CDIGITS: A LARGE ISOLATED ENGLISH DIGIT DATABASE

Yaxin Zhang, Mylene Pijpers, Roberto Togneri, Mike Alder
Center for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia

Abstract

This paper described a large isolated English digit database which was designed for the training and evaluation of statistical algorithms and neural networks. 1108 speakers (575 males and 533 females) were recorded in the UWA campus under office environment.

INTRODUCTION

Stochastic modeling approaches have been successfully used in the speech recognition area as the dominant strategies during the past decade. They are preferable over other approaches not only because they always give higher system performance, but also, more importantly, because they can model the recognition units together with the variations of the speech signal, that always occur for different speakers and repetitions from the same speaker. It is extremely important in the speaker-independent case in which the same utterance is normally spoken by a wide range of speakers with different gender, age, speaking habits, and especially different accent. It is obvious that a speaker-independent speech recognizer is required to deal with a large range of variations of the speech signal in both the time domain and frequency domain. For this purpose, a database with a large number of speakers is always required for the training.

CDIGITS was designed for the stochastic modeling approaches and the probabilistic neural network approaches for speaker-independent isolated English digit recognition of The Center for Intelligent Information Processing Systems (CIIPS) of The University of Western Australia. The database design was aimed at a large amount speakers other than a large vocabulary range. The vocabulary consists of isolated English digits from zero to nine and oh, 11 words. 1108 speakers were recorded, 575 males and 533 females. Each speaker read the isolated digits twice. All speakers are adults, aged about 17 to 60, and are educated. The data sampling rate is 16 KHZ with sampling precision of 14 bits. A 7.8 KHZ antialiasing low-pass filter was fixed on the data acquisition board.

The data was collected within the campus of The University of Western Australia. The recording took place in the CIIPS laboratories, a reading room of the university library, and a dark room in the undergraduate student Guild building.

Due to the recording performed in the general office environments, the data has quite high background noise. The average signal-to-noise-ratio is about 30 dB. This database was classified into 4 groups according to the SNR value and the noise level of data for each speaker. It has been used as a part of test sequences for speech recognition experiments in CIIPS speech processing group. We expect that this database would be used for the speech recognition system evaluation for the tasks of speaker-independent isolated word recognition in an office environment.

THE DATA ACQUISITION SYSTEM

The data acquisition system is inexpensive and portable. It consisted of a IBMPC-486 compatible computer, a TMS320C30 DSP board, a playback speaker, a unidirectional micro phone, and a control deck. The software is based around the TMS320C30 Evaluation Module (hereafter referred to as the EVM board). The EVM board is a half-length board that installs in a vacant slot in a PC. Only the data acquisition facilities of the EVM board were used.

The EVM board has two important parameters that govern the low pass filter, the sampling frequency, and the high pass input filter which is switchable. These two parameters are referred to as **a** and **b**. The governing equations are:

$$Sampling\_frequency = \frac{3750kHZ}{a*b}$$

$$Lowpass\_frequency = \frac{46.875kHZ}{a}$$

$$Highpass\_frequency = \frac{28.125kHZ}{a*b}$$

The maximum sampling rate may be 19.2kHZ, but the maximum lowpass filter cut off frequency is 5.5kHZ. According to Nyquist criterion this would mean the maximum sampling rate that could be used would be 11kHZ to avoid aliasing. However, there is no evidence to imply that the 5.5kHZ maximum is correct, as the specifications do not set any such maximum. Thus the default setting for **a** and **b** were chosen at **a**=6 and **b**=39. This gives the best resolution of sampling rate, 16.02kHZ, which unfortunately does not quite make the 19.2kHZ. The reason for this is that for these values, the lowpass cutoff is 7.812kHZ and a decrement in **a** would exceed the maximum 19.2kHZ, and a decrement in **b** would give a lowpass cutoff of more than half the sampling rate, so aliasing would become a problem. Both decrementing **a** and decrementing **b** would give a faster sampling rate, but this exceeds the operating levels. These values give a high pass cutoff of 120HZ. As 120HZ may remove some information from the speech, the highpass filter is switched off by default, although this means that a 50HZ is able to come through the power supply.

The EVM board also has a microphone gain option of either ±1.5 or ±3 volts. As the microphone used does not have a very high output, the gain was set at ±1.5 volts. The last function the EVM board has to offer for A/D and D/A conversions is a $\frac{\sin x}{x}$ filter. This has been switched on by default.

The control deck is a mixer/graphic equalizer. It has facilities for 2 microphones, and two auxiliary channels, as well as gain control for each input and a master gain control. There is also a bass and treble equalizer, and a decibel power meter. The master gain at 5 was set and with the microphone gain at approx 6 for loud yet undistorted sounds.

DATA RECORDING

Because the recording was performed in general office environments, the background noise from the computer fan, hard disk driver, and general talking from outside, *etc*, are very high although we used a unidirectional microphone.

The operation of the recording was based on software. In each recording a speaker interacted with the keyboard or uttered a word in front of the microphone following the promoting commands shown on the computer screen. The speaker initiated each recording by hitting a key and uttering the word within the 1.15 second time limit. The software automatically created for each speaker a directory which included the all words recorded from the speaker and the information about each speaker. Each word was stored separately as a file. A directory was named using the speaker's name (5 characters from the first name and 3 from the last name). A data file was identified by the word recorded. Furthermore, the speaker recodings were plotted in separate male and female directories.

DATABASE FORMAT

The data format was base on a simple header plus data. Each word file consists of a 128 bytes header followed by the row data of short integers. The header construction is shown as follows.

Items	*Number of bytes*	*Description*
Type_of_file	[4]	*In this case we used* **.dat**.
Speaker_id	[30]	*The name of the speaker, firstname, lastname.*
Sample_freq	[6]	*The sampling frequency,* $default = 16013Hz$.
Speech	[30]	*What was said.*
Preemphasis	[1]	*Not used in this case.*
Sample_bits	[3]	*The number of sample bits,* $default = 14$.
Signed_data	[1]	*Signed data* **Y** *or* **N**, $default =$ **Y**.
Reserved	[53]	*For future use.*

After recording the data was transferred, with appropriate byte swapping, to a SUN-SPARC workstation for further processing. A secondary storage medium like tape and Magnetic-Optical disk were used to archive and access the data off-line.

Speaker	richard whightman
Sex	Male
Age	22
Recorded at	Thu Dec 03 10:59:58 1992
Word files recorded	zer0, one0, two0, ..., nine0, oh0
Appended at	Thu Dec 03 11:02:06 1992
New files added	zero1, one1, two1, ..., nine1, oh1

Table 1: An example of information file

DATABASE MANAGEMENT

Because the data was collected from volunteer speakers, it was very difficult to control the speaker's loudness, habits and speaking status. The amplitude of some data was too high, some too low, some with intolerable noise, and some totally useless. Unacceptable data was removed and the remaining data was in three categories.

(1) **Acceptable**, 977 (518 males and 459 females).

(2) **Problematic**, 109 (4 H_females, 19 H_males, 64 L_females, and 32 L_males).

The "H" indicates the amplitude of data is too high which means at least one sample in the data file is greater than 8191 or less than -8192. The "L" indicates the amplitude of data is too low which means the maximum absolute amplitude in the data file is less than 500.

(3) **Once only**, 22 (6 females and 16 males). Only one repetition was remaining as acceptable.

Normally 23 files form a directory which includes 22 utterances from one speaker and an information file called **_info.txt**. The 22 utterance files are named za.dat, 1a.dat, 2a.dat, ..., 9a.dat, oa.dat, and zb.dat, 1b.dat, 2b.dat, ..., 9b.dat, ob.dat. The first character of a file name indicates what word is it, and the second one, a or b, indicates whether it belongs to the first or second repetition. The information file includes the speaker's name, gender, age, recording time, as well as word files recorded. Table 1 is an example of an information file.

Since some speakers have same names as others, not only the surname, but also the given name, we renamed each speaker's directory using the alphabets after the data was transferred to the SUN workstation. For male speakers, the directory names were specified by a capital character followed by a lowercase character from Aa, Ab, ..., Az, Ba, Bb, ..., Bz, ..., Tx (518). While for female speakers, the directories were named by two lowercase letters from aa, ab, ..., az, ba, bb, ..., bz, ... rq (459). The data management construction is shown as Figure 2.

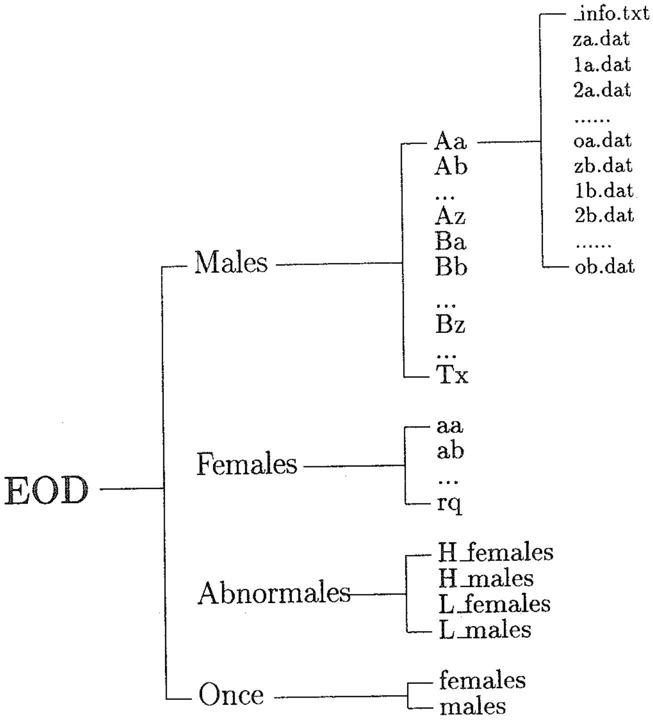

Figure 1, The construction of CDIGITS database management

DATABASE ANALYSIS

The signal-to-noise ratio (SNR) is a very important feature of a database. We used the formula

$$SNR(dB) = 10 \log \frac{E_s}{E_n} \quad (1)$$

and following method to estimate the SNR of each data file. In the formula E_s is the average energy of a data file and E_n is the average energy of background noise of a data file. E_s was calculated for a whole data file while E_n was estimated from the first 100 samples and last 100 samples of a data file, where we normally expect only background

Group number	SNR range	Number of speakers
1	55dB$\le$ SNR	83
2	40dB$\le SNR <$55dB	422
3	30dB$\le SNR <$40dB	366
4	SNR$\le 30dB$	131

Table 2: A classification of speakers according to the SNR of data files

noise. Since each data file has 18432 samples,

$$E_s = \frac{1}{18432} \sum_{i=1}^{18432} s_i^2 \tag{2}$$

and

$$E_n = \frac{1}{200}[\sum_{i=1}^{100} s_i^2 + \sum_{i=18333}^{18432} s_i^2] \tag{3}$$

Because we collected the data from different places with different background noise level, the data files have large variation in noise level. Based on above equations, we classified the data into four groups corresponding to different ranges of signal-to-noise ratio. These are shown in Table 2. Since each speaker's directory has 22 data files, a speaker was classified into a certain group only when the SNRs of all data files in the directory were above the lower limit.

CDIGITS has been used in CIIPS speech recognition group for the system training and evaluations [3,4].

ACKNOWLEDGEMENT

This work was supported in part by The University Fee-Waiver Scholarship and The University Research Studentship of the University of Western Australia. Authors would like to thank the data acquisition software developer Jason Waddell.

REFERENCES

1. Lai, E., Carrijo, G., Bennett, R., Togneri, R., and Alder, M. (1990), "An English Language Speech Database at the University of Western Australia", Proceedings of ICASSP-90, pp. 101-104.

2. Waddell, J. (1992), "Speech Data Acquisition", A Bachelor degree thesis in Dept. of EEE, The University of Western Australia.

3. Zhang, Y., Alder, M., & Togneri, R. (1993), "Using Gaussian Mixture Modeling for Phoneme Classification", Proceedings of First Australian and New Zealand Conference on Intelligent Information Systems, Dec. 1993, pp. 649-652.

4. Zhang, Y., & Alder, M. (1994), "An Improved Training Procedure for Speaker-independent Isolated word recognition", Proceedings of 1994 International Symposium on Speech, Image Processing and Neural networks, April 1994, Hong Kong. pp. 722-725.