A MULTILEVEL FRAMEWORK TO MODEL THE INHERENTLY CONFOUNDING NATURE OF SENTENTIAL F0 CONTOURS FOR RECOGNIZING CHINESE LEXICAL TONES

Jin-Song Zhang[†], *Keikichi Hirose*[‡] *and Satoshi Nakamura*[†]

[†] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
[‡] Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan
{jinsong.zhang, satoshi.nakamura}@atr.co.jp, hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper presents a multilevel framework to cope with the complex variations in Chinese sentential F0 contours in order to recognize lexical tones. Tone nucleus model is to get rid of the influence of intrinsic F0 transition loci at sub-syllable level. Pitch anchoring concept is used to normalize tonal F0 contours at syllable level. Hypo- and Hyper- intonation model is to account for the interplay of tone coarticulation and higher level prosodic effects. The whole approach achieved significant higher performance than the conventional method.

1. INTRODUCTION

Chinese is known as a tonal language, in which each syllable is associated with one of four pitch patterns which have phonemic roles. The same syllables with different pitches become different morphemes or different words. Automatic recognition of tones from the fundamental frequency (F0) contours have a number of applications: Chinese speech recognition and understanding, voice name dialing, prosodic labeling (tones) of available speech database, computer aided language learning (CALL) systems for foreigner Chinese learners and etc..

In the past decades, a number of studies have been made to this issue [1, 2, 3, 4, 5]. However, except that the task of isolate syllables or short length words (2∼4 syllables) was easy to achieve high performances [1, 2], other studies showed that tone recognition was rather difficult for the continuous speech [3, 4, 5], with a performance rather lower than that of word recognition. The major reason can be ascribed to the complex variations existing in the sentential F0 contours, which originate from the mechanical-physiological realization of compound intonation functions [6, 7]. On the one hand, besides the linguistic information like tones and stress, F0 is also used to convey complex para-linguistic information like focus and prosodic phrasing, non-linguistic information from speaker's emotion and age [6, 8]. Thus the nature of information in the one-dimensional acoustic feature F0 is *inherently confounding*[7]. On the other hand, F0 contours, which reflect the periodicals of the successive human vocal cords' vibrations, are to vary due to *articulatory constraints* that determine how the intonation functions can be implemented[6].

Due to these reasons, we believe that an efficient method for recognizing lexical tones should own proper modeling power to deal with the complex variations from either intonation or mechanical-physiological originations, besides building statistical models for tones only. And we have carried out a series of studies at different levels, including sub-syllable F0 segmentation, syllable level normalization, and syllable and phrase level coarticulation, in order to deal with F0 variations from various factors for tone recognition in recent years.

In [9], we focused on the F0 variations of intrinsic F0 loci, which are produced non-deliberately but occurring as the transition loci to the deliberate F0 targets. We proposed to recognize tones using only the features of *Tone nucleus* in order to prevent the interference from them. In [11, 12], we proposed that pitch anchoring might play an important role in discriminating Chinese lexical tones. The proposal suggested that F0 heights normalized by the neighboring tonal F0s can be used for tone recognition. In [10], we focused on the F0 variations of tonal coarticulations and the influences from higher level events. Although rather stable coarticulation patterns, either assimilatory or dissimilatory, may exist to a pair of neighboring tonal F0 contours, they may be interfered by some kinds of "*break strengths*", which seem to be possible phrase boundaries, or other higher-level prosodic events. We referred to the normal tonal coarticulations as hypo-articulations, and the broken articulation as hyper-articulations. Each of these proposals separately brought about significant improvement to tone recognition performances. But a fully integration of the separate approaches have not yet been presented.

This paper is to give the whole framework combining all the three isolate approaches for tone recognition. At subsyllabic level, tone nucleus is used to against influence from intrinsic F0 loci; At syllable level, tonal F0s are normalized by neighboring tonal F0s; At above syllable level, Hyper-context dependent tonal HMMs were adopted to model the "breaking" effect to train more accurate tonal HMMs. For the same test set, the combination approach achieved the highest tone recognition performance.

THIS RESEARCH WAS SUPPORTED IN PART BY THE TELECOMMUNICATIONS ADVANCEMENT ORGANIZATION OF JAPAN.

2. MODEL MULTI-LEVEL F0 VARIATIONS

The four basic lexical tones are known to have different F0 patterns: Tone 1 with high and flat F0 contour, Tone 2 with a rising contour, Tone 3 with a low and dipping contour and Tone 4 with a falling contour. However, except for isolated ones, tones in continuous speech can rarely take these standard patterns. Both their shapes and height are subject to severe variations. The following three models are proposed to deal with F0 variations of different level factors: sub-syllable F0 transition loci, tonal coarticulations and high-level initiation effects.

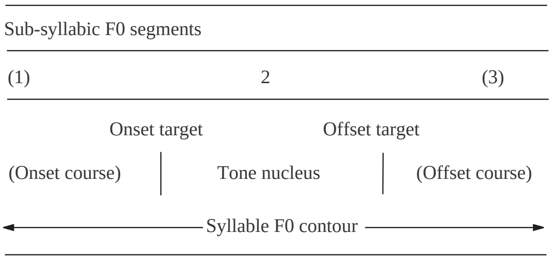

Fig. 1. Illustration of Tone Nucleus Model. Optional F0 segments are indicated by parentheses, only the tone nucleus is obligatory.

2.1. Tone Nucleus Model

Tone nucleus model is a F0 segmental structure model for systematically accounting for F0 variations at syllabic level [9]. As illustrated in Fig. 1, it suggests that a syllable F0 contour may consist of three segments: onset course, tone nucleus and offset course. Among the three segments, only the tone nucleus is obligatory, whereas the other two are intrinsic F0 transition loci, which are articulatory transition F0s non-deliberately produced, and their appearances are optional.

- Tone Nucleus: the segment contains the most critical information for tone perception. The beginning and ending points of a tone nucleus correspond to the Tone onset and Tone offset, which may take pitch values as given in Table 1.

targets	Tone 1	Tone 2	Tone 3	Tone 4
Onset	H	L	L	H
Offset	H	H	L	L

Table 1. Pitch targets of the four lexical tones. "H" and "L" depict high and low pitch targets respectively.

One merit of tone nucleus model is that it provides a systematic view of sub-syllabic F0 variations which originate from segmental phonations. A syllable with a voicing Initial may show a significant F0 transition locus to its onset target, whereas a syllable of the same tone may not has the locus due to its voiceless Initial part. Under the framework of Tone Nucleus model, we may only focus on the Tone Nuclei for recognizing tones, while discarding other transition loci. For example, the second syllable [yi4] in the example utterance has a significant rising locus to its high onset target, shown in (b) of Fig. 2. If we discard the transition F0 loci and only keep tone nuclei, as shown in (c) of Fig. 2. The left F0 contours seem to conform more to the standard tonal F0 patterns than the original ones.

Statistical distribution studies showed that tone nuclei have reliably distributive features, and an automatic approach has been developed in [9] to detect tone nucleus for each syllable when phone-level segmentation is available.

2.2. Anchoring-based F0 Normalization

Although F0 heights (in Hz or logarithm) are assumed to be useful for discriminating between tones of different register levels (such as Tone 1 and Tone 3), numerous well-known phenomena showed that it should be not the right way. For example, it is common that F0s of low pitch by a female speaker may be still much higher than the F0s of high pitch by a male speaker. Furthermore, even in one utterance, a high-pitch in a latter position may own lower F0s than a low-pitch tone in a former position. For the example in (b) of Fig. 2, we could not discern any significant differences between the F0 heights of the sentence-beginning syllable *wo3*, which has low pitches, and those of the sentence-ending syllable *hao4*, which has a high onset pitch.

Aiming at finding a more efficient feature for discriminating the lexical tones, we adopted the psycho-acoustic perception findings to make the following anchoring hypothesis [11, 12] :

- Relative F0 difference between the offset point of the first lexical tone and the onset of the second lexical tone may be an important discriminating cue for high or low pitch, besides the direct cue of a gliding F0 contour.

Based on this hypothesis, a lexical tone in continuous speech can also be acoustically characterized using the patterns given in Table 2, besides using the flat, rising, dipping or lowering F0 patterns.

Lexical Tones	Onset gap	Offset gap
Tone 1	≥ 0	≥ 0
Tone 2	≤ 0	≥ 0
Tone 3	≤ 0	≤ 0
Tone 4	≥ 0	≤ 0

Table 2. Anchoring based feature patterns for the four basic lexical tones in continuous speech.

- Onset gap: the difference between the onset F0 and the offset F0 of preceding lexical tone.
- Offset gap: the difference between the offset F0 and the onset f0 of succeeding lexical tone.

Furthermore, we proposed two methods to normalize a syllable F0 contour in continuous speech. For the ith frame in one lexical tone,

- Left-to-right: $\log F0'_i = \log F0_i - \log F0$ of the preceding tone offset,
- Right-to-left: $\log F0''_i = \log F0_i - \log F0$ of the succeeding tone onset.

The panels (d) and (e) in Fig. 2 illustrate the F0 contours of Left-to-right and Right-to-left normalizations respectively. In (d), H onsets stay higher or nearby 0, and L onsets lower or nearby 0. We may note that the 4th and 6th syllables, both of Tone 3, originally have higher onset values than the final syllable of Tone 4 in (c). But turned out to go to lower regions than the final Tone 4 in (d). Similarly, normalized F0 contours in (e) also show to be consistent with the anchoring patterns in Table 2. The normalized F0 contours $\log F0'$ and $\log F0''$ can be combined with the normal F0 features to do tone recognition.

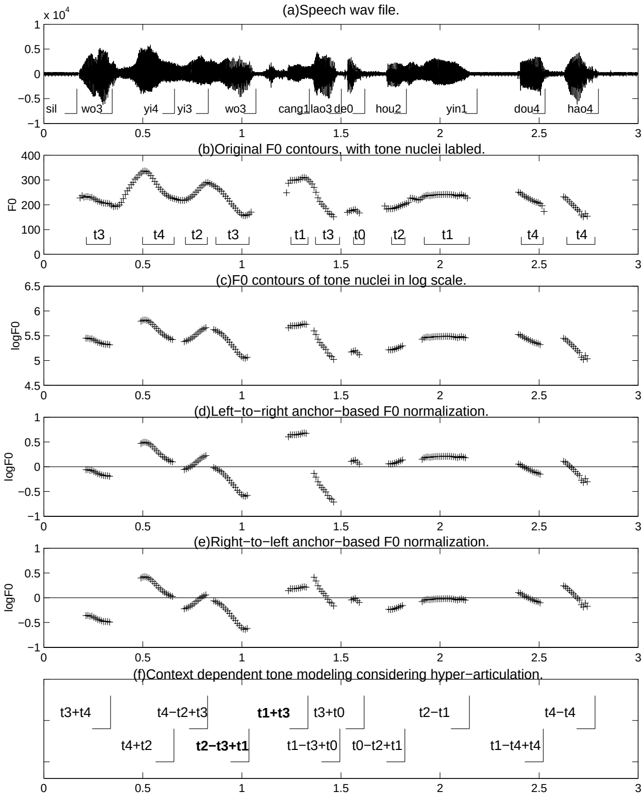

Fig. 2. Illustration of sentential F0 processing under the proposed multi-level framework.

2.3. Hypo- And Hyper-articulation F0 Model

It is well known that tonal F0 contours are subject to contextual variations. However, few studies were able to show a clear image about F0 variations due to an interplay of tonality, contextual tone and high-level prosodic events such as foci and phrasing structures. In our approach for tone recognition, we proposed an efficient framework to guide training of tonal acoustic models, which was named as Hypo-articulation and Hyper-articulation F0 model.

- Hypo-articulation: there seems to be one specific coarticulation F0 pattern for any pair of tones, which perhaps results from the economical articulation rule.
- Hyper-articulation: high-level events may act as a force to break a hypo-articulation.

Offset of	Onset of			
	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	A, A	D, A	D, A	A, A
Tone 2	D, D	D, A	D, A	D, D
Tone 3	D, A	D, A	D, A	D, A
Tone 4	A, A	D, D	D, D	A, A

Table 3. Defined Hypo-articulation patterns in our training of tonal acoustic models. "A" stands for assimilatory effect, while "D" for dissimilatory effect.

Table 3 gives the defined hypo-articulation patterns for each pair of the basic lexical tones with respect to the onset F0 of the first lexical tone and the offset of the second tone. Assimilatory effect indicates that the preceding offset and the succeeding onset show to be assimilated. Dissimilatory effect indicates that the two points appear to depart from each other. For the last two syllables in (c) of Fig. 2, the L offset of the 1st Tone 4 was raised to a higher place, while the H onset of the 2nd Tone 4 was dragged to a lower place, thus they are assimilated. If two neighboring tones do not show their hypo-articulation pattern, they are regarded to be hyper-articulated due to some higher-level effect. For the 4th and 5th syllable tones in (c) of Fig. 2, the onset of the Tone 1 was not lowered, thus hyper-articulated.

Tri-tone context dependent (CD) models can be used to model the hypo-articulation F0 variations. And [10] proposed to use mono-tone, bi-tone to model the hyper-articulation F0 variations. For the mentioned hyper-articulated pairs, a bi-tone *t1+t3* was used to model the broken hypo-articulations between the 4th and 5th tones, instead of *t3-t1+t3*, as shown in (f) of Fig. 2.

3. TONE RECOGNITION EXPERIMENTAL RESULTS

3.1. Experimental Set-up

Tone recognition experiments were carried out on data of a female speaker (0f) in the corpus HKU96. 500 utterances from cs0f0001 to cs040500 were used as training set, while 200 utterances from cs0f0501 to cs0f700 were used as testing set. F0 was extracted by integrated F0 tracking algorithm (IFTA) with a frame shift of 10ms. Phonetic segmentation were assumed available, and achieved by force alignment of the database using Initial and Final acoustic models. Tone nuclei were detected using the algorithm introduced in [9].

Continuous density HMMs with left-to-right configuration were used as lexical tone models. The number of states for the four basic tones is 5, and that for the neutral tone is 3. Mixture number is 6 per state. The standard feature vector has $\log F0$, frame energy and their 1st, 2nd order time derivatives.

3.2. Tone Recognition Experiments:

Comparison recognition experiments have been made with respect to the factor of different acoustic features and the factor of different context dependent strategies. The feature specification includes three kinds:

- Full syllabic features: Acoustic features of the whole syllables are used.
- Tone nucleus I: Acoustic features of the tone nuclei are used.
- Tone nucleus II: Anchoring-based normalized F0 features: $\log F0'$ and $\log F0''$, were appended to the standard feature vector.

The context dependent strategies include:

- CI: Context independent tonal HMMs. There are only 5 tonal HMMs.
- CD: Context dependent tonal HMMs as in [4]. The number is 176.
- CDH: Context dependent tonal HMMs developed under the framework of Hypo- and Hyper-articulation. The number of HMMs is 235, including the 176 CD ones plus other 59 additional ones with the context of utterance boundaries.

Tonal	Recognition correct rates (%)		
HMMs	Full syllable	Nucleus I	Nucleus II
CI	75.3	81.5	85.5
CD	76.2	83.1	85.6
CDH	79.1	85.7	87.3

Table 4. Average correct rates for the four basic and the neutral tones in all nine tone recognition experiments.

Method	T1	T2	T3	T4	Avg.
CI Full syllable	69.2	76.4	70.0	85.3	75.2
CI Nucleus I	83.6	84.5	68.0	90.7	81.7
CI Nucleus II	87.9	87.9	84.6	91.0	87.9
CDH Nucleus II	87.0	89.7	93.2	92.7	90.7

Table 5. Correct rates for the four basic lexical tones in four representative recognition experiments.

3.3. Discussions

Table 4 summarizes recognition results in tone correct rates for all nine experiments, each for one combination of a kind of feature and a kind of context dependency. Table 5 give detailed performances for the four basic lexical tones in four representative experiments. Based on the results, we say,

- Using tone nuclei to recognize tones improved performances significantly. In the approach of CI HMMs, the technique brought by 6.5% absolute improvement for the four basic lexical tones when compared with the standard approach using full syllabic features.
- The anchoring-based normalization of syllable F0 contours brought further significant improvements to the approach of tone nucleus. In the approach of CI HMMs, it gave another 6.2% absolute improvement to the tone nucleus approach for the four basic lexical tones.
- The conventional tri-tone CD approach brought improvements about 1% in the experiments when compared with the CI HMMs. We considered these improvements as limited, and we mentioned that similar results reported in other site [4]. We regarded that the conventional CD be inappropriate to model the coarticualtion variations of lexical tones due to the confounding interplay nature of F0 feature.
- The CDH HMMs brought more improvements than the CD HMMs when compared with CI HMMs. When using the anchoring-based normalized features, the CD HMMs only got slight 0.1% improvements than CI, while the CDH still got 1.8% improvements. This indicates that the Hypo- and Hyper-articulation framework is more efficient to cope with syllable and higher-level coarticulation variations to develop robust tone HMMs.

The whole framework finally achieved 90.7% recognition rates for the four basic tones, equal to a relative error reduction of 62.5% compared to the 75.2% of CI HMMs using full syllabic features.

4. CONCLUSION

This paper presents a multi-level framework to cope with the complex sentential F0 variations for recognize Chinese lexical tones. The significantly improved tone recognition performances showed its efficiency. In the future work, we will apply it to the task of speaker independent recognition task, and incorporate it to a large vocabulary Chinese speech recognition system.

5. REFERENCES

[1] W.-J. Yang, J.-Ch. Lee, Y.-Ch. Chang, H.-Ch. Wang, ”Hidden Markov Model for Mandarin lexical tone recognition”, IEEE Trans. on ASSP, Vol. 36, No. 7, July, 1988, pp.988-992.

[2] Ch.-F. Wang, H. Fujisaki and K. Hirose, ”Chinese four tone recognition based on the model for process of generating F0 contours of sentences”, Proc. of ICSLP90, pp.221-224.

[3] Y.-R. Wang and S.-H. Chen, ”Tone recognition of continuous Mandarin speech assisted with prosodic information”, J. Acoust. Soc. Am. 96(5), Pt. 1, Nov. 1994, pp.2637-2645.

[4] H.-M. Wang, T.-H. Ho, R.-Ch. Yang, J.-L. Shen, B.-R. Bai, J.-Ch. Hong, W.-P. Chen, T.-L. Yu, and L.-sh. Lee, ”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, 1997, pp. 195-200.

[5] ,Y. Cao, Y.-G. Deng, H. Zhang, T.-Y. Huang and B. Xu, ”Decision tree based Mandarin tone model and its application to speech recognition”, ICASSP, 2000, pp.1610-1613.

[6] H. Fujisaki, ”Prosody, Models, and Spontaneous Speech”,In Y. Sagisaka, N. Campbell and N. Higuchi, editors, Computing Prosody: computational models for processing spontaneous speech, New York: Springer-Verlag, 1997. pp.27-42.

[7] B. Granstrom, ”Applications of Intonation - An overview”, ESCA workshop on Intonation: Theory, Models and Applications, Athens Greece, Sep. 1997, pp.21-24.

[8] Y. Xu, ”Effects of tone and focus on the formation and alignment of F0 contours”, Journal of Phonetics, Vol.27, No.1, 1999, pp.55-105.

[9] K. Hirose and J.-S. Zhang, ” Tone recognition of Chinese continuous speech using tone critical segments ”, Proc. of Eurospeech’99, Budapest, Hungary, Sept. 1999, pp.879-882.

[10] J.-S. Zhang and H. Kawanami, ” Modeling carryover and anticipation effects for Chinese tone recognition ”, Proc. of Eurospeech’99, Budapest, Hungary, Sept. 1999, pp.747-750.

[11] J.-S. Zhang and K. Hirose, ”Anchoring hypothesis and its application to tone recognition of Chinese continuous speech”, Proc. of ICASSP, 2000, pp.2741-2744.

[12] J.-S. Zhang, S. Nakamura and K. Hirose, ”Discriminating Chinese lexical tones by anchoring F0 features”, Proc. of ICSLP 2000, Vol. II, pp. 87-90.