A ROBUST TONE RECOGNITION METHOD OF CHINESE BASED ON SUB-SYLLABIC F0 CONTOURS

Jin-song Zhang and Keikichi Hirose

(zjs,hirose)@gavo.t.u-tokyo.ac.jp
Department of Information and Communication Engineering
School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan

ABSTRACT

This paper proposes a scheme of using F0 contours of vowel nuclei to discriminate Chinese lexical tones. The authors suggest that the F0 contour fragment of a vowel nucleus of a syllable contributes most to tone perception of the syllable. To correlate the F0 contour with the phonemic components of a syllable, a tone-based syllabic structure is also proposed. Tone recognition experiments on a speaker independent dissyllable words task proved the effectiveness of the proposed method. Better performance over approaches observing full syllabic F0 contours indicates that the proposed method is a more robust tone recognition method.

1. INTRODUCTION

A syllable in Mandarin Chinese corresponds to a morpheme, and each syllable is associated with one of four main lexical tones(1st,2nd,3rd and 4th). The total number of the tonally differentiated syllables is around 1300. When the tones are disregarded, the number of the basic syllables is only around 410, indicating that homophonous morphemes are quite numerous. Accurate tone recognition should be helpful for continuous Chinese recognition and understanding[1].

The tones are mainly realized through different patterns of fundamental frequency contours (F0 contours)[2, 3]. But the F0 contour of an utterance may be much more complex than a concatenation of a series of syllabic tones[2, 3, 4]. According to the theory of “contour interaction”, an output F0 contour of Chinese speech results from the interplay of syllabic tones, stress, tune, and intonation[2, 4], and syllabic F0 contours may deviate from the standard patterns due to tone neutralization, tone sandhi, stress, phrase tune and sentence intonation. For examples, tone neutralization which means a syllable loses its tone is very common in spontaneous speech, concatenation of two 3rd tones may change into the concatenation of 2nd and 3rd tones, stress on a syllable of 3rd tone may dip its F0 contour, and etc. Thus a tone recognizer should deal with changes in syllabic F0 contours to be robust.

To overcome the changes in syllabic F0 contours, the reported tone recognizers[1, 5, 6, 11] suggested the following methods to improve recognition performance:

- Robust acoustic features for tone recognition: F0 and its first and second derivatives, short time energy, substitution of F0 offset value for different speakers, etc.
- Robust tone modeling techniques: statistical models like discrete HMMs and continuous HMMs, or neural networks.
- Compensate for tone sandhi and intonation: context dependent modeling techniques, allotonal models, prosodic models, F0 modifications to compensate the declination.

All of the suggested techniques are helpful, but the tone recognition problem remains an unsolved challenge to Chinese speech recognition and understanding.

One common point of the above mentioned methods is using the entire F0 contour of a syllable for syllabic tone recognition. But through F0 contour analysis and perceptional experiments[7, 9],it was found that segments of different locations in a syllabic F0 contour may contribute differently to tone perception for the syllable; the F0 contour segment of the latter portion of the syllable contains critical information for tone perception and may be called the tone-critical segment; and the early portion of the F0 contour is subject to variation. As the lexical tones in an utterance(if they are not neutralized) are considered to be relatively independent of intonation, it is reasonable to assume that the tone-critical segments in the final portions of syllables with tones will retain the basic patterns of the lexical tones.

In light of the results in [7, 8, 9], we suggest using tone-critical segments of syllabic F0 contours as the key feature for tone recognition. Since definition to the tone-critical segments is not clearly given in [7, 8, 9], we define segments of vowel nuclei as the tone-critical segments of the syllables. We also introduce a tone nucleus based syllabic structure for the segmentation. Experiments on disyllabic words indicate the effectiveness of the proposed method.

2. F0 CONTOURS OF SYLLABLE-FINAL NUCLEI

In this section, we will correlate the segmental F0 contours with phonemic components in a syllable for the purpose of tone recognition.

2.1. Four lexical tones

The basic structure of a Chinese syllable is (C)V with a tone assigned to it. The two components of a syllable are ordinarily referred as Initial and Final. The Initial can be a consonant or none, and the Final can be a nuclear vowel (or diphthong), or vowel with a preceding glide, or a vowel plus a dental nasal -n or a velar nasal - ng. No other consonants excepts the dental and velar nasals occur in the syllable-final position.

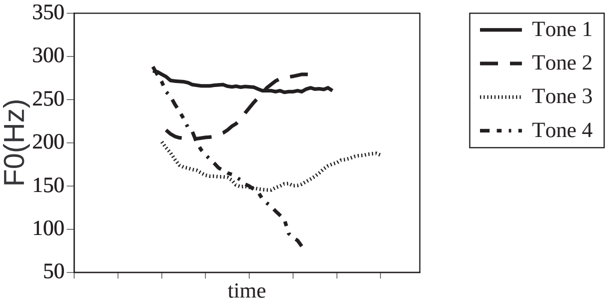

Figure 1. Typical F0 patterns of four basic lexical tones.

The four basic lexical tones in Chinese are mainly realized through difference in their F0 contours[2, 3]. F0 contours in figure 1 depict the typical F0 patterns for the four tones. Distinct features for each tone are generally considered as:

- 1st tone, high level: an almost level F0 contour which is positioned high.
- 2nd tone, high-rising: a rising F0 contour, starting mid and ending high.
- 3rd tone, low-dipping (or falling-rising): a concave circumflex F0 contour, starting mid, falling abruptly down to low, then rising to the mid.
- 4th tone, falling: a falling pitch contour, starting high and ending low.
- A register effect needs to be emphasized to discriminate the 1st and the 3rd tone, because an allotone of 3rd tone with low level F0 contour appears in continued speech[7].

Tonal F0 contours depicted in figure 1 are rarely retained in actual continuous speech. Even in isolated syllables, a syllable with voiced Initial and a velar nasal ending -ng may change the F0 contour into those in figure 2.

2.2. A tone-based syllabic structure

Although [7] described in detail the relationship between different segments of a syllabic F0 contour and tone perception, the exact location of the tone-critical segment was left undefined. A definition is required for the purpose of automatic tone recognition based on the tone-critical segments.

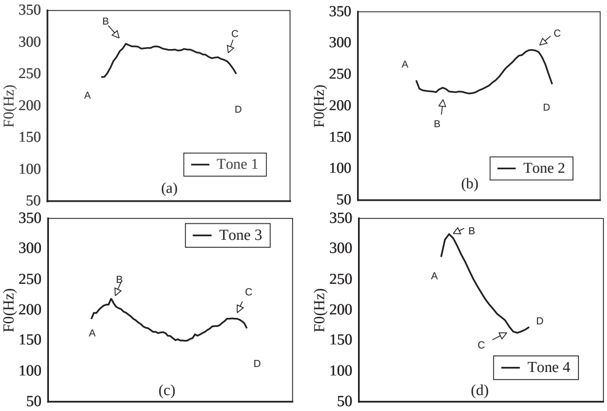

Figure 2. F0 contours of a syllable "meng" for four tones', "meng" has a voicing Initial "m" and a velar nasal ending "-ng".

Based on the reported results in [7, 8, 9] and our observations on both vocal cord and vocal tract events, we believe that "the latter portion of F0 contour of a syllable "mentioned in [7, 8, 9] corresponds to that of Final nucleus of the syllable, whereas "the early portion of the F0 contour of the syllable "corresponds to that of the Initial and transition period between the Initial and Final sounds. We agree with [9] that says, *"In both production and perception, a tone is always aligned with the syllable it is associated with."* This states that there is a correlation between suprasegmental and segmental events in tonal languages.

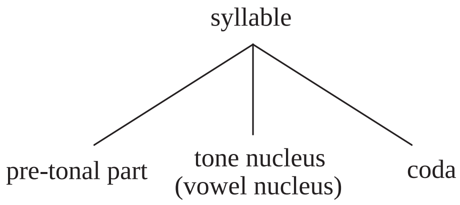

Figure 3. The proposed syllabic structure.

Differing from the claim in [8, 9] that the F0 contour at the end of a syllable converges to a contour that conforms to the underlying tone specifications, we suggest that F0 contours at the end portion of a syllable is subject to variation. Tseng also observed such kinds of variations, which she named "fry register portions"[3]. Such variations may be due to the gliding effect from the preceding tone to the following tone as shown in figure 4, or the resting period of vocal cord vibration[10] as those "C-D" indicated in figure 2.

Figure 3 shows a tone-nucleus-based syllabic structure correlating the relations between segmental F0 contours and underlying phonemic events. A syllable consists of three parts: the pre-tonal part, the tone nucleus (Final nucleus) and the coda part.

- Pre-tonal part: This part spans the Initial (silence is also considered as an Initial) and the gliding period from Initial to the Final nucleus. The F0 contour of this portion may be caused by the transition from the previous tone (figure 4) or a taking off of the vibration of vocal cord from rest points (indicated "A-B" segments in figure 2). It may or may not conform to the syllabic tonal F0 contours. F0 contour of this part contributes very little to the tone perception of a syllable.

- Tone nucleus (Final nucleus): As the F0 contour of the Final nucleus of a syllable carries the syllabic tone information, we call it the tone nucleus. This part may span the major portion of the Final of a syllable. F0 contour and spectral features of this portion differentiate the Final tonally. In continuous speech, F0 contour of this portion should keep the tone-critical features mentioned above if the tone is not reduced. The "B-C" segments of F0 contour in figure 2 can be seen to have consistent patterns similar to those in figure 1.
- Coda part: indicates the coda of the Final of a syllable. The F0 contour of this portion may conform to the syllabic tonal F0 contour or deviate from it. It is also less important for tone perception of the syllable. Examples are shown in figure 2 as "C-D" segments.

The tone-nucleus-based syllabic structure provide us with the segmentation information for F0 contours based on spectral features. Figure 4 illustrates an example of segmentation of a disyllable word. We can see the clear correlation between the spectral features and F0 contours; the F0 contours in the Final nuclei retain the basic forms of their lexical tones, whereas the F0 contours of the other parts do not.

3. TONE RECOGNITION BASED ON F0 CONTOUR SEGMENT OF SYLLABIC FINAL NUCLEI

The segmentation of a syllable into the above-mentioned subsyllabic units can be realized through an HMM phoneme recognizer, whereas the HMMs need to be trained through training data labeled according to the tone-nucleus-based subsyllabic units. After the F0 contour segments of syllabic nuclei are located, they can be used as the cue features for the tone recognition.

3.1. Segmentation using an HMM recognizer

The training data were first hand-labeled according to the subsyllabic units described above, then were used to train HMMs of subsyllabic units. The spectral features used in the experiment were 12th-order mel-frequency cepstral coefficients, log energy and their time derivatives. Final nuclei are segmented using ordinary speech recognition or forced phoneme alignment.

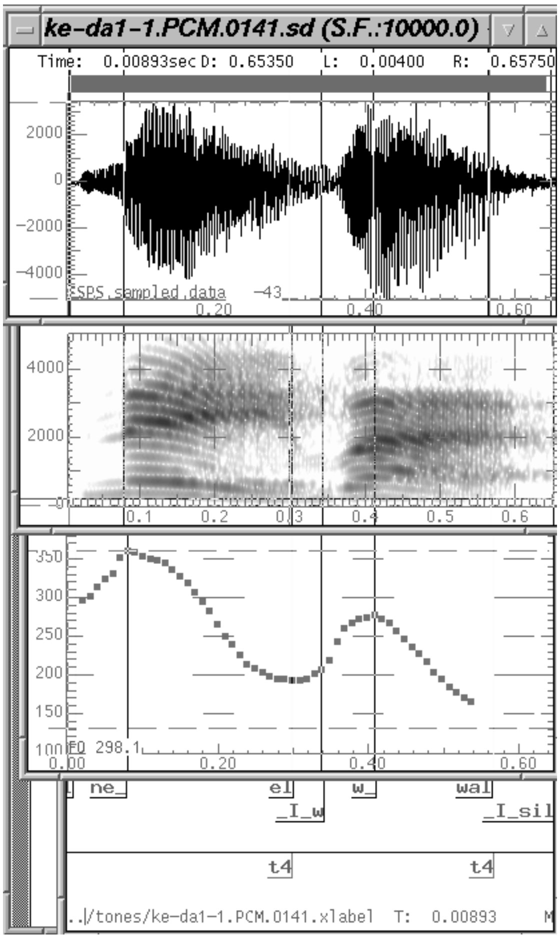

Figure 4. An example of segmentation based on the proposed syllabic structure, where "eI" and "waI" are the nuclei of the respective syllables. Each of F0 contours of the nuclei keeps the basic form of the fourth tone(falling tone).

In our experiment on disyllabic words, phonemic models for pre-tonal parts are context dependent of the head of the Finals, models for the Final nuclei are context independent, and models for the codas are context dependent with respect to the Initials of their succeeding syllables. Segmentation information for the subsyllabic units is a by-product of speech recognition.

3.2. Tone recognition

Only the four basic tones were recognized in this experiment, as the fifth tone (Tone 0) cannot be correctly recognized solely from F0 contours. We adopted HMMs based on their demonstrated good performance [1, 5]. Specifications of the tone recognizer include:

- Number of models: one model for each tone type, totally 4. No allotonal models were used in the experiment.
- Parameters: log F0 and its first- and second- order time derivatives.
- Adaptation method for different speakers: substitution of speaker's offset calculated as the average F0 value of F0 contour fragments of the 1st Tone .

In the tone recognition stage, the F0 contour fragments of the Final nuclei are matched with the tone models, and those of other parts are ignored.

3.3. Experimental Results

A Chinese disyllable corpus consisting of the utterances of 80 Chinese disyllables read by 11 speakers (8 females and 3 males) were used to test the method. This corpus belongs to the Putonghua corpus created in University of Science and Technology of China. The tone concatenations of the vocabulary cover all of sixteen bi-tonal sequences, and tone of second syllable of some words was neutralized by some spekaers. The speakers were from different areas in China. Each speaker read each word once.

Segmentation accuracy depends on the training of the HMMs for the phonemic models. In order to get good location information for tone nuclei, a larger training set was used in the segmentation experiment than in the tone recognition; only 2 females and 1 males remained at the test set. Though the recognition accuracy for disyllabic words disregarding tones was acceptable(99.6% for the training set and 96.4% for the test set), we corrected some major errors of segmentation which was mainly due to an insufficient training dat for the neutralized tones.

In the tone recognition experiment, data of 3 females and 2 males were used as the training set, and data of the remaining 6 speakers formed the test set. The average recognition accuracy for four basic tones was 97.63% for the training set, and 98.16% for the test set. To evaluate the performance of the proposed method, we also present here the results of comparative experiments where full F0 contour was used for recognition. They are:

- Method 1: syllabic tone models only for four basic tones.
- Method 2: additional allotonal models were introduced into method 1 for Tone 3 and Tone 4.
- Method 3: bi-tonal models were introduced to compensate for the coarticulation effects.

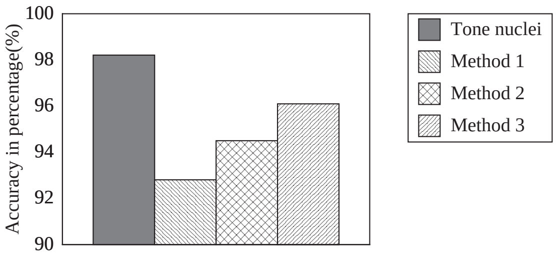

Figure 5. Tone recognition accuracy of the 4 experiments.

Figure 5 depicts the tone recognition accuracy for the four methods. Findings from the experimental results are:

(1) The proposed method exceeded all three tone recognition methods using full syllabic F0 contour. This indicates the robustness of the proposed method.

(2) Performance improved when allotonal models were used and compensation of bi-tonal coarticulation effects were compensated for.

(3) It is reasonable to presume that the performance of the tone-nucleus F0 contour based method may be further improved using additional allotonal models and compensations for the tonal coarticulation.

4. CONCLUSION

We suggested that the F0 contour segment of the Final nucleus of a syllable is the tone-critical segment of a syllabic F0 contour, and proposed using spectral features to locate it. A tone recognition experiment showed that the subsyllabic F0 contour of the final nucleus ia a robust cue discriminating Chinese lexical tones in disyllable words. We plan to apply this method to tone recognition for continuous speech.

REFERENCES

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

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

[3] Chiu-yu Tseng "An acoustic phonetic study on tones in Mandarin Chinese" Special publications No.94, Academia Sinica, Taiwan, 1990.

[4] X.N. Shen "The prosody of Mandarin Chinese" University of California Press, 1990.

[5] X.H. Hu and K. Hirose "Tone recognition of Chinese disyllables using Hidden Markov Models" Trans. Information and Systems of IEICE, Vol. E78-D, No. 6, June 1995, pp.685-691.

[6] L.Zhao, Y. Kobayashi and Y. Niimi "Tone recognition of Chinese continuous speech using continuous HMMs" in Japanese, Journal of Japan Association of Acoustics, Vol.53, No.12, 1997, pp.933-940.

[7] D. H. Whalen and Yi Xu, "Information for Mandarin Tones in the Amplitude Contour and in Brief Segments" Phonetica, 1992,49, pp.25-47.

[8] Yi Xu, "Contextual tonal variations in Mandarin" Journal of Phonetics, 1997,25, pp.61-83.

[9] Yi Xu and Q.E.Wang "What can tone studies tell us about intonation?" ESCA Workshop on Intonation: theory, Models and Applications, Athens Greece, Sept. 1997, pp.337-340.

[10] J. Ni and R. Wang "Modeling the control mechanism for generating the rise-fall pattern in F0 contours", ACTA Acoustic, 1996,Vol.21, No.6, pp.863-871.

[11] Y.R. Wang and S.H. Chen "Tone recognition of continuous Mandarin speech assisted with prosodic information" JASA. 96(5), Pt.1, Nov. 1994, pp.2637-2645.