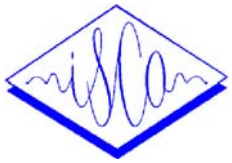

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TONE RECOGNITION OF CHINESE CONTINUOUS SPEECH USING TONE CRITICAL SEGMENTS

Keikichi Hirose and Jin-song Zhang
Department of Information and Communication Engineering
School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
hirose@gavo.t.u-tokyo.ac.jp, zjs@gavo.t.u-tokyo.ac.jp

10.21437/Eurospeech.1999-214

ABSTRACT

This paper presents our approach to automatically detect tone nuclei, and to use their features for recognizing lexical tones of Chinese continuous speech. We have suggested that Fundamental frequency (F0) contour of a syllable usually consists of three segments: onset course, tone nucleus and offset course. Among them, only tone nucleus contains key features for tone discrimination, hence the tone critical segment of a syllable. The other two segments result from physiological transition effect of human vocal cords, and are affected largely by adjacent tones in continuous speech. The tone nucleus can be detected out by a two-process scheme; the first process segments a syllable F0 contour by Segmental Clustering algorithm, and the second one finds tone nucleus according to knowledge rules on supra-segmental features. Tone recognition performance can be improved by using tone nucleus features and discarding others. Tone recognition experimental results proved the advantage of our method over the conventional ones.

1.INTRODUCTION

Chinese is a typical tonal language, where a syllable corresponds to a morpheme and basically has a (C)V phoneme structure with a lexical tone. The consonantal part C and vowel part V are referred to as the Initial and the Final respectively. The lexical tone is mainly associated with the Final. The Initial can be either a consonant or a null, while 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 (Tone 1, Tone 2, Tone 3 and Tone 4) and a neutral tone. These tones are characterized by different F0 contours except the neutral tone. F0 contours of tones in isolated syllables have stable shapes. However, contours in continuous speech vary due to various reasons, such as voicing of initial consonants, tonal coarticulation, sentential intonation, and etc. The variations cause problems to the conventional tone recognizer in which F0 features play important roles, and, therefore, robust tone recognition method is still not available up to now.

On the other hand, there are many important applications for automatic tone recognition. Firstly, accurate tone recognition is greatly helpful for Chinese speech recognition systems, since there are a large number of homophone words when discarding tone information. Secondly, Chinese speech intonation shows larger undulations in F0 contours due to lexical tones as compared with non-tonal languages, such as English and Japanese. Underlying tones need to be known to interpret the sentential intonation structure, which plays an important role in realizing smooth human-machine communication. Thirdly, automatic tone recognizer is useful in a computer-aided-language-learning (CALL) system, since learning lexical tones is very hard for foreigners.

Conventional methods for tone recognition usually use the entire portion of syllabic F0 contour as key features for recognition[1]. However, through analyses of tone features and perceptual experimental results, we have suggested in [2,3] that a syllabic F0 contour can be divided into three segments: onset course, tone nucleus and offset course. These segments contribute differently to tone perception. F0 contour of the tone nucleus contributes the most, and is called as tone-critical segment, whereas those of other two segments have limited influence on tone perception. Compared with F0 contours of the other two segments, that of tone-critical segment keeps more stable shape. Therefore, recognizing tones based on the tone-critical segment was proposed as a robust tone recognition method.

One essential prerequisite for the proposed method is that tone nuclei need to be located before their application to tone recognition. In [2] we have tried to use phoneme recognizer's output to locate tone nuclei. Although it worked, hand labeling of tone nuclei for the training database was too time-consuming. Instead, we newly developed a new scheme to automatically locate tone nuclei in continuous speech, which only uses F0 features together with phoneme and syllable segmentation information which are assumed to be available from conventional acoustic recognizer's output.

Effectiveness of the proposed tone nuclei detection scheme and tone recognition method was tested through tone recognition experiments on a female's utterances in the corpus HKU96 published by the University of Hong Kong. Tone recognition accuracy of the proposed method was compared with that of the conventional one using full syllable F0 contours through experiments of both context independent (CI) and context dependent (CD) tone HMMs. The proposed method achieved

higher accuracy by over than 5 percentages in all cases than the conventional ones.

2. FOUR LEXICAL TONES AND VARIATIONS

The four basic lexical tones can be represented by its onset and offset F0 values listed in Table 1, or F0 contour shapes showen in Fig. 1. Due to physiological articulatory constraints of human vocal cord vibration, these patterns are rarely kept unchanged when uttered in continuous speech. For example, syllable F0 contours like those in Fig. 2 are often observable in continuous speech, and even in isolated syllable cases.

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

Table 1. Onset and Offset F0 values for four basic lexical tones.

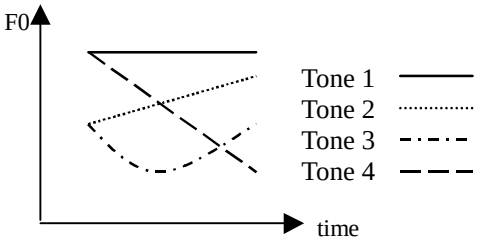

Fig. 1. Typical F0 contour shapes for 4 basic lexical tones.

According to our former discussions in [2,3], a syllable F0 contour can be assumed to consist of three segments. They are:

- **Onset course:** F0 contour segment before tone onset. It represents the transition locus of vocal cord vibration to tone onset.
- **Tone nucleus:** F0 contour segment between tone onset and offset.
- **Offset course:** F0 contour segment containing transition locus of vocal cord vibration after the tone offset.

We can see only tone nuclei delimited by vertical sticks in Fig. 2 keep the typical patterns of their associated tones, whereas other segments deviate. Hidden Markov models (HMM) are often used as acoustic models for the lexical tones [1], and F0 and its first and second time derivatives are usually selected as the acoustic features. Since state transitions of HMMs are sensitive to feature dynamics, the segments with F0 rise or fall may confuse the HMM-based tone recognizers if they are not parts of tone nuclei. For examples:

1. The rising onset course of Tone 1 may lead to an error of Tone 2.
2. The falling onset course of Tone 2 make the whole contour into a dip shape, similar to that of Tone 3.
3. The rising offset course of Tone 4 makes it like a Tone 3.

Hence, if tone nuclei can be detected then interference from other segments can be avoided in tone recognition.

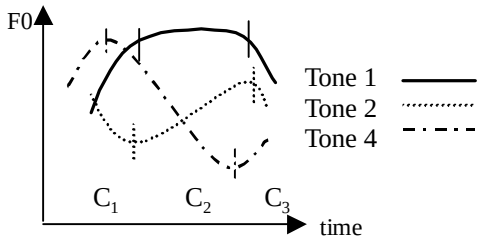

Fig. 2. Illustration of F0 contours with transition loci for Tone 1, Tone 2, and Tone 4. The medium segment delimited by vertical sticks (tone onset and offset) in each contour represents tone nucleus of the tone. Tone 3 has much more complicated variations than other 3 tones, and is not dealt here. C_1, C_2 and C_3 represent onset courses, tone nuclei and offset courses.

3. TONE NUCLEUS DETECTION SCHEME

In order to detect syllable tone nuclei in continuous speech, we developed an automatic scheme consisting of two processes.

(1) F0 contour segmentation based on F0 features.

(2) Tone nuclei decision based on knowledge rules.

This scheme can deal with F0 contour variations of Tones 1, 2 and 4, but not Tone 3. The reason lies in that Tone 3 has more complex contextual changes than the two-point features [Table 1] can describe, it is difficult to locate tone nuclei for various shapes of Tone 3. Furthermore if tone nuclei of other three tones can be detected, the confusions among the four tones can be reduced a lot.

Based on our assumption of 3-segment structure, one syllable F0 contour should contain no more than three segments with different dynamic directions, i.e. rising, falling, or flat. In other words, a syllable F0 contour may be modeled by the concatenation of 3 or less slant lines. If we check Fig. 2 once more, we may find this is very reasonable.

3.1 F0 contour segmentation based on segmental clustering algorithm

Segmentation of F0 contours is not so difficult for those of Tone 2 and Tone 4 in Fig. 2, because there are clear peak or valley points between rising (or falling) and falling (or rising) segments. But it is difficult to segment Tone 1 using the same method, since the change is gradual and there exists no peak or valley points. Furthermore, F0 contours are quasi-continuous curves with undulations, there are possible local peaks and valley points. Therefore, robust segmentation method of syllable F0 contours is necessary.

Segmental clustering algorithm seems to be a good choice for segmenting a syllable F0 contour into 3 segments, since each segment can be represented by a cluster.

Let $F0_1$, $F0_2$, ...,$F0_N$ denote the frame points of a syllable F0 contour. The *i*th point $F0_i$ has a two-

dimensional feature ($f0_i$, $\Delta f0_i$), where,

- $f0_i$ indicates F0 value, and
- $\Delta f0_i = f0_i - f0_{i-1}$ indicates delta F0 value.

The algorithm is to cluster all frame points of a syllable F0 contour into 3 sequential segment clusters [Fig. 2] according to the minimum distance rule. The resulting segment clusters are the 3 segments we are searching for. Each segment cluster contains M_i sequential frame points, and is represented by a normal distribution $\boldsymbol{N}(\mu_i, \Sigma_i)$, $i=1,2,3$. The following Mahalanobis distance is used to measure the distance between frame point $F0_j$ and segment cluster C_i.

$$dis(F0_j, N(\mu_i, \Sigma_i)) = (F0_j - \mu_i)^t \Sigma_i^{-1} (F0_j - \mu_i)$$
$$i = 1, 2, 3. \quad j = 1, 2, ..., N.$$

Algorithm:

Step 1: Initialization: dividing a tone F0 contour into 3 segments with equal length: $M_i = \frac{N}{3}$, $i = 1, 2, 3$

Step 2: Calculating distribution parameters (μ_i, Σ_i), for each segment cluster based on the Maximum Likelihood rule.

Step 3: Re-assign each frame point to the 3 segment clusters according to the minimum distance rule. Viterbi search is used to do the re-assignment:

$$\delta_1(1) = dis(F0_1, C_1), \delta_1(2) = \delta_1(3) = \infty,$$
$$\text{For } 2 \le t \le N,\ 1 \le i \le 3$$
$$\delta_t(i) = \min_{1 \le j \le i}[\delta_{t-1}(j) + dis(F0_t, C_i)],$$
$$\psi_t(i) = \arg\min_{1 \le j \le i}[\delta_{t-1}(j)],$$
$$\text{Termination:}$$
$$\text{total distance} = \delta_N(3),$$
$$\text{Path (class sequence) backtracking:}$$
$$\text{For } t = N-1, N-2, \cdots, 1$$
$$i_t^* = \psi_{t+1}(i_{t+1}^*).$$

i_t^* is the new segment cluster index of the frame point $F0_t$.

Step 4: Decide whether the result satisfy convergence requirement for the total distance. If it does not, repeat step 2 and step 3 until convergence.

3.2 Tone nuclei decision based on knowledge rules

The 3 segments obtained by the above segmentation method are then analyzed to decide if they are possible tone nucleus. Generally speaking, the analyses are carried out by the following methods.

Step 1 (Segment integration): The neighboring segments having similar dynamic features (rising, falling, or flat) 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 50 ms[4].

Step 4 (Location rule): tone nucleus should be in the latter portion of a syllable F0 contour [5].

Step 5: Other rules based on statistical distribution analyses.

Step 6 (Smoothing): Widen the detected tone nucleus region into other two segments if there exist voicing points with features similar to those of tone nucleus.

Most of the above methods are done based on prosodic features. For example, Fig.3 illustrates some of the F0 features used in deciding whether a rising segment in a dip shape F0 contour is the tone nucleus of Tone 2 or not.

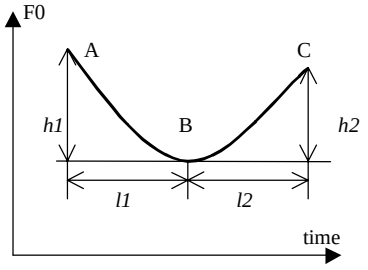

Fig. 3: Illustration of some F0 features for a dip shape contour.

They have meanings like:

- *l1* and *l2:* the lengths of falling and rising segments in voicing frame points.
- *h1:* relative F0 decrease of falling segment.
- *h2:* relative F0 increase of rising segment.

Besides these, there are also other kinds of features used for tone nuclei decision.

Although not all of the possible tone nuclei can be found, most of them have been successfully detected. Effectiveness of tone nuclei detection algorithm can be estimated by its contribution to tone recogniton.

4.TONE RECOGNITION EXPERIMENTS

Tone recognition experiments for continuous speech have been carried out on a female speaker (0f) in data corpus HKU96, published by the University of Hong Kong. 500 utterances titled from cs0f0001 to cs0f0500 were used as training set, 200 utterances from cs0f501 to cs0f700 were used as testing set. The number of syllables is 6419 for training set and 2567 for testing set, and accordingly, 12.84 syllables per a utterance as the average. The utterances have average speech rates ranging from 3.8 to 4.9 per a second, and sometimes local rates exceed 5 syllables per a second. Acoustic feature vector used in the recognition is a 6-element vector consisting of F0, "rms" power and their first and second time derivatives. F0 and "rms" power were calculated for every frame with 10ms step. Normalization by the utterance level was applied to "rms" power.

The corpus offers phoneme, syllable and lexical tone

labels together with sentence text transcriptions. Lexical tone labels were manually checked, since some labels need to be corrected according to its real tonality. For example, due to the well known tone sandhi rules for two Tone 3's, the first one changes to Tone 2. Also, due to vowel intrinsic F0 rules, syllable "Yi" with Tone 1 is usually uttered with a tonality of Tone 4 or Tone 2. Except these two types, no other modifications were made, even though syllable F0 contour is severely altered due to possible neutralization or other reasons.

The number of tone HMMs in context independent (CI) experiment is 6, among them 4 HMMs are for the four basic lexical tones, 1 for the neutral tone and 1 for silence. The number of tone HMMs in context dependent (CD) experiment is 176[1]. Among them 5×5×5 models are for tones at the middle of an utterance, 4×5 for tones at the top of an utterance, 5×5 for tones at the end of an utterance, and 6 for isolated syllables and silence. Continuous density HMMs used in the experiments have a left to right configuration. Number of states is 5 for the 4 basic tones, and 3 for the neutral tone and silence. Mixture number is 6 for middle states and 2 for beginning and ending states in the case of 4 basic tones, and less number of mixtures for the neutral tone and silence HMMs. Due to insufficient training data for the context dependent case, CD HMMs have tied transition matrices. They are first interpolated from CI HMMs, and then re-estimated according to the training data labeled with tri-gram tones.

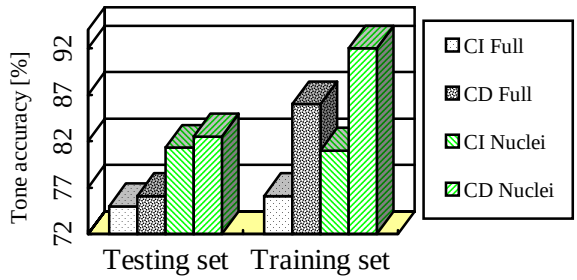

Fig. 4 Tone recognition accuracy of tone recognition experiments.

Recognition results for the conventional method using full syllable F0 contour and the proposed method using tone nucleus are shown in Fig.4 for the both cases of CD and CI HMMs. From the results, we may conclude:

1. The proposed method using tone critical segment features outperformed the conventional method using features of full syllable F0 contour for lexical tone recognition in all the cases. This result is consistent with the results obtained for disyllabic words [2], and, therefore, we can say that the tone-critical segments (tone nuclei) posses key features for lexical tone recognition. Tone recognition based on tone nucleus is a more robust method than the conventional ones.
2. Achievement by the proposed methods also indicates that the tone nucleus detection scheme worked appropriately.
3. CD HMMs increased tone recognition accuracy in both training and testing sets, indicating they are effective in modeling coarticulation effects to some extent. Although the gaps between performances of training set and that of testing set in the case of CD HMMs are probably caused by the insufficient training, we also note that recognition rate increase brought to the testing set by the CD HMMs was not prominent. Only around 1% was obtained in both full syllable F0 contour and tone nucleus cases. This is also consistent with those results reported in [1]. We ascribed this problem to incorrect modeling of tone coarticulation effects by the conventional context dependent models. In [6], further discussion of this question is given with the proposal of new modeling technique of tone coarticulation.
4. Even the highest tone recognition rate (82.5%) for testing set is still rather low for actual use. Besides the possible incorrect modeling of tone coarticulation effects, the reason also lies in that the neutral tone received no specific consideration. When neutral tone is ignored, tone recognition accuracy for the four basic tones reaches above 86%.

5.CONCLUSIONS

We applied a new method focusing on tone nuclei to recognize lexical tones of continuous Chinese speech. Experimental results proved the method is more robust than the conventional ones which observe full syllable F0 contours. They also confirmed the effectiveness of the proposed tone nuclei detection method. Our next work is to search for modeling techniques more suitable for the neutral tone and coarticulation effects.

REFERENCES

[1] H. M.Wang, et al, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulaary using limited training data", IEEE Trans. on SAP, 5, No.2, 1997,195-200.

[2] J.S. Zhang and K. Hirose, "A robust tone recognition method of Chinese based on sub-syllabic F0 contours", ICSLP98, Sydney, Australia, Dec. 1998,703-706.

[3] J.S. Zhang , G. Kawai and K. Hirose, "Subsyllabic tone units for reducing physiological effects in automatic tone recognition for connected Mandarin Chinese", to appear in ICPhS 99, San Francisco, USA, Aug. 1999.

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

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

[6] J.S. Zhang, H. Kawanami. "Modeling carryover and anticipation effects for Chinese tone recognition". to appear in Eurospeech99, Budapest, Hungary, Sept. 1999.