TONE ARTICULATION MODELING FOR MANDARIN SPONTANEOUS SPEECH RECOGNITION

Jian-lai Zhou, Ye Tian, Yu Shi, Chao Huang, Eric Chang

Microsoft Research Asia

{jlzhou, t-yetian, yushi, chaoh, echang}@microsoft.com

ABSTRACT

Tone modeling is an unavoidable problem in Mandarin speech recognition. In continuous speech, the pitch contour exhibits variable patterns, and it is strongly influenced by its tone context. Although several effective methods have been proposed to improve the accuracy for tonal syllables in Mandarin continuous speech recognition, many recognition errors are caused by poor tone discrimination capability of the acoustic model [1][2][3][4]. Furthermore, the case becomes worse for the recognition of spontaneous speech. In this paper, we will report our work on tone articulation modeling. Tone context dependent models are used to model unstable pitch patterns caused by co-articulation in continuous speech. Corresponding acoustic features are investigated as well. Our methods are evaluated on two test sets: one is reading-style speech data, the other is spontaneous. The experimental results show that for the test set of casual speech, the proposed method turns out to be more effective than tone context independent model, while they are comparable for the test set of reading-style speech. Several factors which have potential to improve the proposed method are discussed in the final part in this paper.

1. INTRODUCTION

Mandarin Chinese is a kind of tonal language. Each Chinese character corresponds to a syllable which is associated with a lexical tone. Usually, there are 5 tones for a base syllable. Within a syllable, the vowel part underlies tone information. Syllables with different tone have different meanings. Because the tone plays a role in distinguishing lexical meaning for a family of syllables, tone recognition has been investigated for many years. In all published work, the methods of tone pattern recognition can be categorized into two classes. The first is two-step method, and second one-step [2].

In two-step scheme, the base syllable is recognized using some standard methods first, then the syllable boundary information is used to locate tone area. Hence, tone recognition has been converted into a problem similar with isolated word recognition. Lots of approaches [6][7] including HMM, neural net and other rule based methods can be used for this task. Finally, the output of tone and syllable recognizer is combined to form tonal syllable result. In addition, the scores of tone recognizer can help the base syllable recognizer to optimize the search path [5]. The merit of two-step method is that it is easy to utilize the tone context dependent (CD) model both in training and decoding.

One-step method was proposed in [1], and successfully applied in several commercial systems [1][3]. Further, in [2], one-step method was improved by the main vowel concept. It addressed the problem of large size of the phoneme set for some tonal languages. Via the main vowel method, the number of tonal phonemes in a phoneme set got reduced. The experiments exhibited that the speech recognition system can achieve better performance using a smaller tonal phoneme set. One-step method is more suitable in continuous speech recognition because patterns of pitch contour observed in isolated word change strongly. In casual continuous speech, co-articulation phenomenon becomes more serious. Although some phonetic rules are designed to describe these tone articulation phenomena, in statistical modeling methods, it's hard to combine these rules in a systematic framework.

An obvious idea to model tone articulation is using bi-tone or tri-tone concept, which is similar with tri-phones. But it will result in a very large phoneme set. For example, the phoneme set introduced in [3] contains 157 tonal finals. If we just consider the left tone context to expand them in bi-tone models, there should be about 157*5=785 tonal phonemes. It is impossible to build a state-of-the-art speech recognition system based on such a large phoneme set. The derived tri-phones must be represented by astronomical parameters. Even though main vowel method [2] can be used to reduce the size of the phoneme set, the resulted number of phonemes is 263. It is still too large to make the model fully estimated by limited training data.

With the assistance of phonetics, an abstract representation of a tonal syllable in Mandarin was proposed [11]. This concept results in a compact tonal phoneme set of Mandarin. Based on that, a phoneme set which can model tone articulation is introduced in this paper. A pitch and duration based feature set is also used, which helps tonal syllable recognition together with conventional MFCC feature.

2. A NEW REPRESENTATION APPROACH OF THE MANDARIN TONE

In current tone recognition scheme, Mandarin tone is categorized as five classes from tone 1 to tone 5. Both one-step and two-step methods adopted this representation approach. On the other hand, in conventional Chinese phonetics research [8],

four basic tones can be represented by its onset and offset pitch values as “High” and “Low”, as shown in Fig. 1.

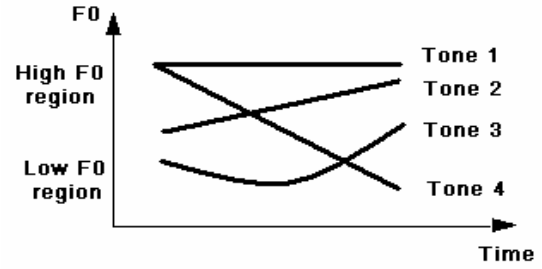

Figure 1: Typical tone patterns for four basic tones

Hence, four basic tones can be encoded by the combination of “High” and “Low”. Based on this idea, a new Mandarin phoneme set was designed in [11]. In this phoneme set, each tonal syllable was split into three parts: consonant initial, front part vowel and back part vowel. Actually, the glide sound was merged into consonant part to create a new initial. Each final in a syllable was divided into two parts. In Table 1, examples of the decomposition approach are listed.

Table 1: the decomposition of a syllable into a phoneme string

Syllable	Phoneme string		
ba1	b	aH	aH
ba2	b	aL	aH
ba3	b	aL	aL
ba4	b	aH	aL
bian1	bi	aH	nnH
bian2	bi	aL	nnH
bian3	bi	aL	nnL
bian4	bi	aH	nnL

The resulted phoneme set contains 97 phonemes. We did lots of experiments including syllable-loop and large vocabulary recognition to verify that this kind of representation is valid in continuous speech recognition. In Table 2, the syllable loop recognition result was listed. Where, the phoneme set 1 is the traditional tonal phoneme set with initial and final structure [3] containing 187 phonemes, and the phoneme set 2 is from the new representation method containing 97 phonemes. More detailed description and results can be found in [11].

Table 2: the performance of new tone representation

Phoneme set	Number of phones	Error rate	
		Base Syllable	Tonal Syllable
Set 1	187	22.36%	41.60%
Set 2	97	21.5%	43.32%

Here, based on the syllable loop recognition result, at least we can conclude that in acoustic level, the new tone representation is comparable with the traditional one in continuous speech recognition. The further conclusion is that the tone representation using “High” and “Low” pitch position can reflect the nature of tone perception, and it can collaborate with one-step method for Mandarin speech recognition.

3. TONE PATTERNS IN SPONTANEOUS SPEECH

In above analysis, each basic tone was represented by the combination of “High” and “Low”. This is true in isolated speech, but tone patterns are varied in continuous or spontaneous speech due to physiological articulation. The shape of the pitch contour of syllables in spontaneous speech is influenced by its context pitch values.

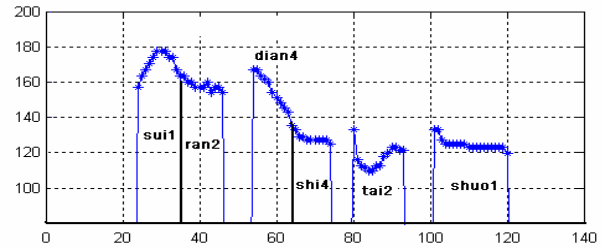

Figure 2: an example of pitch contour in spontaneous speech

In Figure 2, an example of tone articulation phenomenon is shown. Although the second syllable “ran2” should have a contour from a low pitch position to a high position, the virtual contour begins from a relative high position. So the pattern of tone 2 in this case looks like tone4. The reason is that its left tone context is tone 1, and at the ending part of “sui1”, the pitch value still stays in high region. Similar case happens in syllable “shi4”.

There have been some works to discuss this kind of phenomenon [9]. Based on statistical results, the authors suggested that the early portion of the pitch contour of a syllable always varies with the ending pitch value of the preceding syllable, whereas the later position converges to the contour that seems to conform to the purported underlying pitch values. That is to say, the influence of left tone context is greater than right tone context. This assumption will become the basis of our tone articulation modeling scheme. In addition, an approach of using tone critical segments was proposed, in which the pitch contour segments related with transition between two syllables were discarded, only the segments named as “tone nucleus” was used for tone recognition [10]. It made tone recognition more accurate and robust. Since the transition segments will confuse recognizer, and it’s hard to abandon these parts in one-step framework, we need to model this phenomenon in one-step method to improve tone recognition.

4. TONE ARTICULATION MODELING WITH ONE-STEP FRAMEWORK

With the assumption that for a syllable, its left pitch context plays an important role to determine the shape of pitch contour of itself, the simplest way to model tone context is using left tone dependent bi-tone units to form a phoneme set. For example, for a tonal phoneme “a1”, it will have 4 or 5 variations: “a11”, “a21”, “a31”, “a41” and “a 51”. The resulted phoneme set is much larger than the phoneme set reported in all state-of-the-art Mandarin speech recognition systems. The speech recognition prefers a smaller phoneme set because it

can be trained more thoroughly with the same amount of training data.

4.1 The phoneme set

According to the tone representation with "High" and "Low" pitch position described in Section 2, we can find that some effects of tone context are same, and they can be combined into one model. For example, the onset of pitch contour of tone 1 should begin from a "High" position. If its left context is tone 3, tone 4, or tone 5, the pitch contour on the onset should have a transition segment from "Low" to "High" if articulation phenomenon happens. In the same way, if the left context of tone 1 is tone 1 or tone 2, there should be no obvious jump on the onset of pitch contour. In Table 3, the extension of tonal phoneme "a1" is listed.

Table 3: the extension of "a1" by its left tone context for tone articulation modeling

Left tone context	Tonal phoneme	Extended phoneme
1,2	a1	aH1
3,4,5		aL1
1,2	a2	aH2
3,4,5		aL2
1,2	a3	aH3
3,4,5		aL3
1,2	a4	aH4
3,4,5		aL4

The final result is that the number of tonal phoneme in the phoneme set is doubled. In our experiments, we use main vowel method to design our baseline phoneme set which contains 86 phonemes. After adding left tone context dependent phonemes, the size of phoneme set increases to 129. It must be pointed out that the tone 5 is not considered in this scheme. The reason is that the pattern of pitch contour of tone 5 is too complex to be summarized.

4.2 Transcription based on the new phoneme set

A problem of applying the new phoneme set is the transcription. In available transcription, usually character or syllable information is given. For a tonal syllable, it's hard to determine the phoneme string based on the new phoneme set for the given transcription. We used following iterative algorithm to convert syllables in the transcription into the phoneme string based on the new phoneme set:

1) *Using a baseline phoneme set to train a tri-phone based system*
2) *Aligning all training data by the acoustic model gotten in step1.*
3) *Adjusting transcription: representing each syllable by tone dependent phoneme sequence.*
4) *Training a new system based on the transcription from step 3.*
5) *Aligning all training data by the acoustic model gotten in step4.*
6) *If the iteration number is less than a threshold, return to step 3. Otherwise, the training is finished*

So far, the step 3 is implemented by some rules. For example, if the left context of the syllable "ta1" is silence, the resulted phoneme string should be "sil t aH1". If the left tone context is tone 4, and the duration between two voiced segments is less than a threshold, the string is "… t aL1". In addition, in step 5, the alignment is carried out under a lexicon with multi-pronunciation. We hope the system can determine the best path automatically.

4.3 Acoustic feature

In our previous work, we combined MFCC and pitch based features into one feature vector for tone modeling [3]. Some evidence proved that the pitch duration and long span pitch are useful for tone recognition as well [5][10]. To model the articulation of tone, we need add all useful information in feature vector. In one-step method, the syllable boundary is unknown during decoding procedure. We utilized the information of voiced/unvoiced segmentation provided during pitch tracking to construct duration and long span pitch information, and integrate them into feature vector.

The duration information is represented as:

$$l_t = \frac{t - t_0}{D} \tag{3.1}$$

Where l_t is the location of the t^{th} frame in the current voiced segment, t is time index, and t_0 is the frame index of the start of current voiced segment. D is a constant which normalizes the value of l_t. In our system, D=15, it is the average length of voiced segment in training data.

The long span pitch feature is:

$$F_t^{'} = F_t - F_N^{''} \tag{3.1}$$

Where F_t is the pitch value at the t^{th} frame, and $F_N^{''}$ is the average pitch value of last N frames of preceding voiced segment. In our experiment, N=10.

Comparison experiments were carried out on our baseline phoneme set (86 phonemes). Four kinds of configurations of the feature vector are compared as shown in Table 4. The baseline feature set is conventional MFCC feature with it delta and acceleration, the dimension is 39. According to the work in [12], the final two dimensions of the acceleration of MFCC contribute little to the recognition. We replace them by smoothed pitch and delta pitch to form feature set 2. In feature set 3 and 4, the information of duration l_t and long span pitch $F_t^{'}$ is added into the feature vector by an incremental way. In Table 4, we can observe that the final feature set with all feature listed above provides the best performance. Hence, it is adopted in our phoneme set comparison experiments. By our experiments, similar results have been observed on other configurations of phoneme sets.

Table 4: Comparison among all feature sets

Feature Vector	Dimension	Error rate	Error reduction
MFCC	39	48.3%	-
MFCC+p+Δp	39	43.5%	9.94%
MFCC+p+Δp+ l_t	40	42.5%	12.1%
MFCC+p+Δp+ l_t+ $F_t^{'}$	41	41.19%	14.7%

5. EXPERIMENTS

Chinese syllable loop experiments were done to compare the capability between tone context independent (TCI) and tone context dependent (TCD) phoneme sets without any effect of language model.

The training data is reading-style speech and contains about 80 hours' data from 250 male speakers. There are two test sets: test set 1 is reading-style speech, containing 500 sentences from 25 male speakers. The speaking rate of test set 1 is 3.8 syllables per second. Test set 2 is spontaneous speech, containing 570 sentences from 8 male speakers. Test set 2 has a speaking rate of 5.1 syllables per second, much higher than test set 1. Standard HTK was used as experiment platform. In decoding, a multi-pronunciation lexicon was used, in which, each syllable contains two kinds of pronunciations, corresponding to two variations of each tonal phoneme. The recognition results on two test sets can be found in Table 5 and 6.

Table 5: Comparison between tone context dependent/independent modeling on test set 1

Phoneme set	Size of phone set	Error rate of tonal syllable	Error reduction
TCI phone set	86	41.19%	-
TCD phone set	129	41.13%	~0.0%

Table 6: Comparison between tone context dependent/independent modeling on test set 2

Phoneme set	Size of phone set	Error rate of tonal syllable	Error reduction
TCI phone set	86	62.28%	-
TCD phone set	127	55.98%	10.1%

6. CONCLUSION AND DISCUSSION

The representation of tone by "High" and "Low" pitch position is adopted to establish a phoneme set by which tone articulation phenomenon in Mandarin spontaneous speech can be modeled. Comparing with the tone context independent (TCI) system, the tone context dependent (TCD) phoneme set exhibits better performance on the test set of spontaneous speech, and comparable capability on the test set of reading-style speech. The experiment results are very reasonable: the tone articulation phenomenon in reading-style speech is not as serious as in spontaneous speech, so the proposed method turns out to be more effective for dealing with causal speech. In addition, the proposed method gives a phoneme set with acceptable size. In one-step method of tone recognition, it is suitable for building a large vocabulary, speaker independent speech recognition system.

For the proposed method, there are still aspects for improvement. So far the rules of converting tone independent phonemes into tone dependent ones are very rough. Some statistical methods should be used for refining the transcription. Further, we have found the tone articulation phenomenon is related with the duration of consonant initial in a syllable. This may be helpful to determine the conversion in transcription. In addition, the role of language model has never been considered in our current experiments. So a true large vocabulary system based experiment should be done. These action items will be our future work.

7. 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, pp 1543-1546, 1997.

[2] C. J. Chen, Haiping Li, Liqin Shen, Guokang Fu, "Recognize Tone Languages Using Pitch Information on the Main Vowel of Each Syllable", *Proc. ICASSP 2001*, Volume 1, 2001.

[3] Eric Chang, Jianlai Zhou, Shuo Di, Chao Huang, Kai-fu Lee, "Large Vocabulary Mandarin Speech Recognition With Different Approaches in Modeling Tones", *Proc. ICSLP 2000*, October, 2000.

[4] H. Huang, and Frank Seide, "Pitch Tracking and Tone Features for Mandarin Speech Recognition", *Proc. ICASSP 2000*, Istanbul, 2000.

[5] Frank Seide and N. Wang, "Two-Stream Modeling of Mandarin Tones", *Proc. ICSLP 2000*, October, 2000.

[6] W. J. Yang, J. C. Lee, Y. C Chang, and H.C. Wang "Hidden Markov Model for Mandarin Lexical Tone Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 36, NO. 7, pp 988-992, July 1988.

[7] Jin-song Zhang and Keikichi Hirose, "Anchoring Hypothesis and its Application to Tone Recognition of Chinese Continuous Speech", *Proc. ICASSP 2000*, 2000.

[8] W. S. Y. Wang, "Phonological Features of Tone", *International Journal of American Linguistics*, 33.2, pp.93-105, 1967.

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

[10] Keikichi Hirose and Jin-song Zhang, "Recognition of Chinese Continuous Speech Using Tone Critical Segments", *Proc. EUROSPEECH 1999*, 1999.

[11] Chao Huang, Yu Shi, Jianlai Zhou, Min Chu, Terry Wang and Eric Chang, "Segmental Tonal Modeling for Phone Set Design in Mandarin LVCSR", *Submitted to Proc. ICASSP 2004,* 2004, Montreal

[12] Thomas Eisele, Reinhold Hdeb-Umbach, Detlev Langmann, "A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition", *Proc. ICSLP 1996*, 1996.