TONE RECOGNITION WITH FRACTIONIZED MODELS AND OUTLINED FEATURES

Ye Tian, Jian-Lai Zhou, Min Chu, Eric Chang

Microsoft Research Asia
{t-yetian, jlzhou, minchu, echang}@microsoft.com

ABSTRACT

In this paper, different feature extraction and tone modeling schemes are investigated on both speaker-dependent and speaker-independent continuous speech database. Tone recognition features can be classified as detailed features which use the entire F0 curve, and outlined features which capture the main structure of the F0 curve. Tone models of different size, ranging from very simple one-tone-one-model tone models to complex phoneme-dependent tone models, have different ability to characterize tone. Our experiments show two conclusions: 1) the detailed information of the F0 curve is not necessary for tone recognition. The outlined features can not only reduce the number of parameters, but also improve the accuracy of tone recognition. The subsection average F0 and ΔF0, proposed in this paper, is shown to be effective outlined features. 2) The one-tone-one-model scheme is not sufficient. Building phoneme-dependent tone models can highly improve the tone recognition accuracy, especially for speaker-independent data. Thus we suggest use fractionized models trained with outlined features for tone recognition.

1. INTRODUCTION

Chinese is known as a syllabic and tonal language and tone recognition plays an important role and provides very strong discriminative information for Chinese speech recognition [1]. In this paper, we investigate the feature extraction and tone modeling problems.

Traditionally, detailed tone features such as the entire F0 curve are used for tone recognition. In [2], the observation sequence $(F0, \Delta F0)$ of the whole syllable are used with a combination of VQ (Vector Quantization) and HMM (Hidden Markov Model) for tone recognition.

To reduce the number of parameters and improve the robustness, some outlined feature can be extracted from the F0 curve to make tone recognition. In [4], Features are fed into a multi-layer preceptor pattern recognizer for tone recognition, and features include the duration of the F0 contour, the mean of three uniformly divided log-energy sub-contours, and the intercept and slope of three uniformly divided F0 sub-contours. In [5], the voice portion of syllable is divided into 16 subsections and the temporal pitch variation within the syllable is used for tone recognition. In [6], the fuzzy C-means algorithm is applied to tone recognition, and the feature is obtained by curve fitting. The curve fitting function is $f_t = a + bt + ct^2$, and parameter $\{a, b\}$ is used for tone recognition. The phonetic evidence is that the F0 contour of tones can be simply represented by standard templates as shown in table 1. Moreover, as stated in [3], "the recognition of contour tones is crucial in the analysis of certain types of tone systems if we are to capture all and only the consistent characteristics in the phonological structure", and "Over-differentiation would only lead to chaos."

Table 1. Onset and offset F0 value of basic lexical tones

	Onset f0	Offset f0	Direction
Tone1	High	High	Level
Tone2	Low	High	Rising
Tone3	Low	Low	Falling then Rising
Tone4	High	Low	Falling

Tone models of different size, ranging from very simple one-tone-one-model tone models to complex phoneme-dependent tone models, have different ability to characterize tone. Most of the previous algorithms use one-tone-one-model tone models, that is, only five tone models and each model corresponds to one tone. To make tone model more precise, one-tone-one-model scheme can be further subdivided to fractionized tone models according to certain rules. From our observation, we found that in continuous speech the F0 curve of one tone may be different when tone is attached to different phoneme. Thus we expect higher performance for phoneme-dependent tone models.

In previous literatures, there is neither comparison between detailed features and outlined features, nor comparison between fractionized tone models and one-tone-one-model tone models. In this paper, we make these comparisons on both speaker-dependent and speaker-independent Chinese continuous speech database.

0-7803-8484-9/04/$20.00 ©2004 IEEE

According to our experiments, we try to answer the following questions:

1. Is the detailed information of the F0 curve useful for tone discrimination in continuous speech? Can we capture the consistent characteristics in the phonological structure?
2. Are phoneme-independent tone models sufficient for continuous speech recognition? Is it helpful to build phoneme-dependent tone model?

The paper is organized as follows: we investiagates the features problems in section 2 and the models problem in section 3, with the conclusions given in section 4.

2. DETAILED FEATURES AND OUTLINED FEATURES

Tone recognition features can be classified as detailed features which use the entire F0 curve, and outlined features which capture the main structure of the F0 curve.

2.1 Detailed features

In this representation, all the time frame's F0 values of the entire phoneme portion are used to form the observation sequence. Commonly ΔF0 curve are also used, and the observation vector is (F0, ΔF0).

If the phoneme has totally N frames, the number of total parameters used for tone recognition is 2*N.

2.2 Outlined features

To reduce the number of parameters and improve the robustness, some outlined features can be extracted from the F0 curve to make tone recognition.

2.2.1 Curve fitting features

Curve fitting uses a specific function to approximate the real F0 curve of the entire phoneme.

The one-order linear regression use $f(t) = at + b$ to fit the $F0$ curve. The objective of the curve fitting is to achieve the minimum fitting error. The fitting parameters $\{a, b\}$ are used for tone recognition.

The two-order linear regression [6] uses $f(t) = at^2 + bt + c$ to fit the $F0$ curve and the fitting parameters $\{a, b, c\}$ are used for tone recognition.

2.2.2 Subsection outlined features

In subsection outlined representation, the F0 curve of the entire phoneme is divided into several subsections and each subsection is represented by certain parameters.

Assume that time frames $\{F_{0,k_s}, F_{0,k_s+1}, ..., F_{0,k_e}\}$ belong to the subsection k, where k_s and k_e are the start and the end time frame index of the subsection. The following parameters can be extracted for the subsection k:

1) Subsection slope and intercept [4]:

For each subsection's $F0$ curve, we can use linear regression to obtain the slope $S_{F0}{}^K$ and intercept $I_{F0}{}^K$, that is, using $f(t) = S_{F0}{}^K * t + I_{F0}{}^K$ to approximate the $F0$ curve belongs to this subsection.

According to [7], define $X = [0,1,..., k_e - k_s]$, $Y = [F_{0,k_s}, F_{0,k_s+1}, ..., F_{0,k_e}]$, then

$$S_{F0}{}^K = \frac{\left[\sum_{i=1}^{k_e-k_s+1} (X_i - \overline{X})(Y_i - \overline{Y})\right]}{\left[\sum_{i=1}^{k_e-k_s+1} (X_i - \overline{X})(X_i - \overline{X})\right]} \quad (1)$$

$$I_{F0}{}^K = \overline{Y} - S_{F0}{}^K \overline{X} \quad (2)$$

, where $\overline{X}, \overline{Y}$ is the average value of X and Y, respectively.

2) We propose two new outline features: subsection $\overline{F}_0$ and $\Delta\overline{F}_0$:

$$\overline{F}_0{}^k = \frac{1}{k_e - k_s + 1} \sum_{i=k_s}^{k_e} F_{0,i} \quad (3)$$

$$\Delta\overline{F}_0{}^k = \frac{1}{k_e - k_s + 1} \sum_{i=k_s}^{k_e} \Delta F_{0,i} \quad (4)$$

We also tested several other subsection features, including subsection energy, subsection duration, and energy-weighted F0. We do not present them here because their performance is limited.

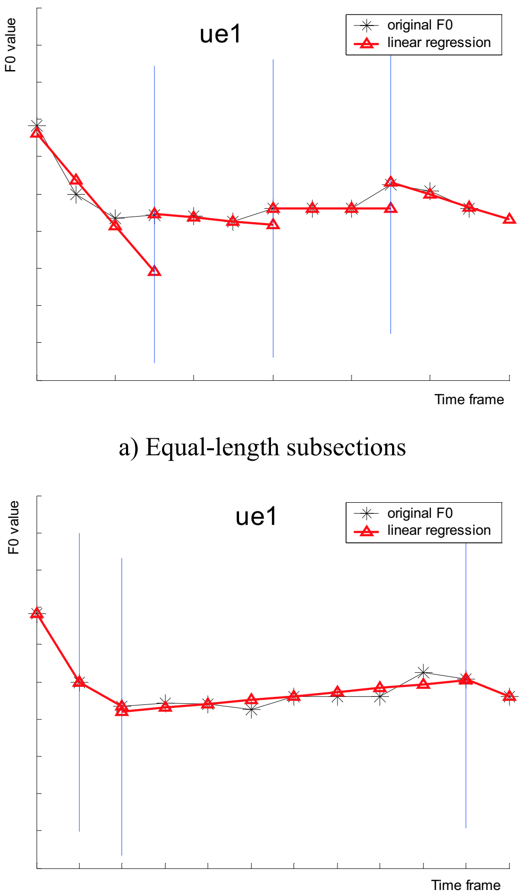
In [4], the subsections are equal-length. In our point of view, the division of subsections can also be optimum selected according to certain rule. We propose a new algorithm referred to the *best-fit division algorithm*. For each possible group of subsections, the linear regression error of each subsection is evaluated. The total linear regression error is calculated as the sum of each subsection's linear regression error. If all possible subsections are ransacked, the best-fit division is the one corresponding to the minimal total linear regression error. As shown in Fig 1, the $F0$ curve of "ue1" is approximated by 4 linear functions. The Fig 1a) use equal-length division and Fig 1b) use the best-fit division.

2.3 Experiments and analysis

We test tone recognition features on both speaker-dependent and speaker-independent Chinese continuous speech database.

Table 2. Tone recognition accuracy with different features
(The one-tone-one-model tone models are used. N is the frame number of the syllable)

Features used		Feature-size / syllable	Model	Tone recognition accuracy	
				Speaker-dependent	Speaker-independent
Detailed feature	(F0,ΔF0) by frame	2*N	HMM	77.18%	62.50%
Curve fitting features	one-order linear regression	2	GMM	75.68%	46.21%
	two-order linear regression	3	GMM	71.29%	54.75%
Subsection outlined features	4 equal-length subsections ($\overline{F0}$, $\overline{\Delta F0}$)	8	GMM	**85.21%**	**65.93%**
	4 equal-length subsections (slope, intercept)	8	GMM	83.81%	65.19%
	best-fit division subsections (slope, intercept)	8	GMM	77.35%	59.55%



a) Equal-length subsections

b) The best-fit division algorithm
Fig.1 the F0 curve of "ue1" is approximated by 4 linear functions.

The speaker-dependent database is chosen because it is recorder originally for text-to-speech purpose and the F0 is well recorded by an EGG device. Thus we can remove the influence of pitch extractor error. The database totally contains 14476 sentences and 216118 syllables. We selected 13047 syllables as test data and others as train data.

The speaker-independent database contains 250 speakers and 1400879 syllables. The F0 curve of each sentence is extracted by the pitch extractor given in [8]. We selected 8561 syllables of 2 speakers as test data and others are used as train data.

The HMM and GMM are implemented by HTK [9]. The state observation probability is 16 Gaussian mixtures. One-tone-one-model tone models are used.

The results are listed in Table 2. From the table, we can see the following points:

1. The $\overline{F0}$ and $\overline{\Delta F0}$ of the equal-length subsections are the best outlined features for tone recognition. They outperform the detailed features. This means that the main value and direction are the most important characteristics for tone recognition. The detailed information is not necessary for tone recognition in continuous speech. Because the F0 curve of continuous speech contains so much variation including co-articulation effect, syllable stress, and sentence intonation, ignore these details is helpful. It also shows that the subsection $\overline{F}_0$ and $\Delta\overline{F}_0$ can capture the consistent characteristics in the phonological structure.
2. Although the best-fit division algorithm fits the original F0 curve more precisely, its performance is inferior to the equal-length subsections. The reason is that the detailed information is ignored in the equal-length subsections but remained in the best-fit division algorithm. This shows again that the detailed information is useless for tone discrimination.

3. FRACTIONIZED MODELS AND ONE-TONE-ONE-MODEL MODELS

In the above experiments, we use one-tone-one-model scheme. In this section, we will compare the tone recognition accuracy of tone models of different size.

3.1 Tone modeling

Table 3. Tone recognition accuracy with different tone models

Tone models	Features	Tone recognition accuracy	
		Speaker-dependent	Speaker-independent
One-tone-one-model tone models (5 models)	(F0,ΔF0)	77.18%	62.50%
	(F0,ΔF0,MFCC)	72.98%	65.23%
Monophone-dependent tone models (54 models)	(F0,ΔF0,MFCC)	80.46%	69.64%
Triphone-dependent tone models (12824 models)	(F0,ΔF0,MFCC)	**86.03%**	**81.01%**

From our observation, we found that in continuous speech the F0 curve of one tone may be different when the tone is attached to different phoneme. Thus we expect higher performance for phoneme-dependent tone models. In this section we will consider the following tone modeling schemes:

1. One-tone-one-model tone models. There are 5 tone models and each for one tone.
2. Monophone-dependent tone models. We use tonal phonemes to make large vocabulary speech recognition. Thus the same tone in different tonal phonemes is different modeled. There are 54 tone models.
3. Triphone-dependent tone models. We use tonal phonemes and create context-dependent triphone to make large vocabulary speech recognition. Thus the same tone in different triphone is different modeled.

There are totally 12824 triphone models whose center phoneme is tonal phoneme. States are tied to share the data and make the robust parameter estimates since there will be insufficient data associated with many of the states. There are only 4981 states after tied.

The later two schemes are speech recognition based tone recognition. We propose them here to compare one-tone-one-model tone models and fractionized tone models.

3.2 Experiments and analysis

We compare the four tone modeling schemes on the same databases as used in section 2.3. For phoneme-dependent tone models, the feature vector (F0,ΔF0,MFCC) is used because with the increase of model number, vector (F0, ΔF0) is not sufficient for discriminate the phoneme-dependent models. For one-tone-one-model scheme, tone recognition accuracy on both (F0,ΔF0) and (F0,ΔF0,MFCC) are evaluated to make sure the improvement of tone recognition accuracy is not due to MFCC feature.

The results are listed in Table 3. From the table, we can see that:

1. The one-tone-one-model scheme is not sufficient for tone recognition in continuous speech. Building phoneme-dependent tone models can improve tone recognition performance.
2. The tone recognition accuracy is higher for more fractionized tone models. Triphone-dependent tone models achieve the best performance, and the improvement is notability especially for speaker-independent data.

4. CONCLUSIONS AND FUTURE WORK

Our experiments show that we suggested using fractionized models and outlined features for tone recognition. Outlined features can reduce the interference caused by various modifications such as co-articulation effect, syllable stress, and sentence intonation, thus can improve the performance. Fractionized tone models use *a prior* knowledge to partition the whole tone space and reduce the confusions among different tones, can also improve the performance.

Future work can be focused on how to apply such conclusions on large vocabulary speech recognition system, such as using outlined features instead of the detailed features to improve tonal syllable recognition accuracy.

5. REFERENCES

[1] L. S. Lee, “Voice Dictation of Mandarin Chinese”, *IEEE Signal Processing Magazine*, V14, n4, 1997, pp.63-101.

[2] W. J. Yang, J. C. Lee, Y. C. Chang, H. C. Wang, “Hidden Markov model for Mandarin lexical tone recognition”, *IEEE Trans. Acoustics, Speech, and Signal Processing*, v36, n7, 1988, pp.988-992.

[3] S. Y. Wang, "Phonological Features of Tone", *International Journal of American Linguistics*, v33, n2, 1967, pp.93-105.

[4] S. H. Chen, Y. R. Wang, “Tone recognition of continuous Mandarin speech based on neural networks”, *IEEE Trans. Speech and Audio Processing*, v3, n2, 1995, pp.146-150.

[5] L. Tan, P. C. Ching; L. W. Chan, Y. H. Cheng, B. Mak, “Tone recognition of isolated Cantonese syllables”, *IEEE Trans. Speech and Audio Processing,* v3, n3, 1995, pp.204 -209.

[6] J. J. Li, X. D. Xia, S. S. Gu, “Mandarin four-tone recognition with the fuzzy C-means algorithm”, *in Proc. FUZZ-IEEE 1999*, v2, pp.1059-1062.

[7] D. Y. Zhu, *Multivariate statistical analysis*, DongNan University Press, 1999

[8] E. Chang, J. L. Zhou, C. Huang, S. Di and K. F. Lee, "Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones", *in Proc. ICSLP 2000*, vol. 2, pp. 983-986.

[9] Steve Young et al., *The HTK book*, 2002.