• A PHONETIC LABELING METHOD BASED ON HIDDEN MARKOV MODELS AND VITERBI DECODING PROCEDURE

Hsiao-Chuan Wang, Rong-Liang Chiou, Shiang-Kai Chuang, and Yi-Fen Huang
National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

This paper presents a phonetic labeling method for the processing of Mandarin speech database produced in Taiwan. These speech data are collected through the telephone networks. Each utterance has been transcribed into Chinese characters and Pinyin sequence. The proposed phonetic labeling method will mark the syllable boundaries in an utterance and assign the phonetic symbols to every syllable. Each syllable is partitioned into two sub-syllables, i.e. the initial and the final. These sub-syllables are represented in hidden Markov models. The labeling process is accomplished by using Viterbi decoding procedure. The labeling correctness is also automatically examined.

1. INTRODUCTION

The phonetic labeling is a process to align the phonetic transcripts with their corresponding acoustic boundaries. To avoid this tedious and time-consuming process by manual operation, especially in dealing with a large size of speech database, an automatic or semi-automatic labeling method is desirable. In the phonetic labeling process, segmentation of acoustic signal is essential. Many phonetic labeling methods have been proposed [1-5]. The typical approach is to detect the boundaries of phonetic units based on some acoustic parameters, such as the energy, the pitch, and the spectra. Once the boundaries are detected, a proper phonetic symbol is assigned to each segment according to its acoustic characteristics. For automatic labeling systems, the detection and transcription process is somewhat like the speech recognition process.

This paper presents a phonetic labeling method for MAT (Mandarin speech across Taiwan) speech database using the technique of hidden Markov model (HMM) and Viterbi decoding. Each Mandarin syllable is expressed in terms of sub-syllables, i.e. the right-context-dependent (RCD) initial part and the context-independent (CI) final part. Each sub-syllable is modeled by an HMM. The feature distribution of an HMM state is modeled by a mixture of Gaussian densities. Using Viterbi decoding procedure, we can decode a given utterance into state sequence so that the syllable and sub-syllable boundaries are located. For each decoded CI final segment, the pitch is calculated to determine the tone of syllable.

The correctness of labeling process is automatically examined by calculating the duration of each syllable and the distance between this syllable and its corresponding state models. If the duration is beyond a preset range or the distance is beyond a preset threshold value, the syllable is marked as an incorrect one and further manual correction should be performed. A small set of speech data which have been manually labeled is used to evaluate the accuracy of our proposed method. The experiment shows that the labeling accuracy is about 90% for the allowed tolerance of 16 ms and about 94% for the allowed tolerance of 32 ms. The labeling accuracy is close to the syllable correctness calculated by auto-examination algorithm. This implies that the labeling accuracy can be estimated by auto-examination during the automatic labeling process.

2. MAT PROJECT

MAT (Mandarin speech across Taiwan) project is a speech data collection project to collect Mandarin Chinese through the public switching telephone networks in Taiwan [6]. This project is conducted by a group of researchers from several universities and research organizations. The goal is to generate a speech database of 5000 speakers. The speech signal is stored as a 16-bit linear PCM waveform with 8 kHz sampling rate. The speech materials are designed for generating speech models for speech recognition. The phonetically balanced words and sentences were extracted from two text corpora of 77324 lexical entries and 5353 sentences [7]. Besides, the database also contains the numbers pronounced in different ways, such as date, time, price, telephone number, and car plate. The prompting sheets are generated to guide speakers to input their speech data. In addition to the designated prompting materials, a set of some questions is provided to gather the information of speakers. These include speaker’s gender, age, language background, education level, and residence.

Accompanying with MAT database, an editing program, called VEDITOR, is designed for operating in WINDOWS environment. VEDITOR can display the speech waveform of in a speech data file, as well as the file header parameters, on the monitor screen. Users can edit the file header, modify the length of waveform, and play back the speech in an interactive mode. By the playback function, a trained operator can check and edit the transcribed Chinese characters and Pinyin transcripts for a recorded utterance.

3. PHONOLOGY OF MANDARIN CHINESE

Mandarin is a syllabic and tonal language. Each Chinese character is pronounced as a monosyllable. The Mandarin syllable can be expressed in terms of the initial, the final, and the tone [8]. If the tones are ignored, the syllable is referred as a base syllable. The number of base syllables is 408. The tone can be specified by the pitch contour of a syllable. In many of Mandarin speech recognition systems, the tone recognition and the base syllable recognition are performed separately. The base syllable in partitioned into two parts, i.e. the initial and the final. The initial is a consonant and the final is the vowel portion. Some of syllables may have no initial consonant, and are referred as syllables with null initial. Some finals are not explicitly indicated by a vowel, and are referred as syllables with empty vowels. In Mandarin, there are 21 initials (not including the null initial) and

38 finals (not including two empty vowels). Table 1 shows initials and finals in Mandarin syllables.

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

Table 1. Initials and Finals in Pinyin symbols

In a syllable, the beginning portion of final is a transition portion which is affected by its preceding consonant. The duration of initial is usually short. Therefore, a more reliable phonetic unit to represent the initial is to include the beginning portion of its following vowel. This phonetic unit is referred as a right-context-dependent (RCD) initial. The remained portion of final is considered as a context-independent (CI) final. By this definition, Mandarin speech has 94 RCD initials and 40 CI finals (including 2 empty vowels). For the syllable without initial consonant, the beginning portion of final is a transition from silence to vowel. We call this transition portion the null initial. Totally there are 38 null initials.

4. HIDDEN MARKOV MODELS FOR SYLLABLES

We consider the RCD initials, the CI finals, and the null initials as the basic units of Mandarin speech. These basic units are the sub-syllable units and modeled by HMMs. Besides, a silence state is used to represent the pause between syllables or the silence portions before and after the utterance. The number of states for each basic unit and the total number of states are shown in Table 2.

Basic Unit	Number of States	Number of Units	Total states
RCD initials	3	94	282
CI finals	4	40	160
Null initials	2	38	76
Silence	1	1	1
			519

Table 2. Number of states in the HMMs

The speech signal is sampled in 8 kHz sampling rate. The frame size for signal processing is 256 points (or 32 ms in length) and overlapped by 128 points (or 16 ms in length). 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. The Log-Eng has been normalized by its maximal value in an utterance so that the effect of loudness difference is eliminated. The MFCCs are calculated by applying the fast Fourier transform (FFT) algorithm to each frame so that the spectrum of a framed signal is obtained. The spectrum is partitioned in Mel-frequency scale. This is accomplished by passing the signal through a set of triangular band-pass filters. Then the discrete cosine transform (DCT) algorithm is applied to convert these 20 Mel-frequency spectra in logarithmic values into cepstrum. The first 12 MFCCs as well as 12 delta MFCCs are calculated. Together with a delta Log-Eng and a delta-delta Log-Eng, a set of 26 parameters is used to represent the feature vector of a frame. Hence, an utterance is represented by a sequence of feature vectors.

In this study, the training data are extracted from MAT-160, i.e. the database of 160 speakers. This database contains five sub-databases;

(1) MATDB-1 short spontaneous speech,
(2) MATDB-2 numbers pronounced in five different ways,
(3) MATDB-3 isolated syllables,
(4) MATDB-4 isolated words of 2 to 4 syllables, and
(5) MATDB-5 phonetically balanced sentences.

The speech data from MATDB-4 and MATDB-5 are used for model training. There are totally 35,843 syllables in these two sub-databases. The HMM of each sub-syllable is generated by using these training data. The model of each state is represented as a mixture of 4 Gaussian densities.

5. PHONETIC LABELING PROCESS

In an MAT file, the PCM waveform of utterance is stored in binary format. The file header contains parameters to specify the speaker background, signal condition, and speech quality. The transcripts of utterances are also stored in the file header in terms of Chinese characters and Pinyin symbols. For those read speech, the transcript is automatically obtained from the prompting sheet since speakers are asked to pronounce the words or sentences listed on the prompting sheet. For those spontaneous speech, speakers answer the questions generated by system, and the transcript of each answer is input by an operator. All the transcripts have been checked manually so that we can assume that most of them are correct. If a transcript does not exactly correspond to the speech, this transcription error can be checked during the labeling process.

For an utterance with known transcript, we concatenate the HMMs of sub-syllables to form a sequence of state models. The silence state which can be skipped is inserted between syllables and placed at the beginning and the end of the utterance. By using this state model sequence, we perform Viterbi decoding procedure to align the frames to the state models. Then the boundaries of all the syllables and sub-syllables are located. For each syllable, the syllable duration and the distance between a frame and its corresponding state model are calculated during Viterbi decoding procedure.

In order to check the correctness of labeling process, we examine the syllable duration and its average distance from state models [9]. The syllable duration is defined as the number of frames belonging to this syllable. Let N denote the number of syllables in an utterance and L_i be the duration of i-th syllable, $1 \le i \le N$. The average duration and its standard deviation in an utterance are calculated and denoted as L and ΔL, respectively. The syllable duration must be inside the range;

$$Max(8, L-2\Delta L) \le L_i \le Min(40, L+\Delta L) \qquad (1)$$

Otherwise, we mark this syllable as an incorrect one. Equation (1) has bounded the syllable duration between 128 and 640 ms.

The distance between a syllable and its corresponding HMM is defined as the average distance between the frame and its corresponding state model. Let the distance between a frame and its corresponding state model be d_k. The distance of i-th syllable from its model is calculated by;

$$D_i = \frac{1}{L_i}\sum_k d_k \qquad (2)$$

Similarly, we calculate the average distance between syllable and its HMM and the standard deviation, and denote them by D and ΔD, respectively. The distance between a syllable and its corresponding HMM must be smaller than a threshold value;

$$D_i \leq D + 2\Delta D \qquad (3)$$

Otherwise, this syllable is marked as an incorrect one.

6. EXPERIMENTS

In this study, the training data are extracted from the Mandarin speech databases MAT-160. This database contains utterances provided by 160 speakers. Two sub-databases, MATDB-4 and MATDB-5, are used as training data. There are totally 12955 syllables in MATDB-4 and 22888 syllables in MATDB-5. These training data are used for generating the state models. The databases, Test-500 and Test-1000, which are provided by the speakers different from the speakers in the training data are used for testing. Test-500 contains 4726 syllables and Test-1000 contains 9411 syllables. In order to know the accuracy of labeling process, two manually labeled databases, Manu-100A and Manu-100B, are produced. Manu-100A contains 100 utterances with totally 570 syllables. These utterances are extracted from MATDB-4 and MATDB-5. Manu-100B contains 100 utterances with totally 593 syllables which are extracted from Test-1000. These data files are manually labeled so that the syllable boundaries are exactly marked. Table 3 shows the detail contents in the training and test databases.

Database	Isolated syllables	Isolated words	Sentences	Total syllables
MATDB-4	0	4800	0	12955
MATDB-5	0	0	1599	22888
Test-500	50	150	300	4726
Test-1000	100	300	600	9411
Manu-100A	0	75	25	570
Manu-100B	0	75	25	593

Table 3. Contents in the databases

The experiments are arranged as follows;

(a) Inside test
-- The speech data files in MATDB-4 and MATDB-5 are used for automatic labeling test. This is to verify the correctness of model generation and labeling process. Equations (1) and (3) are used for auto-examination.

(b) Outside test
-- Same process as inside test is performed except the speech data files in Test-500 and Test-1000 are used. This is to estimate the flexibility of proposed labeling process when other speech data files are labeled.

(c) Selected data test
-- Same process as inside test is performed on the speech data files in Manu-100A and Manu-100B. This is to verify auto-examination method as comparing with the result of next experiment.

(d) Accuracy of syllable segmentation
-- The data files in Manu-100A and Manu-100B are used for this experiment. The boundaries marked by our proposed labeling method will be compared with those manually labeled boundaries. The labeling accuracy of a syllable is specified in different allowed tolerances.

The experimental results are shown in Table 4.

Database	MATDB-4	MATDB-5
Number of utterances	4800	1599
Number of syllables	12955	22888
Syllable errors	183	1672
Syllable error rate	1.41%	7.31%

(a) Inside test

Database	Test-500	Test-1000
Number of utterances	500	1000
Number of syllables	4726	9411
Syllable errors	305	765
Syllable error rate	6.45%	8.13%

(b) Outside test

Database	Manu-100A	Manu-100B
Number of utterances	100	100
Number of syllables	570	593
Syllable errors	33	53
Syllable error rate	5.76%	8.94%

(c) Selected data test

Database		Manu-100A	Manu-100B
		Accuracy	
Allowed tolerance	0 ms	70.4%	54.3%
	16 ms	89.0%	89.9%
	32 ms	93.2%	94.6%
	48 ms	96.4%	95.9%
	64 ms	97.6%	96.2%
	80 ms	99.8%	99.9%

(d) Labeling accuracy

Table 4. Experimental results

Table 4(a) shows that the syllable error rate for isolated words is much smaller than that for sentences. 92.69% of syllables in the sentences are considered correct as they are examined by Equations (1) and (3). For isolated words, the correctness can be as high as 98.59%.

Table 4(b) shows that the syllable error rates are 6.45% and 8.13% for Test-500 and Test-1000, respectively. Since this is an outside test, the correctness is somewhat less than the correctness of inside test. The result is quite reasonable. This shows the flexibility of applying this proposed labeling method to large speech database.

Table 4(c) is the test result of two small databases which have been selected for manually labeling. The result of Manu-100A is comparable with the result in Table 4(a) because the data files are from MATDB-4 and MATDB-5. The result of Manu-100B is close to the result of Test-1000 in Table 4(b). The error rate of outside test is greater than that of inside test. The syllable correctness based on auto-examination for Manu-100A and Manu-100B is 94.24% and 91.06%, respectively.

Table 4(d) shows that the labeling accuracy is 93.2% for Manu-100A with allowed tolerance of 32 ms and 89.9% for Manu-100B with allowed tolerance of 16 ms. These two values are close to the syllable correctness shown in Table 4(c) based on auto-examination, i.e. 94.24% and 91.06%, respectively. In other words, the labeling accuracy can be roughly estimated by using the proposed auto-examination algorithm.

7. CONCLUSION

This paper describes a phonetic labeling method for Mandarin speech database. By modeling the sub-syllables of Mandarin speech in hidden Markov models, the labeling process is accomplished by using Viterbi decoding procedure. An auto-examination algorithm is proposed to check the correctness of labeling process. The experimental results show that the proposed labeling method can achieve the labeling accuracy around 90% for allowed tolerance of 16 ms. The labeling accuracy can be roughly estimated by an auto-examination algorithm. This proposed method can be applied to a large speech database.

ACKNOWLEDGMENTS

This research has been sponsored by the National Science Council, Taiwan, under contract numbers NSC-87-2213-E-007-031 and NSC-88-2213-E-007-048.

REFERENCES

[1] N. parlangeau and A. Marchal 1996. AMULET: Automatic multisensor speech labeling and event tracking: study of the spatio-temporal correlations in voiceless plosive production. *Proceedings ICSLP'96*, Philadelphia, PA, USA, vol. 3, pp. 1720-1723.

[2] P. Fetter, U. Haiber, and P. regel-Brietzmann 1997. A low cost phonetic transcription method. *Proceedings EuroSpeech'97,* Rhodes, Greece, vol. 2, pp. 811-814.

[3] O Mella and D. Fohr 1997. Semi-automatic phonetic labelingof large corpora. *Proceedings EuroSpeech'97*, Rhodes, Greece, vol. 4, pp. 1731-1734.

[4] F.C. Chou, C.Y. Tseng, and L.S. Lee 1998. Automatic segmental and prosodic labeling of Mandarin speech database. *Proceedings ICSLP'98,* Sydney, Australia, vol. 4, pp. 1263-1266.

[5] I. Gholampour and K. Nayebi 1998. A new fast algorithm for automatic segmentation of continuous speech. *Proceedings ICSLP'98,* Sydney, Australia, vol. 4, pp. 1555-1558.

[6] H.C. Wang 1997. 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.

[7] C.Y. Tseng 1995. A phonetically oriented speech database for mandarin Chinese. Proceedings ICPhS'95, Stockholm, Sweden, vol. 3, pp. 326-329.

[8] C.N. Li and S.A. Thompson 1981. *Mandarin Chinese: A functional reference grammar*, University of California Press.

[9] H.C. Wang and R.L. Chiou 1998. A semi-automatic phonetic labeling method for MAT speech database. *Proceedings First International Workshop on East-Asian Language Resources and Evaluation,* Tsukuba, Japan, pp. 173-178.