A PRELIMINARY TEST OF MAT-160 SPEECH DATABASE IN CONNECTED SYLLABLES RECOGNITION

Rong-Liang CHIOU and Hsiao-Chuan WANG

Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan 30043
Tel: +886-3-574-2587, Fax: +886-3-571-5971, E-mail: hcwang@ee.nthu.edu.tw

ABSTRACT

A project to collect Mandarin speech data across Taiwan (MAT) will generate a speech database of 5000 speakers. A sample database of 160 speakers, called MAT-160, is extracted for non-profit distribution and preliminary studies. This paper presents a preliminary study on using MAT-160 for connected syllable recognition. It shows not only the technique of channel compensation in telephone speech recognition, but also the utilization of MAT-160 database in speech researches. MAT-160 contains about 42,000 Mandarin syllables in 10,560 speech files. It includes 407 base syllables in Mandarin speech which can be used for training all the necessary sub-syllables for Mandarin speech recognition. Several channel-effect compensation methods are investigated for comparison.

1. INTRODUCTION

Mandarin Speech across Taiwan (MAT) is a speech data collection project conducted by a group of researchers in Taiwan [1]. The speech data were collected at nine recording stations through telephone networks during 1995-1998. The goal is to generate a speech database of 5000 speakers in Taiwan. The spoken materials were designed for generating speech models and evaluating the telephone-based speech recognition systems developed for Mandarin speakers. The contents of the database include answering statements, numbers spoken in different ways, isolated Mandarin syllables, isolated words, and phonetically balanced sentences. A sample database of 160 speakers (81 males and 79 females) is extracted for non-profit distribution and preliminary studies. This sample database is coded MAT-160 which includes 10,560 speech data files with more than 42 thousands of Mandarin syllables.

This paper presents a preliminary test on MAT-160 database. The target is to recognize strings of Mandarin syllables without concerning the tones. The MAT-160 contains 407 base syllables in Mandarin speech. It allows the training of all necessary sub-syllable models for Mandarin speech recognition. In this test, isolated syllables, words, and sentences in MAT-160 are used for generating the speech models. A set of 500 utterances obtained from other 30 speakers through telephone networks is the test database. Several channel effect compensation methods have been investigated.

2. CONTENTS OF MAT-160 SPEECH DATABASE

In MAT project, nine speech data collection stations were set up in different cities. Each station consisted of a personal computer equipped with a telephone interface card, a sound card, and the software for speech data recording and speech file editing. A dedicated file format was designed for MAT speech files. The file header contained the necessary information about the speech data and also the Chinese characters and Pinyin transcripts of the recorded utterance. The PCM data of speech signal were stored in binary format which retained the waveform of the recorded utterance and its preceding and succeeding silent portions of about 0.5 seconds.

The framework of speech material design for MAT project was created by Dr. Chiu-Yu Tseng of Academia Sinica [2]. The materials were extracted from two text corpora of 77324 lexical entries and 5353 sentences. Forty sets of speech materials were produced for generating the prompting sheets. Besides, the database also contained 200 numbers pronounced in five different ways, such as dates, times, prices, telephone numbers, and car plates. The prompting sheets were designed for guiding the speakers to input their speech data. It also asked questions to gather information about the speaker, such as his/her gender, age, language background, education level, and residence. Totally, each speaker has to input 66 utterances in about 6 minutes through a telephone handset in an interactive mode. The speech recording system had been designed to automatically write Chinese characters and Pinyin transcripts onto the file header according to the contents in prompting sheet except the answers to the questions.

The MAT-160 database is further divided into five sub-databases.

(1) MATDB-1 short answers
(2) MATDB-2 numbers spoken in five different ways
(3) MATDB-3 isolated syllables
(4) MATDB-4 isolated words of 2 to 4 syllables

(5) MATDB-5 phonetically balanced sentences

In the following experiments, MATDB-3, MATDB-4, and MATDB-5 are used for generating the sub-syllable models. An additional database of 30 speakers that are different from the speakers in MAT-160 is for testing. This test database contains 200 isolated words and 300 sentences recorded in telephone networks.

3. PHONOLOGY OF MANDARIN

Mandarin is a syllabic and tonal language. Each Chinese character is pronounced as a monosyllable. The structure of Mandarin syllables can be expressed in terms of the initials, the finals, and the tones [3]. If the tones are ignored, the number of distinct syllables is 408. The tones are specified by the pitch contours as described in Table 1.

Table 1 Tones of Mandarin syllables

Tone	Pitch pattern	Notation
Tone-1	High level	
Tone-2	High rising	ˊ
Tone-3	Falling-rising	ˇ
Tone-4	High falling	ˋ
Neutral tone	none	·

Since the tones can be identified by their pitch contours, they are separately processed in most of Mandarin speech recognition systems [4]. Therefore, a Mandarin syllable is usually recognized by its structure of initial part and final part. The syllable without tone is referred as the base syllable. The initial is a preceding consonant and the final is the followed vowel portion. Some of syllables may have no initial consonant, and they are referred as null initials. In Mandarin speech, there are 21 initials (not including the null initial) and 38 finals (not including 2 empty vowels). Table 2 shows all the initials and finals in the Mandarin speech.

Table 2 Initials and Finals in Pinyin symbols

	Pinyin	
Initials	b, p, m, f, d, t, n, l, g, k, h, j, q, x zh, ch, sh, r, z, c, s	
Finals	a	a, ai, au, an, ang
	o	o, ou
	e	e, e(è), ei, en, eng
	er	er
	i	i, ia, io, ie, iai, iau, iou, ian, in, iang, ing
	u	u, ua, uo, uai, uei, uan, uen, uang, ung
	ü	ü, üe, üan, ün, üng

In a syllable, the beginning portion of the final is affected by its preceding consonant. A more realistic approach for identifying an initial is to recognize the right-context-dependent initials (RCD-initials). Totally, there are 94 RCD initials in Mandarin speech. By this arrangement, the phonetic units for Mandarin syllable recognition are 94 RCD-initials and 40 context-independent finals (CI-finals).

4. SUBSYLLABLE MODELS AS RECOGNITION UNITS

Let RCD-initials and CI-finals be the basic units of Mandarin speech. Hidden Markov models are used to model these RCD-initials and CI-finals. In our study, we express each RCD-initial by 3 states and each CI-final by 4 states. For those syllables without initial consonants, 2 states are used to model their null initial part. Besides, a silence state is applied to represent the pause portions in an utterance. The total number of state models is 519 which are specified as follows;

3 states x 94 CD-initials = 282 states
4 states x 40 finals = 160 states
2 states x 38 null initials = 76 states
1 state x 1 silence = 1 state

The speech signal is sampled at the rate of 8 kHz. The frame size for signal processing is 256 points and overlapped by 128 points. The signal is pre-emphasized before Hamming window of 256 points is applied to each frame. Then the logarithmic energy (Log-Eng) and Mel-frequency cepstral coefficients (MFCCs) of each frame are calculated based on the windowed samples. The logarithmic energy has been normalized by its maximal value in the utterance in order to eliminate the effect of different loudness of the speech. The Fast Fourier Transform (FFT) algorithm is applied to each frame to find its spectrum. This spectrum is passed through a set of 20 triangular band-pass filters in Mel-scale. The logarithm of these 20 Mel-frequency spectrum is then converted into cepstrum by discrete cosine transform (DCT) algorithm. The feature vector derived from a speech frame is a vector of 26 elements which includes 12 MFCCs, 12 delta MFCCs, one delta Log-Eng, and on delta-delta Log-Eng.

5. RECOGNITION OF BASE SYLLABLES IN SENTENCES

During the recognition phase, a technique called one-stage dynamic programming is applied to decode an input utterance into a sequence of Mandarin syllables. The speech files in MAT-160 speech database has been manually screened so that the noise has been minimized. In this study, only the fact of channel distortion is concerned. Two approaches are proposed to attack this target. One is to estimate the channel bias and adjust the speech models to the channel environment. The other is to subtract the estimated channel bias from the speech signal so that the channel effect to the signal is minimized. In our experiments, the Bayesian affine transformation and the Bayesian bias transformation [5]

are applied to adjust the speech models. The signal bias removal (SBR) [6] and the Hierarchical signal bias removal (HSBR) [7] methods are used to compensate the channel bias in the signals.

5.1. Bayesian affine transformation

This method applies an affine transformation function,

$$y = \mathrm{A}x + b, \tag{1}$$

where A and b are the estimated transform matrix and bias vector, respectively. Let $Y = \{y_t\}$ be the observation sequence, $S = \{s_t\}$ be the state sequence, and $L = \{l_t\}$ be the mixture sequence. Then the probability of the observation y_t for state n and mixture m is given by

$$P(y_t \mid s_t = n, l_t = m, \eta = (\mathrm{A}, b)) = (2\pi)^{D/2} \left|\mathrm{A}\Sigma_{n,m}\mathrm{A}\right|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(y_t - \mathrm{A}\mu_{n,m} - b)^T (\mathrm{A}\Sigma_{n,m}\mathrm{A})^{-1} (y_t - \mathrm{A}\mu_{n,m} - b)\right\} \tag{2}$$

The maximum a posteriori (MAP) method is used to estimate the parameter set, $\eta = (\mathrm{A}, b)$.

For simplicity, the matrix A is assumed to be a diagonal matrix. Under this assumption, A and b can be solved in closed forms.

5.2. Bayesian bias transformation

If matrix A is an identity matrix, it results in a form of compensation by bias vector only. This is called the Bayesian bias transformation. The probability function becomes

$$P(y_t \mid s_t = n, l_t = m, \eta = b) = (2\pi)^{D/2} \left|\Sigma_{n,m}\right|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(y_t - \mu_{n,m} - b)^T \Sigma_{n,m}^{-1} (y_t - \mu_{n,m} - b)\right\} \tag{3}$$

The maximum likelihood (ML) method can be used to solve for the bias vector b.

5.3. Signal bias removal

Signal bias removal (SBR) is a method based on maximum likelihood algorithm. It estimates the difference between the test environment and the training condition so that the difference is removed during the recognition phase. Let $Y=\{y_t\}$ be the test observation sequence, $X=\{x_t\}$ be the supposed observation sequence in training environment. The difference between these two sequences is

$$\bar{b} = y_t - x_t \tag{4}$$

This difference, or called the mean bias, can be estimated by the following equation,

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu}_t), \tag{5}$$

where $\hat{\mu}_t$ is a codeword of state S_i which gives the maximum observation probability,

$$\hat{\mu}_t = \arg\max P\left\{y_t \mid b, S_i\right\} \tag{6}$$

The bias can be estimated recursively to improve its accuracy. Finally, the adjusted observation is given as;

$$\tilde{x}_t = y_t - (\bar{b}^{(n)} + \bar{b}^{(n-1)} + \cdots + \bar{b}^{(1)} + \bar{b}^{(0)}) \tag{7}$$

5.3. Hierarchical signal bias removal

If we assume that the bias is not a constant in an utterance, we should consider the bias as a frame-dependent vector.

$$\bar{b}_t = y_t - x_t \tag{8}$$

Before we go into the bias estimation, we calculate all the differences between the test frames and the corresponding state model means. Then we cluster the frames into M clusters so that the biases are also defined in M clusters.

$$b_j^c = \frac{1}{T_j}\sum_{l=1}^{T_j}(y_{t_j(l)} - \hat{\mu}_j) \tag{9}$$

where $y_{t_j(l)}$ is a frame belonging to cluster j, and T_j is the total number of frames belonging to cluster j. The clustered bias is calculated by the equation,

$$\bar{b}_t = \frac{\sum_{j=1}^{M} b_j^c w_{t,j}}{\sum_{j=1}^{M} w_{t,j}} \tag{10}$$

where

$$w_{t,j} = \frac{1}{(y_t - \hat{\mu}_j)^2} \tag{11}$$

is a cluster weighting factor. Similar to SBR, the bias vector can be recursively calculated.

6. EXPERIMENTS

The reference models are 519 state models. We use isolated syllables, isolated words, and phonetically balanced sentences in MAT-160 for training the state models. There is about 37,700 Mandarin syllables in the training data. For testing, 500 utterances were collected from 30 speakers (15 males and 15 females) who were different from those speakers in MAT-160. The test speech database includes 200 isolated words and 300 sentences. They are totally 4754 syllables in the test database. The recognition rate is calculated by the equation;

Recognition rate = 1 – (Substitution rate + Deletion rate + Insertion rate).

The experimental results are summarized in Table 3. For the cases of SBR, three types of bias definition are used;

Type I – Only one bias vector is calculated.
Type II – One bias vector is calculated for all

speech models and one bias vector for silence model.

Type III – Three biases are calculated for CI-finals,

RCD-initials, and silence, respectively.

The case of no compensation, referred as baseline test, is also presented for comparison.

Table 3 Syllable recognition rate (%)

Mixture number	4	8	16
Baseline test	35.08	37.02	39.33
Bayesian Affine Tran.	39.84	40.85	42.85
Bayesian Bias Trans.	39.04	40.13	42.01
SBR (Type I)	39.61	41.48	42.83
SBR (Type II)	39.73	41.29	43.02
SBR (Type III)	39.86	42.01	43.82
HSBR	39.48	41.15	43.21

From the experimental results, we find that the recognition rate of baseline test is far below our expectation. The substitution error has contributed most of error rate, i.e. about 50%. The detail of baseline test is shown in Table 4.

Table 4 Baseline test

Mixture number	4	8	16
Insertion error (%)	9.67	9.98	9.41
Deletion error (%)	1.64	1.47	1.60
Substitution error (%)	53.61	51.53	49.65
Recognition rate (%)	35.08	37.02	39.33

It is clear that the state models are not accurate enough for discriminating all the recognition units in Mandarin speech. One of the possible reasons is that we do not know the segmentation accuracy of sub-syllables during the training process. This may cause the inaccuracy in training the reference models. The other fact is that the number of data for training the state models is small. This size of speech data may not be able to generate reliable speech models.

As far as the channel-effect compensation is concerned, the best result is by using SBR Type III. The major improvement is in the reduction of substitution errors. The overall improvement is about 4.7% in the recognition rate. A detail is shown in Table 5.

Table 5 Experimental result of using SBR Type III

Mixture number	4	8	16
Insertion error (%)	8.49	8.53	8.17
Deletion error (%)	1.81	1.45	1.62
Substitution error (%)	49.84	48.01	46.39
Recognition rate (%)	39.86	42.01	43.82

7. CONCLUSION

Several channel compensation methods have been examined for the syllable recognition using MAT-160 speech database. Relatively, SBR method is the most promising one because of its simple implementation and better performance. The experimental result also shows that the substitution error is high. This may be due to the insufficient speech data in MAT-160 for generating the reliable models. However, the database is still good enough for the investigation of some channel effect compensation methods.

ACKNOWLEDGEMENT

This research has been supported by the National Science Council, Taiwan, ROC, under the contract number NSC87-2213-E-007-031.

REFERENCES

1. H.-C. Wang, "MAT – a project to collect Mandarin speech data through telephone networks in Taiwan," *Computational linguistics and Chinese language Processing*, vol.2, no.1, pp.73-90, 1997.
2. C. Y. Tseng, "A phonetically oriented speech database for Mandarin Chinese," *Proceedings of ICPhS'95*, Stockholm, Sweden, 1995, vol. 3, pp. 326-329.
3. C. N. Li and S. A. Thompson, *Mandarin Chinese: A functional reference grammar*, University of California Press, 1981.
4. L. S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal processing Magazine*, vol. 14, no. 4, pp. 63-101, July 1997.
5. J.-T. Chien, H.-C. Wang, and C.-H. Lee, "Bayesian affine transformation of HMM parameters for instantaneous and supervised adaptation in telephone speech recognition," *Proceedings of EUROSPEECH-97*, Rhodes, Greece, September 1997, vol.5, pp. 2563-2566.
6. M. G. Rahin and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 19-30, 1996.
7. M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, vol.3, no.4, pp.107-109, 1996.