IMPROVED TONE MODELING BY EXPLOITING ARTICULATORY FEATURES FOR MANDARIN SPEECH RECOGNITION

Hao Chao, Zhanlei Yang, and Wenju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
{hchao, zhanlei.yang, lwj}@nlpr.ia.ac.cn

ABSTRACT

For the same tone pattern, different articulatory characteristics may make the pitch contour change. This paper applies articulatory features, which represent the articulatory information, as well as prosodic features to the tone modeling. Three kinds of tone models are trained to verify the effectiveness of articulatory features. Tone recognition experiments indicate significant improvement can be achieved when using both articulatory features and prosodic features. After the first pass search of a speech recognition system, tone models using new tonal features are employed to rescoring the N-best hypotheses, and a 6.5% relative reduction of character error rate is achieved.

Index Terms— tone modeling, Mandarin, speech recognition,

1. INTRODUCTION

Different from some languages such as English, Mandarin is a syllabic and tonal language. Each Chinese character can be represented by a syllable plus a tone. Tone provides strong discriminative information for many ambiguous characters， especially for the characters who share the same syllable but have different tone patterns. Incorporating tone information into large vocabulary continuous speech recognition (LVCSR) has been proved useful [1, 2, 3, 4].

There are two ways by which tone information can be integrated into continuous Mandarin speech recognition: embedded tone modeling and explicit tone modeling [4].Compared with the embedded tone modeling, the explicit tone modeling is capable of exploiting supra-segmental nature of tones [1].

In the explicit tone modeling, high accuracy is required for tone classifier to achieve better speech recognition performance. Prosodic features, such as Fundamental frequency (F0), duration and energy have been wildly used for tone modeling. To deal with the co-articulation phenomenon in continuous speech, adjacent syllables are also employed for feature extraction, and then the performance of tone recognition can be improved [5]. Furthermore, it can be observed that the pitch contour of tones may change when they are attached to different manner and place of the articulation events. Considering the influence of manner and place of the articulation events, the phoneme dependent tone models are built in [6].

Different from [6], this paper employs the articulatory information as a form of tonal features which are used in tone modeling. The articulatory features are obtained by hierarchical MLP (Multilayer Perceptron) classifiers. Then, tone models are constructed based on prosodic features and articulatory features. To verify the effectiveness of new tonal features, we established three kinds of tone models: MLP, Support Vector Machine (SVM) and Gaussian mixture models (GMM). Finally, the tone models based on MLP are integrated into the continuous speech recognition system.

The rest of this paper is organized as follows. In Section 2, we introduce how to get articulatory features as well as select prosodic features. In Section 3, several tone modeling methods are presented. Meanwhile, the methods of incorporating tone models into Mandarin speech recognition systems are also introduced. In Section 4, a series of experiments are carried out and results of the experiments are discussed. The conclusions will be described in section 5.

2. PROSODIC FEATURE AND ARTICULATORY FEATURE

2.1. Selection of Prosodic Features

For each syllable, the tone contour features, duration and average energy are used for tone modeling. The tone contour is divided into three sections, and the average F0 value of each section is computed. The three values, as well as the average F0 value of whole syllable, are viewed as the tone contour features for the current syllable.

Considering the impact of co-articulation on tone patterns, for the current syllable, the contour features of its neighboring syllables are also used for tone modeling. Similarly, the tone contour of each neighboring syllable is divided into three sections. The last section of the previous

syllable and the first section of the next syllable are as the transitions of co-articulation. Therefore, the average F0 values of these two sections are combined with the tone contour features of the current syllable. All prosodic features are listed in table 1.

Table 1. *Prosodic features*

1	Tone contour features of the current syllable	4 features
2	Duration of the current syllable	1 feature
3	Energy mean of the current syllable	1 feature
4	Tone contour features of the neighboring syllables	2 features

2.2. Articulatory Features

2.2.1. Impact of articulatory characteristic on F0

Chinese syllable is with the Initial-Final structure. For the same tone pattern, different articulatory characteristics of initials and finals will make the F0 contour different in shape or in level. For example, if a syllable is with unvoiced initial, its F0 values are larger than that of the syllables with voiced initial for the same tone pattern. Figure 1 shows the averaged F0 contours of the four tones of the syllables with unvoiced initial ‘s’ and the syllables with voiced initial ‘l’. Besides, tongue position of vowels in finals also influences the F0 values. Compared with the vowels (such as ‘a’) with low tongue position, the vowels (such as ‘i’) with high tongue position have higher F0 values. Figure 2 shows the averaged F0 contours of the four tones for the syllables that respectively take ‘i’ and ‘a’ as finals. The F0 contours of figure 1 and figure 2 are obtained from 48373 sentences by 55 male speakers.

2.2.2. Articulatory categories of initial and final

Both the initials and the finals can be further divided into several detailed categories according to the manner and the place of articulation. The initials are divided into 4 categories (Table 2).

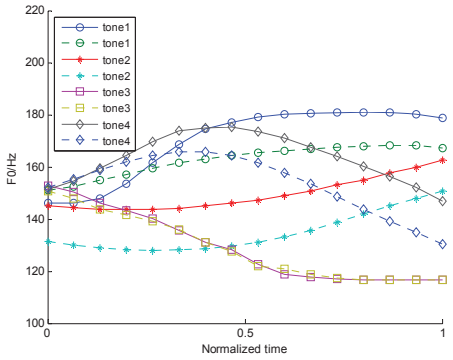

Fig.1. Averaged F0 contours of tones. The solid lines correspond to averaged F0 contours of syllables with the unvoiced initial‘s’ and the dot lines correspond to the voiced initial ‘l’.

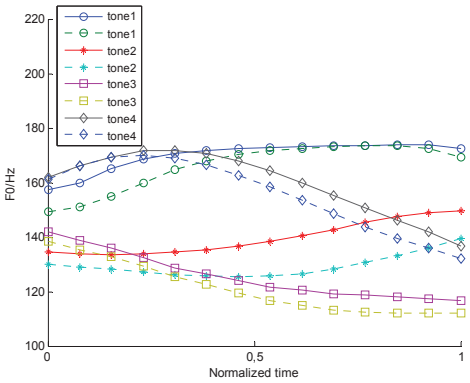

Fig.2. Averaged F0 contours of tones. The solid lines correspond to averaged F0 contours of syllables with the final ‘i’ and the dot lines correspond to the final ‘a’.

For finals, the articulatory characteristic of nasal finals (vowel followed by a nasal consonant, such as ‘ang, an’) is different from that of simple vowels (finals with one vowel, such as ‘a’) and compound finals. However, the articulatory characteristic of some compound finals is similar to that of the simple vowels.

Table 2. *Articulatory categories of initials and finals*

	Categories	Description	
1	m n l r y w	Voiced	Initial
2	b p d t g k	Stop	
3	z c zh ch j q	Fricative	
4	f s sh x h r	Affricate	
5	a ia ua	Simple vowel and tail-dominant	Final
6	e ie üe		
7	o uo		
8	i		
9	u		
10	ü		
11	er		
12	ai uai	head-dominant and centre-dominant	
13	ei uei		
14	ao iao		
15	ou iou		
16	an ian üan uan	Nasal	
17	in en uen üen		
18	ang iang uang		
19	eng ong ing iong		

In a compound final, one of its vowels will dominate the pronunciation process. According to the situation of dominant vowel in compound final, the compound finals are divided into three categories: head-dominant (ai, ei, ao, ou), centre-dominant (“iao, iou, uai, uei”) and tail-dominant (“ia, ie, ua, uo, üe”). In the pronunciation process of a tail-dominant compound final, pronunciation of the first vowel is very short, and the articulatory characteristic depends on its dominant vowel. Thus, the tail-dominant compound final

and the corresponding simple vowel (such as “ia” and “a”) fall into a class. Similarly, pronunciation of the first vowel in the centre-dominant compound final is also short. The centre-dominant compound final and its corresponding head-dominant compound final (such as “uai” and “ai”) fall into a class. Nasal finals can be categorized according to their vowel parts and nasal consonants (“n” and “ng’). Thus, the finals are divided into 15 categories (Table 2).

2.2.3. Extraction of articulatory features

The terms “articulatory features” in this paper are used to calculate the posteriori probabilities of the articulatory categories, conditioned on the acoustic features of a speech segment. To obtain articulatory features, hierarchical MLP classifiers are used. As shown in Figure 3, MLP classifiers in the first level are trained by using standard acoustic features. The two MLP classifiers in the second level are trained using posterior features estimated by the previous MLP classifiers. Each MLP classifier in the first level corresponds to an articulatory category in Table 2. For a frame of acoustic features, a vector of posterior probability outputs can be acquired from these MLP classifiers. Thus, for a speech segment corresponding to the initial or the final of a syllable, a sequence of posterior probability vectors can be obtained. Then, the vector sequence is divided into 3 sections with a 3-4-3 ratio, and the vectors are averaged for each section. The three vectors plus the logarithm of the duration of the speech segment are concatenated as the inputs of an integrative MLP classifier in the second level, which maps them to the segmet-level category probabilities. Integrative MLP-1 is used when the speech segment belongs to initials, and integrative MLP-2 is used when the speech segment belongs to finals. The integrative MLP-1 has 4 output nodes corresponding to articulatory category 1-4, while integrative MLP-2 has 15 output nodes corresponding to articulatory category 5-19. Thus, for a syllable, 4 output values of the integrative MLP-1 and 15 output values of integrative MLP-2 are combined as the articulatory features for tone recognition.

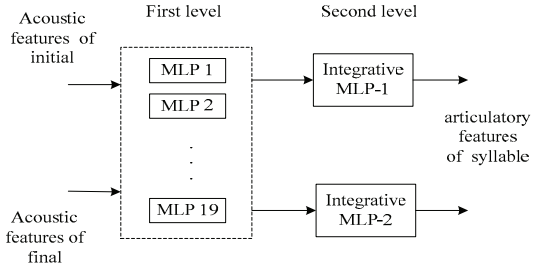

Fig. 3. Estimation of posterior probabilities of articulatory categories using hierarchical MLP classifiers

3. TONE MODELING AND INTEGRATING TONE MODELS INTO LVCSR

3.1. Tone Modeling

To verify the effectiveness of both prosodic features and articulatory features, this paper builds three kinds of models for tone recognition: MLP, SVMs and GMMs. The MLP model has one hidden layer, and the number of hidden nodes is 2N+1, where N is the number of input nodes. LibSVM is used to train the SVM model [7]. For GMM based tone modeling method, several GMMs are constructed to make sure that each tone pattern corresponds to a particular GMM. The number of mixture components in a GMM is 128, and diagonal covariance matrices are adopted

3.2. Integrating Tone Models into LVCSR

The tone models are integrated into the continuous speech recognition system by rescoring the N-best hypotheses. The N-best hypotheses are generated by the first pass recognition, and are reranked with tonal scores integrated. The total score of a hypothesis can be defined as:

$$\psi = \sum_{i=1}^{D} [\psi_{AM}(s_i) + \alpha \psi_{LM}(s_i) + \beta \psi_{TM}(s_i)] \qquad (1)$$

where α is the language model weight, β is the tone model weight, and D is the syllable number of the hypothesis for a candidate path. Thus, the hypothesis with the highest path score is regarded as the best hypothesis.

4. EXPERIMENTS

4.1. Experimental Setup

The data corpus applied in experiments comes from Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development. 83 male speakers’ data are employed for training (48373 sentences, 55.6 hours) and 6 male speakers’ for test (240 sentences, 17.1 minutes). Acoustic features are 12 dimensions MFCC plus 1 dimension normalized energy and their 1st and 2nd order derivatives.

As the baseline, A HMM based speech recognition system is developed by HTK V3.2.1 [8]. The structure of HMM is left to right with 5 states, 3 emitting distributions and no state skipping, except ”sp” (short pause) model with 3 states, 1 emitting distribution. Each emitting distribution is modeled by 16 Gaussian mixtures. A bigram language model with 48188 words is used.

The force alignment, which is applied to training the hierarchical MLPs and tone models, is implemented by the baseline system. In addition, the baseline system generates the N-best hypotheses. When integrating tone models into LVCSR, the boundary information is obtained according to the current hypothetic paths.

4.2. Tone Recognition

We use MLP, SVM and GMM to model the tone patterns, in which force alignment on test utterances is implemented to extract the tonal features. The tone recognizers are tested on 240 utterances from the test set. The experimental results are shown in Table 3.

Table.3. *Accuracy of tone recognition with various features and models*

	MLP(%)	SVM(%)	GMM(%)
prosodic features	74.21	74.54	73.50
prosodic features + articulatory features	80.61	79.43	79.27

As can be seen in table 3, the accuracies of these three models are all improved significantly when articulatory features are fused with traditional prosodic features.

Although good performance has bee achieved by integrating articulatory features into tonal features, further improvement would be obtained if the hierarchical MLP classifiers, as the detectors of articulatory characteristics for speech segments, have higher accuracy.

To obtain more accurate detection results of articulatory features, another method is employed instead of hierarchical MLP classifiers. First, we force align the speech utterance. Thus, every speech utterance is cut into several segments, and corresponding intinial/finial are labeled to every segment. For a syllable, there are two segments: initial segment and final segment. The initial segment and the final segment are categorized according to Table 2: the initial segment is assigned to an initial category and the final segment is assigned to a final category. Then, the articulatory features for this syllable can be obtained: the feature value is set to 1 for the categories that the initial segment or the final segment belongs to, and to 0 for other categories

Table 4 shows the experiment results when the articulatory features are obtained by this method. Because the articulatory features are binary, it is not appropriate to model tones using GMM. Compared with results in the third line of table 3, the accuracies of MLP and SVM are further improved. Obviously, this feature extraction method can not be used when tone information is incorporated into continuous speech recognition because of the impossibility of force alignment. Nevertheless, results in Table 4 demonstrate the greater potential of articulatory features

Table.4. *Accuracy of tone recognition when using binary articulatory features*

Model	MLP(%)	SVM(%)
Accuracy(%)	83.43	82.37

4.3. Rescoring the N-best Hypotheses

Performance comparisons among the baseline and the system incorporated with tonal information are implemented. The tone model used here is MLP classifier. The weight of the tone model is set empirically.

Table.5. *Speech recognition results*

System	Sub.	Ins.	Del.	Err.
Baseline	14.56	0.16	0.48	15.33
Tonal incorporation	13.74	0.22	0.38	14.34

As can be seen in table 5, the character error rate (CER) of the baseline system is 15.33%. When the tone information is merged with the weight of 5.5, a 6.5% relative reduction is achieved. And the results demonstrate that tone cues can provide discriminative information for speech recognition.

5. CONCLUSIONS

This paper firstly investigates the influence of articulatory characteristic on the F0 contour, and then integrates the articulatory information as a form of tonal features into the tone recognition tasks. Results on the tone recognition tasks have shown that the usage of both prosodic features and articulatory features can improve accuracies significantly. When the constructed tonal model is fused with prosodic features and articulatory features, a promising result can be obtained for continuous speech recognition.

6. ACKNOWLEDGEMENTS

This work was supported in part by the China National Nature Science Foundation (No.91120303，No.90820303 and No.90820011) and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

7. REFERENCES

[1] X. Lei, et al, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," in *Proc. ICSLP*, 2006, pp. 1237-1240.

[2] H X Wei, et al, "Exploiting prosodic and lexical feature for tone modeling in a conditional random field framework," in *Proc. ICASSP*, 2008, pp. 4549－4552.

[3] Yao Qian, Frank K. Soong, "A Multi-Space Distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition," *Speech Communication*. vol. 51, pp.1169-1179, 2009.

[4] T.Lee, et al, "Using tone information in Cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing,* vol. 1, pp. 83-102, 2002.

[5] Yao Qian, Tan Lee, "Tone recognition in continuous Cantonese speech using supratone models," *Journal of the Acoustical Society of America*, 121(5):2936-2945, 2007.

[6] Ye Tian, et al, "Tone recognition with fractionized models and outlined features," in *Proc. ICASSP*, 2004, pp. 105－108.

[7] LibSVM—A Library for Support Vector Machines, Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] S. Young et al, *The HTK Book*, Cambridge, 2000.