# TONE RECOGNITION IN MANDARIN SPONTANEOUS SPEECH

*Zhaojie Liu*[1,2] *Pengyuan Zhang*[1] *Jian Shao*[1]
*Qingwei Zhao*[1] *Yonghong Yan*[1] *Ji Feng*[2]
1 ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences
2 Institute of Physics, Chinese Academy of Sciences
zliu@hccl.ioa.ac.cn

## Abstract

This paper reports our study on tone recognition in Mandarin spontaneous speech, which is characterized by complicated tone behaviors. Real-Context is proposed as a new concept used in the tone modeling. First, the "error" data, which may bring negative influences to the tone model, are removed from the training data by an iterative method. Then we cluster the reduced training data into a few subsets to generate a more refined tone model. Gaussian Mixture Model (GMM) is used for the tone modeling. All experiments are based on the spontaneous speech database, Train04. Experimental results demonstrate the effectiveness of the methods.

## 1. Introduction

Mandarin is a kind of tonal languages. Its words are composed of one or multiple mono-syllable units called characters, and each Chinese character corresponds to a syllable associated with a lexical tone. Syllables or words with the same sequence of consonants and vowels have different tones. Usually, Mandarin contains five tones, characterized by syllable-level pitch or fundamental frequency (F0) contour pattern: high-level (tone 1), high-rising (tone 2), low-dipping (tone 3), high-falling (tone 4) and neutral (tone 5). The neutral tone often occurs in word-end or sentence-end contexts in continuous speech and does not have a stable F0 contour, which is not considered in this paper.

Accurate tone recognition plays an important role in automatic Mandarin speech recognition. Although tone recognition has been investigated for many years, relatively high recognition accuracy are only obtained in isolated words and reading speech [1][2][3][4][5][6]. Various pattern recognition methods were applied to tone recognition, which including Hidden Markov Models (HMM) [4], Gaussian Mixture Model (GMM) [7], Decision-tree Classification [8], and Support Vector Machine (SVM) [9]. Approaches fall into two major categories, namely, embedded tone modeling and explicit tone modeling. In embedded tone modeling, Pitch-related features can be added as extra dimensions in the short-time acoustic feature vector. Tone recognition is done as an integral part of the existed system. On the contrary, in explicit tone modeling, tones are independently modeled and recognized in parallel to the recognition of acoustic units. Then the results are combined in a post-processing stage [7]. Since pitch is a supra-segmental feature, which is spanned across segments and lay on a group of voiced segments, the explicit tone modeling may be more effective for modeling tone variations. For instance, the context-tone concept and supra-tone are proposed to model tone explicitly.

Spontaneous speech, as opposed to planned speech, is a more natural way in which people communicate with each other. It usually contains mispronunciations, emotional status, and other unlinguistic utterances. Besides, the speaking rate is relatively fast, leading to more serious articulation. All the phenomena mentioned above would cause tone contours to deviate from their canonical patterns and bring challenges to the tone recognition. Up to now, few efforts have been made on this respect. Recently, [10] performed some experiments to compare the capability between tone context independent and tone context dependent phoneme sets with embedded tone modeling. The purpose of this paper is to generate a refined tone model which can better describe the complex tone patterns in spontaneous speech. A new context unit is proposed to model tonal context influences, and then we cluster the samples so as to gain actual tone patterns. A kind of similar distance between two tones is also used.

The remainder of the paper is organized as follows. We first describe the tone feature and the model used in this paper. Then, the strategy of reducing "error" data in the training data is illustrated in section 3. In section 4, we introduce a new Real-Context concept and briefly present the clustering process. Some experiment results are given in section 5. And finally, section 6 provides the conclusions and some discussions.

## 2. Tone feature and tone model

F0 is one of the most important features of Mandarin tone. Although energy and voicing also carry some cues for tone, the cues are not as obvious as that of F0, especially for continuous speech. In present analysis, we mainly use F0 and its first derivatives as tone features. Tone is realized primarily by the F0 movement across the voiced portion of a syllable. F0 contours of four lexical tones are shown in Fig. 1, which are computed by averaging over 1000 utterances spoken by male speakers. These utterances cover most of the tonal syllables used in Mandarin.

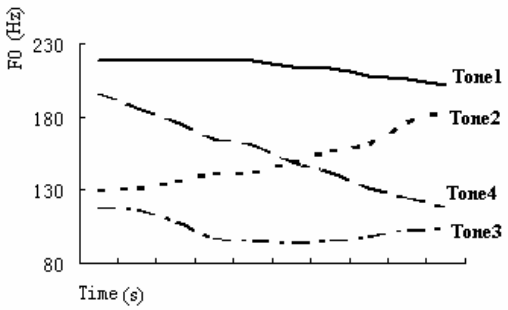


*Figure 1:* F0 contour of four lexical tones

### 2.1. Pitch extraction

Previous research suggested that the height and shape of the F0 contours, rather than the exact values at individual points, are critical for the recognition of Mandarin tones. Therefore, it is unnecessary to design a complex system to extract pitch accurately for the tone feature. A fast and robust pitch tracking algorithm (RAPT) [11] is used in our work. Two-pass normalized cross correlation function (NCCF) is calculated to generate F0 candidates, which is expressed as:

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}} \quad (1)$$

$$(k = 0,1,...,k-1; i = 0,1,...,M-1)$$

Where

$$e_m = \sum_{l=m}^{m+n-1} s_l^2$$

$s_j$ is a sampled speech signal; $i$ is the frame index; $k$ is the lag; $n$ is the sample number in an analysis window; $m$ is the sample number in a frame.

### 2.2. F0 normalization

F0 is a highly variable acoustic feature affected by a number of linguistic and extra-linguistic factors. The dynamic range of F0 greatly depends on the speaker's gender, age and physiological characteristics. It is also related to the speaker's physical conditions, speaking style and emotional status. In Fig. 2, an example of tone contour in spontaneous speech is demonstrated. Within the sentence, F0 spans a range of more than 50Hz. [10] reported that the pitch of an individual adult speaker can range from 100 to 300Hz. Therefore, F0 should be normalized by the speaker independent system. In this paper, F0 is normalized by the following method:

$$F_i = K*(F0_i - \min F0_i)/(\max F0_i - \min F0_i) \quad (2)$$

where $\min F0_i$ and $\max F0_i$ are the minimum and maximum F0 of a sentence and $K$ is a constant.

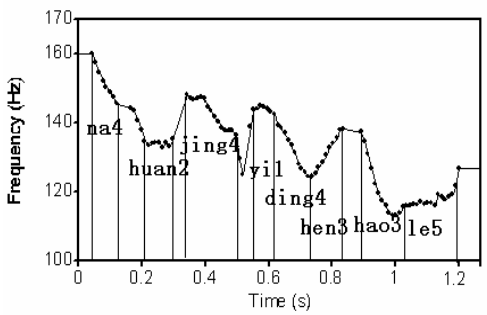


*Figure 2:* Pitch contour in the spontaneous speech

### 2.3. Tone model

In Mandarin syllable, the Final is regarded as voiced whereas the Initial is either voiced or unvoiced. Hence, F0 features for tone recognition are only extracted from the Final segments. In present study, the feature vector, which is represented by a fixed-length vector, contains F0 and its first derivatives. Every unit is divided evenly into three portions, and then the mean value is computed from each portion. Our tone model is trained by using tone labels provided by force-alignment results that should reflect the actual tone pronunciation by the speakers instead of the canonical tone marks associated with the character. Initial/Final boundaries are obtained by aligning the syllable labels with the acoustics model.

Gaussian mixture model (GMM) is employed for our tone modeling. It provides a probabilistic output that can be readily integrated into HMM based ASR systems.

The expectation-maximization (EM) algorithm is used for the estimation of GMM parameters.

## 3. Reducing the "error" data

Except for filled pauses, repairs, hesitations, repetitions, and disfluencies, spontaneous speech also contains other unlinguistic utterances and many noises. All of these phenomena are the main factors that may cause a very low performance of our ASR recognizer and imprecise cutting of Initial/Final boundaries. To make the tone model less affected by the "error" data, we first set a proper threshold to keep the results with high confidence in Initial/Final segmentation, and then use an iterative method to remove the data with very low scores. The methods are presented as follows:

1. Initialize the training data with high confident recognition results.
2. Train the tone model using the training data.
3. Recognize every sample in the training data. The sample, which is recognized to its labeled tone with the lowest score, will be removed from the training data.
4. Decide whether the result satisfies convergence requirement. If it does not, repeat 2 and 3 until convergence

## 4. Tone variations modeling

As we know, the detection of tone variations is a key point in tone recognition. In spontaneous speech, tone variations are too complicated in actual pronunciations to be described by linguistic rules. Indeed, accurate description of tone variations is the precondition of the accurate tone recognition. A great number of works have discussed this respect [10][12]. For example, [12] suggested that unit selection strategy is needed to extend to incorporate tonal context. Their statistical results showed that the influence of the left tone context is greater than the right one. Through analyzing the results, we find that samples of the tone variations account for a fair proportion in the whole database, which would lead to confusion in tone recognition. In order to devise such a strategy, Model units from left-context (L-C) to super-tone are used in this paper. Further, Real-Context (Real-C) is first proposed to model the tone variations, which is defined as follows:

$$\begin{cases} preF - \beta > 0, context \quad is \quad defined \quad H \\ preF - \alpha < 0, context \quad is \quad defined \quad L \\ other, defined \quad M \end{cases} \tag{3}$$

where $\alpha$ and $\beta$ are set in advance, $preF$ is the real pitch position of pre-tone last dimension vector. The context will become H-*, M-*, L-*, each of which is a real context. * indicates the current tone. Real-Context, which is opposite to the tonal context, will reflect the context influence of the pre-tone pitch position in reality. And Real-Context has fewer feature vectors than supra-tone. Fig. 3 shows the difference among them. 1 denotes context tone model, 2 is supra-tone model and 3 is Real-Context model.

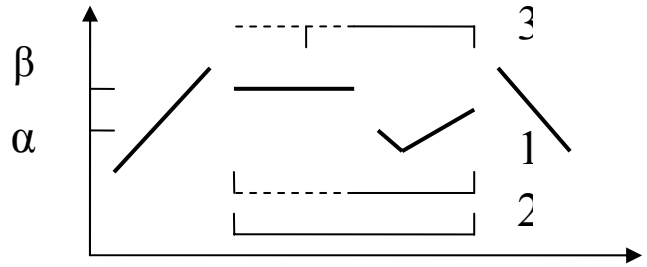


*Figure 3:* Tone Units

In current study, the method based on clustering is presented to model tone explicitly. Before that, we employ a kind of similarity measurements to compute the distance between two tones, which should reveal the similarity of their pitch contour shapes, and also include their different pitch height. The distance is defined:

$$Dis(X,Y) = \sqrt{\sum_{i=0}^{n} ((x_i - \bar{x}) - (y_i - \bar{y}))^2} + \left|\bar{x} - \bar{y}\right|$$

$$X = (x_0, x_1, \ldots, x_n), \qquad \bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{4}$$

$$Y = (y_0, y_1, \ldots, y_n), \qquad \bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$$

In the clustering process, we first classify the tone data into four groups based on their labels. Second, the hierarchical clustering technique is used to find pitch contour patterns of each tone group. Initially, all samples in the group are assigned to their own clusters. Then the algorithm proceeds iteratively, joining two most similar clusters at one stage, and continuing until the condition is satisfied. The advantage of this method is that the number of clusters does not have to be fixed in advance.

## 5. Experiments

For all the experiments report in this paper, we use the Mandarin CTS data collected by Hong Kong University of Science and Technology in 2004, made within China and Hong Kong by mostly college students. The training set, a part of the train04, contains phonetically rich training utterances spoken by male and female speakers. The testing set comprises about utterances which are unseen from the training set. The contents are given in table 1.

*Table1:* Speech database used in this paper

| | Num of sentences | Num of speakers | |
|---|---|---|---|
| | | male | female |
| training | 13046 | 50 | 50 |
| testing | 2729 | 10 | 10 |

Our recognition system is HMM-based. The acoustic models are Initial/Final models with both left and right context dependency. The acoustic feature vector is composed of 12 MFCC plus energy, and their first and second order derivatives. The recognition accuracy for base syllable is 58.9% for all the train utterances.

We perform experiments with different tone units. Table 2 shows the tone recognition results of different models. The baseline (L-C) overall accuracy is 40.8%. The best performance, 43.0%, has been attained for tone 4(T4), which has the highest percentage of distribution among all tones. Tone 3(T3) gets the lowest accuracy of 36.7%. Experiments have also been done with right-context, and the results become a little worse. It is also revealed that the Real-C outperforms L-C with 2.3% absolutely and is slightly better than supra-tone models (di-tone).

*Table 2:* Tone recognition accuracy with different models

| Unit | T1(%) | T2(%) | T3(%) | T4(%) | total(%) | gain |
|---|---|---|---|---|---|---|
| L-C | 42.4 | 39.6 | 36.7 | 43.0 | 40.8 | ref |
| di-tone | 45.6 | 40.3 | 38.4 | 45.5 | 42.9 | 2.1 |
| Real-C | 46.2 | 40.5 | 38.1 | 45.8 | 43.1 | 2.3 |

The experimental results of reducing the "error" data and clustering are given in Table 3. Improvement of 3.3% by reducing the "error" data can be attained, but clustering does not exhibit noticeable performance improvement.

*Table 3:* Reducing the "error" data and clustering

| Unit | Reducing the "error" data(gain) | Clustering (gain) |
|---|---|---|
| di-tone | 45.4 (2.5%) | 46.1 (0.7%) |
| Real-C | 46.4 (3.3%) | 46.7 (0.3%) |

## 6. Conclusion and discussion

In this paper, we have explored tone recognition in Mandarin spontaneous speech. The proposed new Real-Context concept

would be more helpful in modeling the tonal context influence, as shown by more than 2% improvement absolutely. Meanwhile, a refined tone model is generated by reducing the "error" data from the original data, improving the tone recognition accuracy of 3.3%. Furthermore, through clustering the training data to subsets, which may accurately describe the tone variations, we can achieve a tone recognition accuracy of 46.7%.

Nevertheless, there are still aspects of modification for the proposed methods in our study. Although tone recognition accuracy with 5.9% has been absolutely improved, it is still much lower than that of base syllable. Therefore, refining the tone modeling still requires intensive analyses, for example, tone variation rules may be added into the clustering process. Further, additional experiments are needed by other different database.

## 7. Acknowledgements

This work is partially supported by Chinese 973 program (2004CB318106), National Natural Science Foundation of China (10574140, 60535030), and Beijing Municipal Science & Technology Commission (Z0005189040391).

## 8. References

[1] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny and K. Shen, "New methods in continuous Mandarin speech recognition", *5th European Conference on Speech and Communication and Technology, Vol. 3, 1543-1546, 1997.*

[2] H. Huang, and F. Seide, "Pitch Tracking and Tone Features for Mandarin Speech Recognition", *Proc. ICASSP 2000, vol.3, 1523-1526, 2000.*

[3] C.J. Chen, H.P. Li, L.Q. Shen, G.K. Fu, "Recognize Tone Languages Using Pitch Information on the Main Vowel of Each Syllable", *Proc. ICASSP 2001, Vol. 1, 61-64, 2001.*

[4] W. J. Yang, J. C. Lee, Y. C Chang, and H. C. Wang, "Hidden Markov Model for Mandarin Lexical Tone Recognition," *IEEE Trans. ASSP, Vol. 36, 988-992, 1988.*

[5] G. P. Kong, S. N. Lu, "A VQ study on pitch models of disyllable in Mandarin". *ACTA ACUSTICA*, 25(2), 2000.

[6] E. Chang, J. L. Zhou, C. Huang, S. Di, K. F. Lee, "Large vocabulary mandarin speech recognition with different approaches in modeling tones", *Proc. ICSLP 2000*, 983-986 2000.

[7] Y. Qian, "Use of tone information in Cantonese LVCSR based on Generalized Character Posterior Probability Decoding," Ph. D Dissertation, *The Chinese University of Hong Kong, 2005*

[8] Cao Yang et al. Decision-tree based Mandarin tone model and its application to speech recognition. *Proc. ICASSP 2000, 1759-1762*.

[9] S. D. Dinoj et al. "Tone Recognition in Mandarin using Focus", *INTERSPEECH, 2005*, 3301-3304, 2005.

[10] J. L. Zhou, Y. Tian, Y. Shi, C. Huang, E. Chang, "Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition", *Proc. ICASSP 2004, 997-1000, ICASSP 2004.*

[11] A.D. Talkin, Speech Coding and Synthesis, *Elsevier Science B.V., Amsterdam,* chapter A robust algorithm for pitch tracking (RAPT), 495-518, 1995

[12] Y. Xu, Q.E. Wang, "Pitch target and their realization: Evidence from Mandarin Chinese", *Speech communication, Vol.33, 319-337, 2001.*