 75 symbols. It would have been possible to approach the labelling of the database in one of two ways. Given an accent decision from the Shibboleth sentence the choice of symbols could have been restricted to a predefined set for that accent. Alternatively, any of the symbols could be used to label any of the speech. As the purpose of the database was to train acoustic-phonetic models the super-set approach was used.

For the labelling of non-speech sounds a set of special symbols was devised. These covered silence, breath noise, speaker noise, extra speech, an impulse and series of impulses. These symbols could be used either as the main symbol or as a *diacritic* used in addition to the phonetic symbol representing the foreground speech. The standards used for the labelling of the database are described in [2].

6. DATABASE STATISTICS

6.1. Total numbers of callers

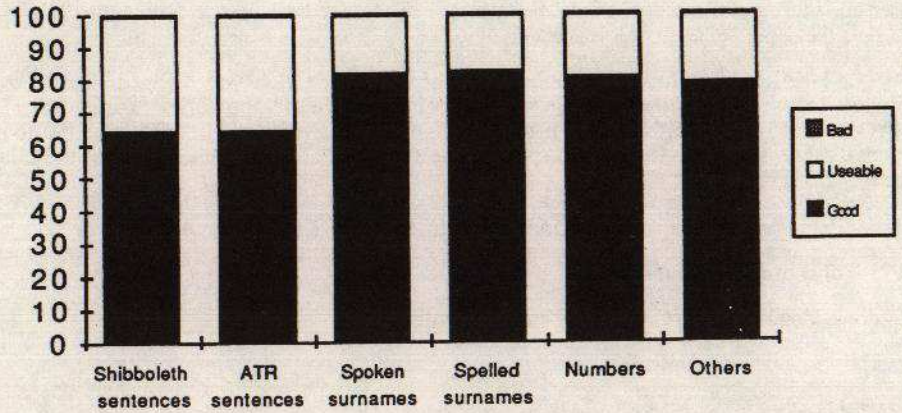
Figure 2. Total number of callers for each age group.



The distribution of female callers is as expected, but the distribution of male callers shows a dip in the range 50-59. This would need to be addressed in future, perhaps by recruiting more males in this age band if the response rate for these people is lower than average.

6.2. Quality decisions

Figure 3. Percentage of utterances of each type in each quality category.



It can be seen from Figure 3 that the most of the data was either in the categories *good* or *usable*. Only 0.6% of the data was lost due to extreme distortion or noise and is in the category *bad*. Of the 24% of data in the category *usable* 18.5% was due to noise contaminants, and 5.5% was due to the talker making an error while reading or speaking the words. It can be concluded therefore that the performance of the callers on this task was generally of a high standard.

7. COMPARISON WITH OTHER SPEECH DATABASES

Table 3. A comparison of speech databases.

DATABASE	Number of participants	Annotation Level	Sentences	Words	Other Speech?	Telephone Line?
SUBSCRIBER	1014	SUB-WORD	7	NAMES	√	√
DARPA ATIS	N/A		DIALOGUE			√
DARPA SWBD	500	WORD	2500	DIALOGUE		√
TIMIT	640	SUB-WORD	10			
N-TIMIT	640	SUB-WORD	10			√
DARPA RM DEP	12		1912			
DARPA RM IND	160		57			
SCRIBE	12	SUB-WORD	50	2 MINUTES	√	
ATR/CSTR	4	SUB-WORD	200	5000	√	
DARPA WSJ	264		WSJ			
BREF (FR)	120		MONDE			
PHONDAT (GR)	40	AUTOMATIC	BERLIN	SOME	√	

Table 3 shows that Subscriber is amongst the largest of the phonetically annotated databases. It is the only database which could be used for sub-word unit speech recognition over the British telephone network.

8. CONCLUSIONS

The method used to recruit and reimburse callers resulted in 81% of those recruited successfully completing their calls. For those callers who completed the calls, 64% of the sentence data and 80% of the other data were *good*, ie followed the prompt exactly and were not significantly affected by line noise, background noise or distortion.

Accent diagnostic sentences should be towards the end of the call to minimise the effect of 'telephone voice'.

Subscriber is a unique resource for the investigation of sub-word unit speech recognition over the telephone network.

9. ACKNOWLEDGEMENT

The authors would like to thank Professor Mervyn Jack of CSTR for the information contained in Section 7 of this paper.

10. REFERENCES

[1] K Edwards et al., *The design and performance of two accent diagnostic "Shibboleth" sentences'*, Proc IOA Speech and Hearing 1992

[2] C M Scott et al., *'Standards for labelling and segmentation of telephone quality speech'*, Proc IOA Speech and Hearing 1992

[3] G P Walker, *'A speech database collection architecture'*, Proc IOA Speech and Hearing 1990, Vol 12: Part 10, pp159-164