SUBSYLLABIC TONE UNITS FOR REDUCING PHYSIOLOGICAL EFFECTS IN AUTOMATIC TONE RECOGNITION FOR CONNECTED MANDARIN CHINESE

Jin-song Zhang, Goh Kawai, and Keikichi Hirose

University of Tokyo, Japan

zjs@gavo.t.u-tokyo.ac.jp, goh@kawai.com, hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper presents our attempt to model physiological transition effect on syllable F0 contour in order to improve lexical tone recognition performance for Mandarin Chinese. We suggested that a syllable F0 contour consists of three segments: onset course, tone nucleus and offset course. Among the three segments, only tone nucleus contains key features for tone recognition, and the other two result from physiological transition effect of human vocal cords. Therefore, the tone recognizer performance can be improved by only focusing on tone nuclei and discarding the other two segments. Segmentation of the three segments was achieved by our proposed method. Context-dependent tonal models, which are trained on tone nucleus features, were also introduced to model contextual tone coarticulation effects for tone recognition. Advantages of the proposed methods were proved through tone recognition experiments of continuous speech of Mandarin.

1.INTRODUCTION

Voice fundamental frequency contours (F0 contours) play an important role in the transmission of linguistic information concerning word meaning and sentence structure. While linguistic information intended by a speaker is discrete, its corresponding F0 contour is semi-continuous both in time and in frequency. In fact, F0 contour of an utterance is usually a quasi-continuous curve with rather complex undulations. This can be ascribed mainly to the transfer characteristics of the physiological and physical mechanisms involved in the control process of the voice fundamental frequency [1]. Therefore, study is necessary to elucidate the relationship between the acoustic features and the linguistic events.

In Chinese, a syllable corresponds to a morpheme and its basic structure is (C)V with a lexical tone assigned to it. The two components of a syllable are ordinarily referred as the Initial and the Final. The Initial can be a consonant or none, and the Final can be a nuclear vowel (or diphthong), or a vowel with a preceding glide, or a vowel plus a nasal –n or –ng. There are four basic lexical tones, each of them being characterized by a distinct F0 contour, which can be represented by its onset and offset F0 values[table 1] [2, 3]. Since F0 contour of a syllable is a quasi-continuous curve which may contain large undulations, there exist no distinct “onset ” and “offset” points in the contour. Therefore engineers use whole syllable F0 contours as acoustic feature to recognize lexical tones[4]. But the undulations in syllable F0 contours may interfere with tone recognition, and lead to low tone recognition accuracy.

Although linguists knew to ignore undulations in F0 contour by using “onset” and “offset” values in analyzing tonal events [5], there have been few discussions on using such idea for tone recognition. Also there are no discussions on automatic detection of the “onset” and “offset” points from syllable F0 contours.

	Onset F0 value	Offset F0 value
Tone 1	High	high
Tone 2	Low	high
Tone 3	Low	low
Tone 4	High	low

Table 1. Representation of F0 contours for four lexical tones by two-point distinct values.

This paper presents our attempt to use the ideas of “tone onset” and “tone offset” to model syllable F0 contour for tone recognition. The approach includes definitions of three sub-syllabic tone units and their segmentation method from syllable F0 contours. The suggested three sub-syllabic tone units are called as“onset course”, “tone nucleus” and “offset course”, respectively, according to their roles in tone realization. Furthermore, another approach to cope with physiological coarticulation effect, where context-dependent acoustic tone models based on tone nuclei, were introduced. Effectiveness of these approaches receives proof from improvements in tone recognition performances.

2.DEFINITION OF SUBSYLLABIC TONE UNITS

For a syllable F0 contour, only its later portion is regarded to bear tone information [6], whereas the early portion is regarded as physiological transition period from the previous tone. Besides this observation, we also found there is often a voicing period in the ending portion of a syllable which also forms a transition period of vocal vibration and contributes nothing to the tonality

[7]. The whole locus of vocal cord vibration for a tone can be assumed to result from a four-stage tone generation process.

2.1 Four-stage tone generation process

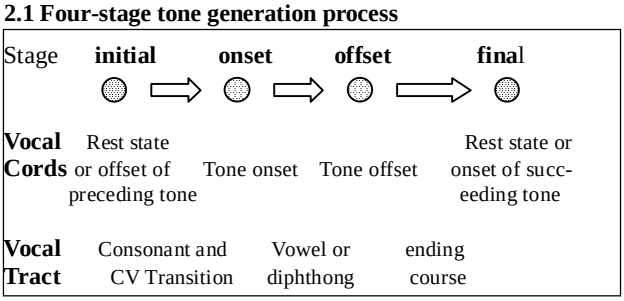

Fig.1 Illustration of four-stage tone generation process. Different physiological states are associated with "Vocal Cords" and "Vocal Tract" during a tonal syllable generation.

Fig.1 depicts the tone generation process and the associated possible physiological states. In an isolated syllable F0 contour, vocal cords are in rest states in the initial and final stages. But in connected speech, the two rest states may be substituted by the offset state of the preceding tone and the onset state of the succeeding tone. There assumed to be rather systematic relationship between vocal cord states and vocal tract states: for instance, period from tone onset to offset state may correspond to syllable Final vowel or diphthong.

2.2 Three sub-syllabic tone units

From the above consideration, we suggested that a syllable F0 contour can be divided into 3 segments and they play different roles in tone realization. We called them: onset course, tone nucleus and offset course, respectively, according to their roles in the tone generation process.

- **Onset course:** segment corresponding to the period of vocal cords vibration from the rest state or offset state of the preceding tone to the onset state of the current tone. Phonemically, this course spans the Initial consonant (or silence) and the transition period to the Final vowel or diphthong. F0 contour of this segment can be predicted by interpolation when the tone onset point is known.
- **Tone nucleus:** segment corresponding to the period from tone onset state to offset state where vocal cord vibrates. Normally, F0 contour of this segment keeps the basic pattern of the tone unless it is affected by high level prosodic factors such as neutralization, contextual effect, focus, phrasing, and etc. Phonemically, this segment covers the nucleus of the Final vowel or diphthong.
- **Offset course:** segment containing transition locus from the tone offset state to the rest or the onset state of the succeeding tone. F0 contour of this segment can also be given by interpolation. Phonemically, this segment holds the ending course of syllable Final.

Among the three segments, only the tone nucleus is essential to a tone realization, whereas the other two segments are optional; their appearance depends on voicing characteristic of Initial consonant, contextual relation, syllable duration and etc. Based on the definitions of these three segments, we can say that a syllable F0 contour can be approximated by no more than 3 slant lines if ignoring micro undulations.

3.AUTOMATIC SEGMENTATION ALGORITHM

In view of the fact that there is to some extent relationship between vocal cord and vocal tract events during generation process of Chinese tonal syllables [8], we once tried to use sub-syllabic unit HMMs of vocal tract features to segment a tone syllable into the three sub-syllabic tone units. Experiments were carried out on a multi-speaker disyllable database with satisfying results. One problem of this method is that it requires the database to be hand-labeled according to the three sub-syllabic tone units. In order to avoid the time-consuming labeling work, we developed an automatic segmentation scheme for continuous speech.

This scheme can presently deal with the F0 contour variations of Tones 1, 2 and 4, but not Tone 3. The reason lies in that Tone 3 has more complex contextual effect than the two-point distinct feature [table 1] can describe. For example, when a tone contour preceding a Tone 3 syllable is raised due to the "anticipatory raise" [6] coarticulation effect, the F0 contour of Tone 3 also changes into a falling shape which resembles that of Tone 4. It is impossible to locate a pair of onset and offset points with similar levels of low F0 values in a falling shape. The segmentation scheme is based on the following understanding about tone nucleus:

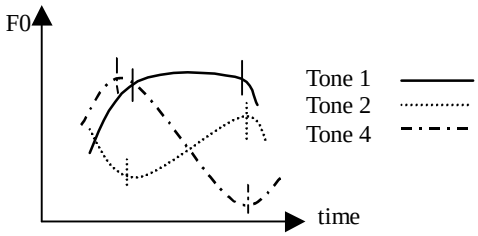

Fig2. Illustration of F0 contours with transition loci for Tone 1, Tone 2, and Tone 4. The medium segment of each contour represents tone nucleus.

①Tone nucleus has a rather stable F0 contour, with a length longer than 50ms.
②Tone nucleus always lie in the later part of a syllable.

The segmentation scheme consists of two processes. During the first process, F0 contour of Final portion of a syllable is divided into three segments, each owning a relatively consistent shape. During the second process, analysis is conducted for the 3 segments to decide if there are any transition loci according to non-F0 features of energy and duration that affect perception.

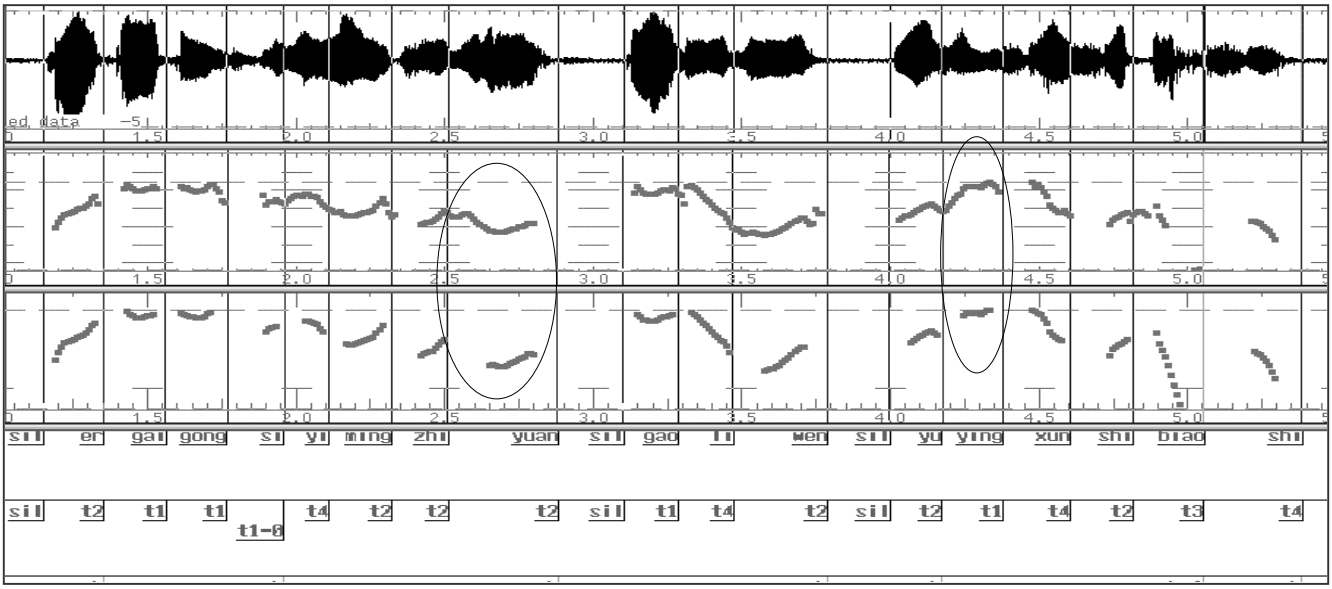

Fig.3 An example of the tone nucleus segmentation result. The second panel from the top depicts original F0 contours and the third one depicts segmented tone nuclei F0 contours. For example, (indicated by two circles), the falling segment in 8th syllable "yuan" with Tone 2, and rising segment in 13th syllable "ying" with Tone 1 were excluded from tone nuclei.

3.1 F0 contour segmentation based on clustering algorithm

In this process, all the frame points in the F0 contour of a syllable Final are automatically classified into one of three segments, each owing a consistent shape. Location information for syllable Initial and syllable Final is assumed to be available, as the by-products of a phoneme recognizer. F0 and its first time derivative act as the features for segmentation. The algorithm is:

step 1: Start from 3 segments with equal length.

step 2: Compute means and variances of the parameters of these segments.

step 3: Re-assign each frame point into the 3 segments according minimum distance rule. (Viterbi search is used to find the path.)

step 4: Repeat step 2 and 3 until convergence.

3.2 Analysis based on non-F0 features:

The three segments are analyzed according to F0 and non-F0 features to find possible transition loci.

step 1: Integration: The neighboring segments having the same inclination (positive, negative, or zero) are combined.

step 2: Sonority rule: Energy features are used to estimate the possible sonority of each segment. Energy of tone nucleus should satisfy the sonority requirement.

step 3: Duration rule: Duration of a tone nucleus should be longer than 50ms.

step 4: Location rule: Tone nucleus should be in the later portion of a syllable F0 contour.

step 5: Smoothing: widen the detected tone nucleus into transition loci when there exist voicing points with features similar to those of tone nucleus period.

Fig. 3 gives an example of original and segmented tone nuclei F0 contours of a continuous utterance. We can see the segmented contours show more consistent shapes with respect to its lexical tone.

4.TONAL COARTICULATION AND TRI-GRAM CONTEXT DEPENDENT MODEL

While the sub-syllabic tone units are useful to model physiological transition effect for tone recognition, they can do nothing about tonal F0 contour variations due to contextual tonal coaticulation effects. This is because contextual tonal coarticulation effects are not local, but cover the whole syllable tone. Studies have revealed that not only the onset and offset points can be lowered or raised, but also inclination can be changed [5,9]. These coarticulation effects result from the physiological and mechanical constraints of human vocal cord vibration. When two tones have quite different F0 values in the conjunction area, the vocal cord vibration cannot change from a state to another state abruptly, instead, is subject to compromise effect [9].

Engineers have already tried to use tri-gram context dependent models to model such kinds of tonal coarticulation effects. For example, a tone series "Tone 1 – Tone 2 – Tone 3 – Tone 4" can be acoustically modeled by context-independent (CI) or context-dependent (CD) models like:

CI	T1	T2	T3	T4
CD	T1-(T2)	(T1)-T2-(T3)	(T2)-T3-(T4)	(T3)-T4

Therefore, in CD models, coarticulation effects on each conjunction area can be modeled by an allotone for a basic tone.

Different from other approaches, we trained context dependent tonal models using the tone nucleus features of each tone.

5.TONE RECOGNITION EXPERIMENTS

700 utterances by a female speaker from the speech corpus HKU96, developed at the University of Hongkong, were used in the tone recognition experiments. The total number of syllables is 8985 (and, accordingly, 12.84 syllables per a sentence on the average). The speaker uttered in a citation form at a speed ranging from 3.8 to 4.9 syllables per a second. Syllable F0 contours keep their original shapes at lower speech rates, but are largely distorted when speech rate exceeded 4.5 syllables per a second. 500 out of 700 utterances were used as the training set and the rest formed the testing set. All speech samples were recorded with 16kHz sampling rate. F0 and energy were calculated for each frame with 10ms step.

Continuous density HMMs with left-to-right configuration were used to model the tones. The number of tone HMMs is 5 (4 basic tone models and a neutral tone model) in the context-independent model experiment, and number of tonal model is 222 (silence is considered) in the context-dependent model experiment. Feature vector consisted of six components: log F0, normalized rms energy, and their first and second time-derivatives. Baum-Welsh and Viterbi algorithms were used for model training and testing respectively. Comparative experiments were carried out for the conventional tone recognition method using whole syllable features, and our proposed method only focusing on tone nucleus features. Tone recognition accuracies for five tones including neutral tone for the test data set are illustrated in Fig.4.

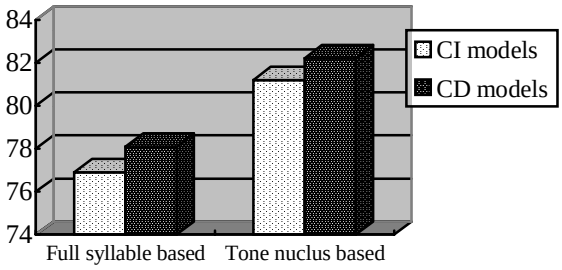

Fig. 4 Tone recognition experimental results

From Fig.4 , the following results were obtained:

1. Our methods using tone nucleus features outperformed the conventional ones both in CI and CD model cases. This result is consistent with that obtained in the experiments for disyllabic words [7]. Therefore, we can conclude that the tone nucleus segment possesses key features for tone recognition. The proposed sub-syllabic tone units are useful for modeling the physiological transition effects on syllable F0 contour.
2. Achievement by the proposed methods also indicates that the tone nucleus segmentation method is done appropriately.
3. Improvement of the recognition rate by CD models is only about 1%, much less than that expected. One possible reason may be insufficient training of the models. Another possibility is that the tri-gram CD models are not adequate for modeling tonal F0 contour variations, resulting from high level influences like sentence intonation, prosodic phrasing, and etc.
4. Even the highest tone recognition rate (82.2%) is still rather low. Besides the possible incorrect modeling of contextual tonal variations, the reason also lies in that neutral tone received no specific consideration. When neutral tone is ignored, tone recognition accuracy reached 86%.

6.CONCLUSION

Physiological transition effect on a syllable F0 contour has never received strict attention in Chinese speech recognition area, even though linguists have known the way to exclude its influence from their analysis of Chinese phonetics for a long time. In this paper we presented our attempt to prevent its interference effect on automatic tone recognition. Furthermore, context dependent tone models trained on tone nucleus features were used to model tonal contextual effects to improve tone recognition performance. Our future work is to search for more robust tone nucleus detection method, and more suitable modeling technique for tonal contextual variations.

REFERENCES

[1] Fujisaki H. and Hirose K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J. Acoust. Soc. Jpn. (E)5, 4, 233-242.

[2] Chao Y.-R., 1968, A grammar of spoken Chinese. Berkeley: University of California Press.

[3] Whalen, D. H., and Xu, Y. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica 49, 25-47.

[4] Wang H.M., et al, 1997. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. IEEE Trans. on Speech Audio Processing, 5, No.2, 195-200.

[5] Sheng X. S., 1990. Tonal coarticulation in Mandarin. Journal of Phonetics, 18, 281-295.

[6] Xu Y., 1997. What can tone studies tell us about intonation? proceedings from the ESCA Workshop on Intonation: theory, models and applications, Athens Greece. 337-340.

[7] Zhang J.-S., and Hirose K., 1998. A robust tone recognition method of Chinese based on subsyllabic F0 contours, proceedings from ICSLP98, Sydney, 703-706.

[8] Lin T., and Wang L.-J., 1991. Chinese phonetic course. University of Beijing Press.0

[9] Xu Y., 1994. Production and perception of coarticulated tones. J. Acoust. Soc. Am. (4), 2240-2253.