S2.17

AN ENGLISH LANGUAGE SPEECH DATABASE AT THE UNIVERSITY OF WESTERN AUSTRALIA

E. M-K. Lai[*], *G.A. Carrijo*[*], *R. Bennett*[*],
R. Togneri[*], *M. Alder*[**] *and Y. Attikiouzel*[*]

* Department of Electrical & Electronic Engineering/
** Department of Mathematics
The University of Western Australia, Nedlands, Western Australia 6009

ABSTRACT

This paper is a report on the content and status of a major speech database collection effort at the Department of Electrical and Electronic Engineering of the University of Western Australia. The goal is to collect a useful set of speech material from a very large number of speakers. These speakers are drawn from a wide cross-section of the local community with a variety of ethnic and education backgrounds. Speech materials include isolated digits and numbers, vowels and voiced phonemes, connected digits, and phonetically balanced sentences. Speech signals are encoded into 16-bit PCM format and stored onto Betamax format video tapes. In the seven months since this project started, speech from 100 speakers have been collected. A statistical break-down of the backgrounds of the speakers is also presented.

INTRODUCTION

Successful research and development of practical speech recognition algorithms and systems depends very much on the quantity and quality of speech data available. A number of large speech databases have been constructed or are under construction in the United States [1-3], France[4-5], the United Kingdom[6] and Japan[7-8]. They are being used mainly for the evaluation and testing of speech recognition algorithms and systems. Some of them have been made available to the speech research community.

Our main research interest is in developing techniques and systems for speaker-independent speech recognition that are practical for use in Australia. This means that the database must contain speech taken from a very large number of speakers. It also means that what is generally considered as the Australian accent must be captured. The ethnic heterogeneity of Australians is well known and must also be taken into account if the sytems are to be useful locally for the general public. Unfortunately, these constraints make the above-mentioned databases unsuitable for our purposes. Moreover, in Australia little effort has been made in collecting large quantity of speech materials for speech recognition research. So far, there is no report of any generally available speech database that is collected in Australia. These factors prompted our decision to construct a large Australian speech database that will eventually be generally available for speech researchers both in Australia and overseas.

This paper is a report on the content and current status of our effort in collecting a large-scale speech database. In the following sections we shall describe the speech materials gathered, the subjects' composition and background, the procedure and method of storage of speech data. We commenced data collection in March 1989. Speech from 110 subjects has been collected up to October 1989. The experiences gained during this period of time and plans for improvements are also discussed.

SPEECH MATERIALS

The choice of speech materials is dictated by our current areas of research interest in automatic speech recognition. They can be divided into 4 separate categories:

(1) isolated digits, numbers and words

This category include digits from zero to nine, numbers from eleven to nineteen and also ten, twenty, up to ninety. Other words such that make up a typical number. They are 'hundred', 'thousand', 'million', 'billion', and 'and'.

(2) vowels and diphthongs

Nineteen major vowels and diphthongs are included in this category. As an aid to pronounciation, legitimate words that start and end with consonants with the vowel or diphthong embedded in-between (i.e. CVC words) are used. The nineteen words are listed in Appendix A.

Dr. Carrijo is on study leave from Universidade Federal de Uberlandia, 38400 Uberlandia, M.G., Brazil with financial support of CNPq, Brazil.

CH2847-2/90/0000-0101 $1.00 © 1990 IEEE

(3) connected digits

In this category there are thirty-five connected digit strings each of which is seven digits long. All possible transitions from one digit to another can be found in this set of digit strings. These digit strings are of the same length as normal telephone numbers in most Australian cities.

(4) phonetically balanced sentences

Six phonetically balanced sentences in which all most commonly used phonemes can be found. The sentences used are listed in Appendix B.

Isolated and connected digit recognition has been one of our major areas of research in speech recognition. It has a number of applications in areas such as numerical data entry, hands-free telephone dialing, telephone directory assistance, and telephone banking systems. Hence categories (1) and (3). Another area of research interest is phoneme-based large vocabulary isolated word recognition and connected speech recognition. The phonetically balanced sentences will provide us with samples of phonemes which will help us in developing pattern-matching and/or rule based phoneme recognition systems. It has been shown that coarse phonetic recognition could reduce the search space in a large vocabulary recognition system by a large amount [9]. By classifying vowels and diphthongs into 3 categories instead of just one further reduces the number of plausible word candidates [10]. The vowels and diphthongs category shall help in our research in this area further.

SUBJECTS

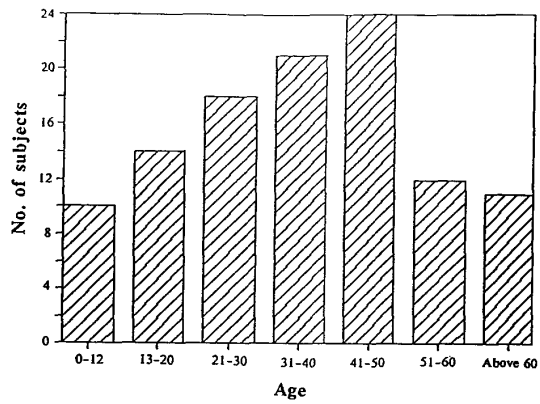
It is our intention to construct a speech database that will be useful in the research and development of techniques and systems for speaker-independent speech recognition that are practical for use in Australia. Therefore the ethnic heterogeneity of Australians must be taken into account. In an effort to attract subjects from a wide cross-section of the Perth community this project was publicised through articles in the state-wide newspaper, the "West Australian", as well as some local suburban newspapers.

Table 1: Childhood Country of Subjects

Country	Percentage of Subjects
Australia	56%
Britain	22%
Africa	7.3%
Other European Countries	4.5%
New Zealand	2.8%
USA	2.8%
India	2.8%
Far East Asia	1.8%

In the seven months since data collection commenced in March 1989, 110 people volunteered, of which 55 are males and 55 are females. Figure 1 shows the age distribution of these subjects. 56% of them are born and grew up in Australia while the rest are from the United Kingdom, other parts of Europe, South Africa, New Zealand, India and Far East Asia. Table 1 shows the percentage of subjects who spent their childhood in each of these countries. The majority (85%) of those who are 21 years of age or above are tertiary educated.

Fig. 1: Age Distribution of the Subjects



RECORDING

The Hardware System

Speech utterances are encoded into 16-bit linear PCM (pulse-code modulated) format sampled at 44.1 kHz using a Sony PCM-501E Digital Audio Processor. The encoded data are stored onto the video tapes through a Betamax format video recorder. After the recording session, the recorded speech data is then played back through the Digital Audio Processor and re-digitized using the voice data acquisition sytem. The voice data acquisition system samples at 10KHz with 12-bit accuracy. Each digitized utterance is played back through the speech output system. Details of these systems could be found in [11]. Figure 2 shows a block diagram of our recording and data acquisition and playback equipment.

Software

There are two main software programs we have developed for use in this project. Both of them are written in the 'C' programming language, except for time-critical portions which are written in 8086 assembly language, on an IBM-PC/AT compatible computer running the MS-DOS operating system. The first program is used for collecting digital data from the voice data acquisition system, for displaying them graphically on the graphic screen, for playing them back through the speech output

system, and for saving them onto disk files. The second program guides the subjects through the recording, giving instructions when necessary and displaying the words in the database list one at a time. There are two parts to this program. The first part is a demonstration designed to familiarise the subject with the format of the recording and adjusts the speed at which words are presented. The second part simply takes the subject through the database list of words during the actual recording session[12].

Procedure

All recordings are done in a quiet room. The room is not acoustically isolated. However, the noise level is negligible. A high quality uni-directional dynamic microphone is placed about 15cm from the lips of the subject. The subjects are requested to provide information on their ethnic and education backgrounds for statistical purposes. They will then go through a demonstration session of the prompting program which will familiarise them with the format of the recording. In the actual recording session, the subject will read all the items in the database once.

Remarks

Our experiences have shown that the data collection program which takes the subjects through the recording session is invaluable. Our subjects come from a wide variety of backgrounds and some may feel intimidated by the equipment around them. The demonstration portion of the program helps them to relax and the pace at which words are presented could be slowed down for those who find the standard pace difficult to follow. As a result their speech is not tense and more closely resemble their normal way of speaking.

CONCLUSIONS

Details of the content and current status of the English speech database collected at the Department of Electrical and Electronic Engineering of the University of Western Australia are presented. Voices from 110 subjects have so far been collected. These subjects come from a wide cross-section of the local community and the data collected will prove to be useful in developing techniques and systems practical in the Australian environment. This is an on-going project and more data are continuously being collected.

ACKNOWLEDGEMENTS

The authors wish to thank Miss L. Avery for her help in programming the user prompting program for data collection.

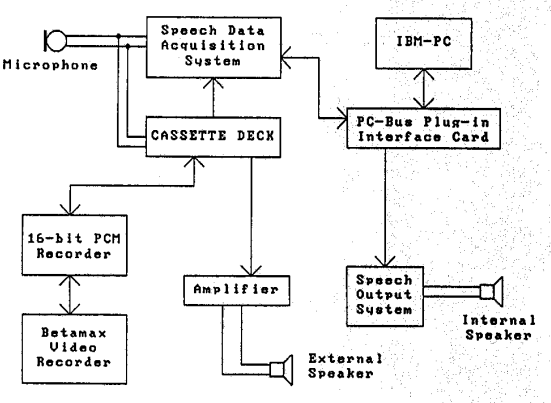


Fig. 2: Block Diagram of Speech Data Collection, Digitization and Playback Equipment Setup

REFERENCES

[1] R.G. Leonard, "A Database for Speaker-independent Digit Recognition", Proc. ICASSP-84, Paper 42.11, 1984.

[2] M.F. Guyote, K.A. Lewis & D. Lijana, "A Speech Database at the United States Air Force Academy", Proc. ICASSP-86, Paper 7.2, pp.313-316, Tokyo, 1986.

[3] P. Price, W.M. Fisher, J. Bernstein & D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", Proc. ICASSP-88, Paper S13.21, pp.651-654, New York, 1988.

[4] R. Carre, R. Descout, M. Eskenazi, J. Mariani & M. Rossi, "The French Language Database: Defining, Planning, and Recording a Large Database", Proc. ICASSP-84, Paper 42.10, 1984.

[5] G. Perennou, "B.D.L.E.X.: A Data and Cognition Base of Spoken French", Proc. ICASSP-86, Paper 7.5, pp.325-328, Tokyo, 1986.

[6] P.C. Millar, I.R. Cameron, A.J. Greaves & C.M. McPeake, "A Very Large Telephone-speech Database Collected Using an Automated Voice-interactive Dialogue", Proc. ICASSP-88, Paper S13.20, pp.647-650, New York, 1988.

[7] S. Itahashi, "A Japanese Language Speech Database", Proc. ICASSP-86, Paper 7.4, pp.321-324, Tokyo, 1986.

[8] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa & T. Wanatabe, "Construction of a Large-scale Japanese Speech Database and Its Management System", Proc. ICASSP-89, Paper S10b.12, pp.560-563, Glasgow, Scotland, 1989.

[9] R. Carlson, K. Elenius, B. Granstrom & S. Hunnicutt, "Phonetic Properties of the Basic Vocabulary of Five European Languages: Implications for Speech Recognition", Proc. ICASSP-86, Paper 51.4, pp.2763-2766, Tokyo, 1986.

[10] E.M-K. Lai, Y. Attikiouzel, "A Comparison of Several Coarse Phonetic Classification Schemes -- Preliminary Results", Proc. 1st Australian Conf. on Speech Science & Technology, Canberra, Nov. 1986, pp.316-321.

[11] E.M-K. Lai, "The Speech Data Acquisition and Output Systems: Hardware and Software", Tech. Report #SP-01/89, Dept. of Electrical & Electronic Eng., The Univ. of Western Australia, Oct. 1989.

[12] L. Avery, "Speech Database Collection Program", Pass Degree Project Report, Dept. of Electrical & Electronic Engineering, The Univ. of Western Australia, Oct. 1989.

APPENDIX A

List of Words in the Vowels and Diphthongs Category

Heat	Hit	Hat	Head
Heart	Hot	Hut	Hoard
Hood	Hoot	Hurt	Height
Hate	Void	Pout	Hoed
Wierd	Cared	Tour	

APPENDIX B

List of Phonetically Balanced Sentences

(1) Measure three young kids for height.

(2) Which boat tour should they join now?

(3) Some vagabonds share an apartment.

(4) How do we go there from here?

(5) Black soot and parks annoy her.

(6) You'll be my love for always.