A MODEL-BASED TONE LABELING METHOD FOR MIN-NAN/TAIWANESE SPEECH

Wei-Chih Kuo, Yih-Ru Wang and Sin-Horng Chen
Department of Communication Engineering
Chiao Tung University
Hsinchu, Taiwan 300, ROC
tel: +886-3-5731822, fax: +886-3-5710116 ext 31822
schen@cc.nctu.edu.tw

Abstract

In this paper, a model-based tone labeling method for Min-Nan/Taiwanese speech is proposed. It takes the mean and shape of syllable pitch contour as two modeling units and considers some major affecting factors that control their variations. By using the EM algorithm to estimate all parameters of the pitch mean and shape models from a speech database, we can decide the best tone sequences pronounced in all utterance of the database. Experimental results showed that it outperformed the VQ classification method which suffers from the interferences resulted from neighboring syllables and from the global prosodic phrase pattern.

1. INTRODUCTION

Prosodic labeling is to mark significant prosodic cues, such as tones and breaks, in all utterances of a database [1]. A large well-labeled speech database is important for both text-to-speech (TTS) and automatic speech recognition (ASR) studies. In TTS, highly natural prosody generation algorithms can be derived from such a database [2]. In ASR, prosodic cues can be used to help correct acoustic decoding errors [3] or to provide useful information for speech understanding [4]. Prosodic labeling can be done manually by well-trained persons or automatically by models. To perform the prosodic labeling of a large database by hand always faces two difficult problems. One is the consistency across the whole database and another is the heavy workload. Consistency of prosodic labeling is usually difficult to maintain when the work is done by several persons or when the work lasts for a long time period. This is especially true when the prosodic cue to be labeled can be affected by the interaction of several linguistic features simultaneously. On the other hand, automatic prosodic labeling needs a sophisticated mathematic model to consider various affecting factors that contribute the variability of the prosodic cue.

The paper addresses the issue of tone labeling of a Min-Nan/ Taiwanese speech database. A well-labeled database should be very beneficial to the learning of word chunking rules in text analysis as well as to the training of a prosodic generating algorithm for TTS. But due to the following two facts, this is a tough task in Min-Nan language processing. One is that syllables can change their tones drastically in the spoken Min-Nan language, and another is that Min-Nan is a colloquial language and does not have a standard written form. In this study, a model-based approach is adopted to solve the problem.

The paper is organized as follows. Section 2 gives a brief introduction to the background knowledge of the Min-Nan/ Taiwanese language. Section 3 presents the proposed tone labeling method. Experimental results to evaluate the performance of the method is discusses in Section 4. Some conclusions are given in the last section.

2. The Min-Nan/Taiwanese Language

Min-Nan is a spoken dialect widely used in the south-eastern China and Taiwan. Just like Mandarin, Min-Nan speech is a syllabic and tonal language. Syllable is the basic pronunciation unit. There exist more than 2000 syllables. Each syllable can be divided into two parts: a base-syllable and a tone. There are only 877 base-syllables and 8 tones including a degenerated one which is not used in modern Taiwanese. The base-syllables have similar initial-final structure like Mandarin base-syllables except that some finals can have stop endings (entering tones). There are 18 initials and 82 finals. Although tonal syllables are the basic pronunciation units, word is the smallest meaningful unit in syntax. Words are composed of several syllables and sentences are formed by concatenating words.

Although Min-Nan/Taiwanese speech has similar linguistic characteristics like Mandarin speech, it is a colloquial language and does not have a standard written form. There exist two popular written forms in Taiwan. One is the Romanization form which uses Roman alphabets to spell each base-syllable and uses a number to specify its tone. This representation is referred to as “羅馬拼音 (Romanization)”. It was used to write the Taiwanese Bible and is used in some specific societies. Its main drawback lies in the difficulty of understanding the text without reading it out. The other, referred to as “漢羅拼音 (mixed script of Han and Roman characters)”, is a hybrid one in which most syllables are represented by Chinese characters with only a small set of special syllables being represented in Romanization form. Text written in this representation is easier to understand so that it is widely used in writing books and text documents. Unfortunately, the system to represent words in Chinese characters is still not standardized nowadays in Taiwan. Except some popular words, people always choose, with their own preference, a string of Chinese characters with similar pronunciations in Mandarin to represent a Min-Nan/Taiwanese word. This makes the text analysis very difficult for Min-Nan/Taiwanese language because of the lack of a standard lexicon.

It is worth to note that Min-Nan/Taiwanese language has two pronunciation styles. The first one, referred to as “白話 (colloguial)”, is widely used for daily conversation. The second, referred to as “文言 (literary)”, is restrictedly used in reading poetry, some numbers, names and so on.

0-7803-8484-9/04/$20.00 ©2004 IEEE

Although there are only 7 lexical tones, the tone pattern of a syllable may change seriously in continuous speech. This is known as tone sandhi. A previous study showed that Min-Nan speech possesses a set of tone sandhi rules [5]. Generally, all syllables except the last one of a word chunk have to change their tones according to the following rules：

$$\begin{aligned} &1 \rightarrow 7 \\ &7 \rightarrow 3 \\ &3 \rightarrow 2 \\ &2 \rightarrow 1 \\ &5 \rightarrow \begin{cases} 7 & \text{south} \\ 3 & \text{north} \end{cases} \\ &4\ (p,t,k) \leftrightarrow 8\ (p,t,k) \\ &4h \rightarrow 2 \\ &8h \rightarrow 3 \end{aligned} \tag{1}$$

Here an arrow indicates the way of tone change, e.g., Tone 2 will change to Tone 1; “north” and “south” mean the northern and southern parts of Taiwan; and “p”, “t”, “k”, and “h” represents the ending phones of entering tones. Besides, four additional rules [5] are used for the cases when a syllable preceding the special character “仔” (/a/) has been changed to Tone 2 or 3：

$$\begin{aligned} &7 \rightarrow 3 \rightarrow 7 \\ &8h \rightarrow 3 \rightarrow 7 \\ &3 \rightarrow 2 \rightarrow 1 \\ &4h \rightarrow 2 \rightarrow 1 \end{aligned} \tag{2}$$

For instances, 鋸(ki3→ki)仔 and 葉(hioh8→hioh7)仔. But there still exists a problem of applying these rules to an input text to obtain the correct tone sequence pronounced in the associated speech. It is that the way to form word chunks from a word sequence is not exactly know.

3. The Proposed Approach of Tone Labeling

The task of tone labeling is to determine the tone sequence pronounced in each utterance of a speech database. Several approaches can be employed to tackle the task. Firstly, a direct one is to do the job manually by hearing and/or by observing the pitch contour. But the approach will suffer from the difficulties of inconsistency and heavy workload as mentioned above. Another approach is to determine the tone sequence by applying the above tone sandhi rules to the associated text. The main problem of the approach is that the way to automatically form word chunks from the word sequence is not know exactly. Besides, determine tones only from texts may suffer from errors. The third approach is to regard it as a classification problem by classifying the pitch contours of all syllables with the same lexical tone using the unsupervised clustering technique such as vector quantization (VQ). A drawback of the approach is that errors may occur because the pitch contour of a syllable in a continuous speech is influenced by many affecting factors other than the tone itself. The fourth approach is to tackle the task by an efficient pitch contour model which can separate all major affecting factors that control the variation of the pitch contour.

In this study, we adopt the last approach by using a statistical pitch contour model [6]. We first represent the pitch contour of each syllable by using a 3-rd order orthogonal polynomial expansion [7]. The basis polynomials used are normalized, in length, to [0,1] and can be expressed as

$$\begin{aligned} &\phi_0(\tfrac{i}{M}) = 1 \\ &\phi_1(\frac{i}{M}) = [\frac{12 \cdot M}{M+2}]^{1/2} \cdot [\frac{i}{M} - \frac{1}{2}] \\ &\phi_2(\tfrac{i}{M}) = [\tfrac{180 \cdot M^3}{(M-1)(M+2)(M+3)}]^{1/2} \cdot [(\tfrac{i}{M})^2 - \tfrac{i}{M} + \tfrac{M-1}{6 \cdot M}] \\ &\phi_3(\frac{i}{M}) = [\frac{2800 \cdot M^5}{(M-1)(M-2)(M+2)(M+3)(M+4)}]^{1/2} \\ &\qquad \cdot [(\frac{i}{M})^3 - \frac{3}{2}(\frac{i}{M})^2 + \frac{6M^2-3M+2}{10 \cdot M^2}(\frac{i}{M}) - \frac{(M-1)(M-2)}{20 \cdot M^2}] \end{aligned} \tag{3}$$

for $0 \le i \le M$, where M+1 is the length of the current syllable log-pitch contour and $M \ge 3$. They are, in fact, discrete Legendre polynomials. A syllable pitch contour $f(\frac{i}{M})$ can then be approximated by

$$\hat{f}(\tfrac{i}{M}) = \sum_{j=0}^{3} \alpha_j \cdot \phi_j(\tfrac{i}{M}) \qquad 0 \le i \le M\,, \tag{4}$$

where

$$\alpha_j = \tfrac{1}{M+1} \sum_{i=0}^{M} f(\tfrac{i}{M}) \cdot \phi_j(\tfrac{i}{M}) \tag{5}$$

The four coefficients are then divided into two parts: α_0 representing the mean and $\begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{bmatrix}$ representing the shape. They are separately modeled. The pitch mean model used can be expressed by

$$Y_n = F_n + \beta_{pt_n} + \beta_{t_n} + \beta_{ft_n} + \beta_{p_n} \tag{6}$$

where Y_n is the observed pitch mean α_0 of the nth syllable; F_n is the normalized pitch mean and is modeled as a normal distribution with mean μ and variance ν; β_r is the compressing-expanding factor (CF) for affecting factor r; t_n, pt_n and ft_n represent respectively the lexical tones of the current, previous and following syllables; and p_n represents the prosodic state of the current syllable. Here prosodic state roughly represents the state of the syllable in a prosodic phrase and is treated as hidden. Note that t_n ranges from 1 to 22 including 7 standard patterns of lexical tones and all their sandhi tones, while both pt_n and ft_n ranges from 0 to 22 with 0 denoting the cases of major punctuation marks {，。,！,；,？、,：} or the non-existence of the previous or following syllable. The CFs for $pt_n = 0$ and $ft_n = 0$ are set to zero because we do not want to count the affection of tone across punctuation mark.

The pitch shape model used can be expressed by

$$\mathbf{Z}_n = \mathbf{X}_n + \mathbf{b}_{pt_n} + \mathbf{b}_{t_n} + \mathbf{b}_{ft_n} + \mathbf{b}_{p_n} \tag{7}$$

where $\mathbf{Z}_n$ is the observed shape vector $\begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{bmatrix}^T$ of the nth syllable's pitch contour; $\mathbf{X}_n$ is the normalized pitch shape vector and is modeled as a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix R.

To estimate the parameters of these two models, an EM algorithm is adopted. The EM algorithm is derived based on the maximum likelihood (ML) estimation from incomplete data with prosodic state and pronounced tone pattern being treated as hidden or unknown. To illustrate the EM algorithm, an auxiliary function is firstly defined in the expectation step (E-step) as

$$Q(\bar{\lambda},\lambda) = Q_1(\bar{\lambda}_1,\lambda_1) + Q_2(\bar{\lambda}_2,\lambda_2) \tag{8}$$

where

$$Q_1(\bar{\lambda}_1,\lambda_1) = \sum_{n=1}^{N}\sum_{p_n=1}^{P}\sum_{t_n} p(p_n,t_n \mid Y_n,\bar{\lambda}_1)\log p(Y_n,p_n,t_n \mid \lambda_1)\ , \tag{9}$$

$$Q_2(\bar{\lambda}_2,\lambda_2) = \sum_{n=1}^{N}\sum_{t_n} p(p_n,t_n \mid \mathbf{Z}_n,\bar{\lambda}_2)\log p(\mathbf{Z}_n,p_n,t_n \mid \lambda_2)\ , \tag{10}$$

N is the total number of training syllables, P is the total number of prosodic states, $p(p_n,t_n \mid Y_n,\bar{\lambda}_1)$, $p(Y_n,p_n,t_n \mid \lambda_1)$, $p(p_n,t_n \mid \mathbf{Z}_n,\bar{\lambda}_2)$ and $p(\mathbf{Z}_n,p_n,t_n \mid \lambda_2)$ are conditional probabilities, $\lambda = \lambda_1 \cup \lambda_2$, $\lambda_1 = \{\mu,\nu,\beta_t,\beta_{pt},\beta_{ft},\beta_p\}$ and $\lambda_2 = \{\boldsymbol{\mu},\mathbf{R},\mathbf{b}_{pt},\mathbf{b}_t,\mathbf{b}_{ft},\mathbf{b}_p\}$ are the sets of parameters to be estimated, and λ and $\bar{\lambda}$ are respectively the new and old parameter sets. Based on the assumption that the normalized pitch mean F_n and shape $\mathbf{X}_n$ are both normally distributed, $p(Y_n,p_n,t_n \mid \lambda_1)$ and $p(\mathbf{Z}_n,p_n,t_n \mid \lambda_2)$ can be derived from the assumed model given in Eqs.(6) and (7) and expressed by

$$p(Y_n,p_n,t_n \mid \lambda_1) = N(Y_n;\mu + \beta_{pt_n} + \beta_{t_n} + \beta_{ft_n} + \beta_{p_n},\nu)\ , \tag{11}$$

and

$$p(\mathbf{Z}_n,p_n,t_n \mid \lambda_2) = N(\mathbf{Z}_n;\boldsymbol{\mu} + \mathbf{b}_{pt} + \mathbf{b}_t + \mathbf{b}_{ft} + \mathbf{b}_p,\mathbf{R}) \tag{12}$$

Similarly, $p(p_n,t_n \mid Y_n,\bar{\lambda}_1)$ and $p(p_n,t_n \mid \mathbf{Z}_n,\bar{\lambda}_2)$ can be expressed by

$$p(p_n,t_n \mid Y_n,\bar{\lambda}_1) = \frac{p(Y_n,p_n,t_n \mid \bar{\lambda}_1)}{\sum_{p'_n=1}^{P}\sum_{t'_n} p(Y_n,p'_n,t'_n \mid \bar{\lambda}_1)}\ , \tag{13}$$

and

$$p(p_n,t_n \mid \mathbf{Z}_n,\bar{\lambda}_2) = \frac{p(\mathbf{Z}_n,p_n,t_n \mid \bar{\lambda}_2)}{\sum_{p'_n=1}^{P}\sum_{t'_n} p(\mathbf{Z}_n,p'_n,t'_n \mid \bar{\lambda}_2)} \tag{14}$$

Then, sequential optimizations of these parameters can be performed in the maximization step (M-step). At the end of each iteration, the pronounced tone pattern for each syllable are re-assigned to one of its possible patterns by

$$t_n^* = \arg\max_{t_n} p(t_n \mid Y_n,\lambda_1)p(t_n \mid \mathbf{Z}_n,\lambda_2) \tag{15}$$

To execute the EM algorithm, initializations of the parameter set $\bar{\lambda}$ are needed. This can be done by estimating each individual parameter independently. Specifically, the initial CF for a specific value of an affecting factor is assigned to be the difference of the mean (mean vector) of Y_n $(\mathbf{Z}_n)$ with the affecting factor equaling the value to the mean of all Y_n $(\mathbf{Z}_n)$. Notice that, in the initializations of CFs for prosodic states, each syllable is pre-assigned a prosodic state by vector quantization. After initializations, all parameters are sequentially updated in each iteration. The iterative procedure is continued until a convergence is reached.

4. Experimental Results

Performance of the proposed model-based Min-Nan tone labeling method was examined by simulation using a single male speaker database. The database contained 255 utterances including 130 sentential utterances with length in the range of 5-30 syllables and 125 paragraphic utterances with length in the range of 85-320 syllables. The total number of syllables was 23,633. All speech signals were digitally recorded in a 20 kHz rate. All speech signals and the associated texts were manually pre-processed in order to extract the required acoustic features and linguistic features.

Four tone labeling methods were then realized and compared. The first one was the manual approach which determined the tone sequence to be pronounced by examining the text. Although the results might contain some errors, we still took them as the reference target because of the lack of a better one. It is referred to as MANUAL. Another two were the VQ-based methods which used 4 (mean + shape) and 3 (shape) orthogonal expansion coefficients of syllable pitch contour as classification features, respectively. They are referred to as VQ-4 and VQ-3. The last one was the proposed model-based method and referred to as MODEL. The RMSEs of the reconstructed pitch contour are 0.815 and 0.286 ms/frame for VQ-4 and MODEL, respectively. Better results of MODEL show the effectiveness of the pitch mean and shape models. Table 1 shows the correct rates of tone labeling for the latter three methods by taking the results of MANUAL as reference target. Correct rates of 50.9, 52.4, and 61.9% were obtained by VQ-4, VQ-3, and MODEL, respectively. Obviously, MODEL outperformed both VQ-4 and VQ-3. It can also be found from Table 1 that Tone 1 and Tone 2 which have a single sandhi tone pattern have better labeling results.

By examining all 22 tone patterns obtained in the pitch mean and shape models, we found that most sandhi tone patterns matched with those tone patterns suggested by the above-mentioned sandhi rules. Figs.1 displays the standard and sandhi tone patterns for lexical Tone 1 and Tone 2. Can be seen from Fig.1(a) (Fig.1(b)) that the shape of the sandhi tone pattern of Tone 1 (2) resemble to the standard pattern of Tone 7 (1). Fig.2 dispalys pitch contour patterns of standard and sandhi tones for Tone 3 and Tone 2. Can be seen from the Fig.2(a) (Fig.2(b)) that all three (two) sandhi Tone 3 (2) patterns resemble to the standard Tone 3 (2) pattern.

5. Conclusions

In this paper, a model-based tone labeling method for Min-Nan/Taiwanese speech has been discussed. It employed a statistical modeling technique to separate some major affecting factors that influence syllable's pitch contours. By the model, we can determine the best tone pattern pronounced in each syllable from its pitch contour with interferences from both neighboring syllables and the global effect of prosodic phrase being eliminated. Experimental results confirmed that it outperformed the VQ classification method. So it is a promising method.

Acknowledgement

This work was supported in part by MOE under contract EX-92-E-FA06-4-4 and in part by NSC under contact NSC92-2213-E-009-128. The authors thank Prof. R. L. Cheng for supplying the lexicon and the text corpus.

References

[1] Li Aijun, "Chinese Prosody and Prosodic Labeling of Spontaneous Speech," Speech Prosody 2002

[2] Fu-chiang Chou, Chiu-yu Tseng, Keh-jiann Chen and Lin-shan Lee, "A Chinese Text-to-Speech Based on Part-of-Speech Analysis, Prosodic Modeling and Non-uniform Units," ICASSP, pp. 923 – 926, 1997

[3] W. J. Wang, Y. F. Liao and S. H. Chen, "RNN-based Prosodic Modeling for Mandarin Speech and Its Application to Speech-to-Text Conversion," Speech Communication 36 (2002) pp.247-265.

[4] Hung-yun Hsieh, Ren-yuan Lyu and Lin-shan Lee, "Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary," ICSLP, vol. 2, pp. 809 – 812, 1996.

[5] R. L. Cheng, Taiwanese pronunciation and Romanization – with rules and examples for teachers and students, Wang Wen Publishing Company, 1993

[6] S. H. Chen, W. H. Lai and Y. R. Wang, "A New Pitch Modeling Approach for Mandarin Speech," submitted to J. Acoust. Soc. Am.

[7] S. H. Chen and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," IEEE Tarns. Communications, vol.38, no.9, pp.1317-1320, Sep 1990.

Table 1 The correct rates of the three tone labeling methods of VQ-4, VQ-3, and MODEL. (unit: %)

Tone (sandhi tones)	1 (7)	2 (1)	3 (2,1)	4 (2,1,8)	5 (7,3,7)	7 (3,7)	8 (3,7,4)	Ave.
VQ-4	61.9	82.9	55.4	40.9	28.1	34.0	33.9	50.9
VQ-3	58.7	84.8	44.1	28.7	43.7	47.2	35.8	52.4
MODEL	72.4	89.3	51.7	55.7	50.6	51.1	41.9	61.9

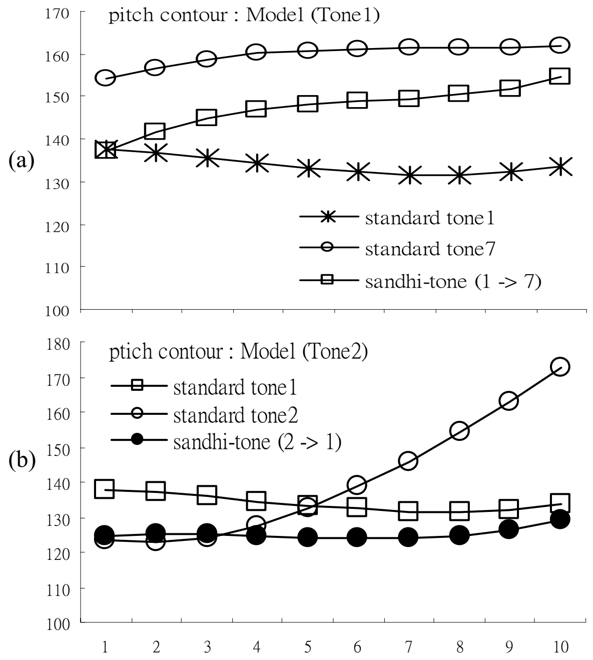

Fig. 1 Comparison of standard and sandhi tone patterns for lexical (a) Tone 1 and (b) Tone2.

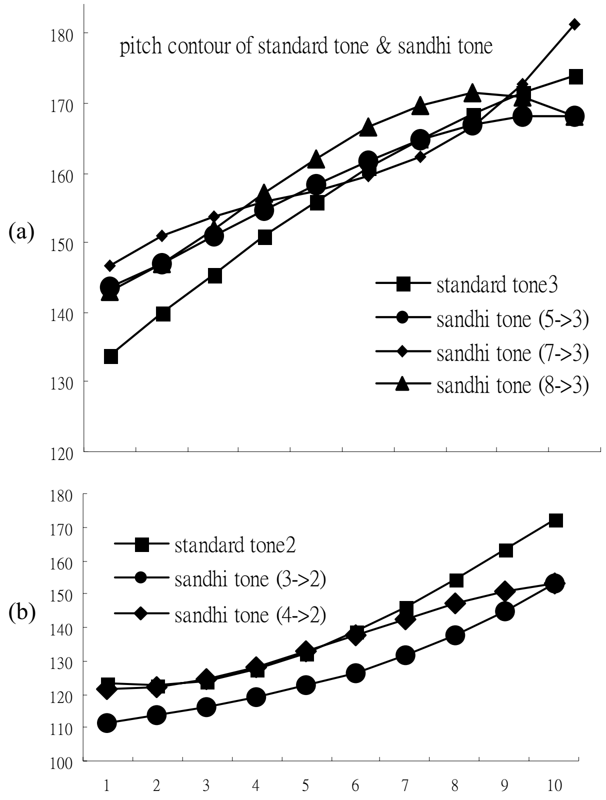

Fig. 2 Comparison of pitch contour patterns of standard tone & sandhi tones for (a) Tone 3 and (b) Tone 2.