Robust Acoustic Modeling of Contextual Tonal F0 Variations on the Basis of Tone Nucleus Framework

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

[†] ATR Spoken Language Translation Research Laboratories
[‡] Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
{jinsong.zhang, satoshi.nakamura}@atr.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents our tone nucleus based framework to model the complex variations in sentential F0 contours to build robust tonal acoustic models. By focusing on tone nuclei, the approach can get rid of the influence of artículatory transition loci on tonal modeling. Anchoring-based Tone nucleus normalization improves the discriminability among the tones. 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 of tone recognition than the conventional method.

1. Introduction

The major difficulty to build robust tonal acoustic models from the fundamental frequency (F0) contours can be ascribed to the complex variations in the sentential F0 contours, which originate from the mechanical-physiological realization of compound intonation functions [1, 2]. On the one hand, besides the lexical tone, F0 is also used to convey high-level linguistic information like stress, focus and prosodic phrasing, non-linguistic information from speaker's emotion and age [1, 3]. Thus the nature of information in the one-dimensional acoustic feature F0 is *inherently confounding*[2]. 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[1]. Therefore, we believe that tonal models should own proper modeling power to deal with the complex variations from both high-level intonation and mechanical-physiological originations, besides building statistical models for tones only. The methods we proposed includes 3 different levels: at syllabic level, *tone nucleus* is considered as the characteristic manifestation of the tonality. At syllable level, *anchoring* based pitch normalization helps to extract more efficient discriminating features. At higher-level, the interaction of tonal coarticulations and higher-level intonation functions is represented by a framework we called *hypo- and hyper-articulations*. Among the three levels, *tone nucleus* provides the basis for the other two techniques.

2. Tone Nucleus Model

Tone nucleus model is a F0 segmental structure model for systematically accounting for F0 variations at syllabic level [5, 6]. 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.

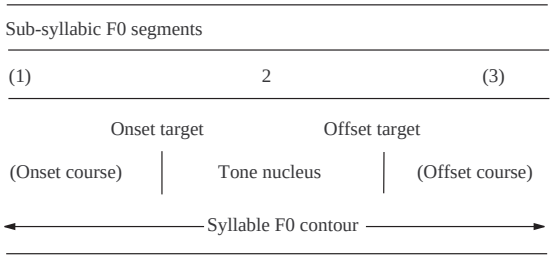

Figure 1: *Illustration of Tone Nucleus Model. Optional F0 segments are indicated by parentheses, only the tone nucleus is obligatory.*

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

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

- 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.

Tone-nucleus model offers a possible systematic framework to deal with F0 variations that result from both articulatory constraints and confounded intonation functions, for tone recognition and intonation function detection. Fig. 2 illustrates F0 contours of two continua of "Tone 2 Tone 2". Due to the voicing (or not) on the initial segment of the second syllables, F0 contours of the second syllables in the two continua differ greatly. The second Tone 2 in (a) has a substantial lowering F0 locus which makes the whole F0 into a lowering-rising dipping shape, very similar to those of Tone 3s. However, according to the Tone-nucleus model, F0 loci "BC" in (a) and (b) of Fig. 2 are the tone nuclei, whereas the F0 locus "AB" in (a) is an articulatory transition. If the articulatory transition is ignored in tone modeling, the above problem may be avoided.

It was found that when a prosodic boundary exists between two neighboring tones, F0 contour of the second

This research was supported in part by the telecommunications Advancement Organization of Japan.

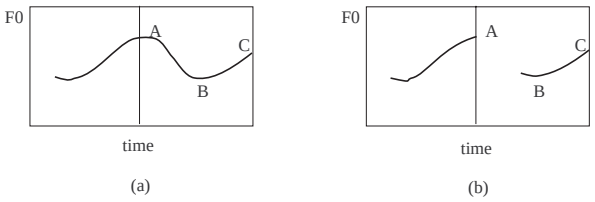

Figure 2: *Illustrations of F0 contours of continua of "Tone 2 Tone 2", with a contrast of voiced initial segment (a) and unvoiced initial segment (b) in the second syllables. The thin vertical lines indicate the syllable boundaries.*

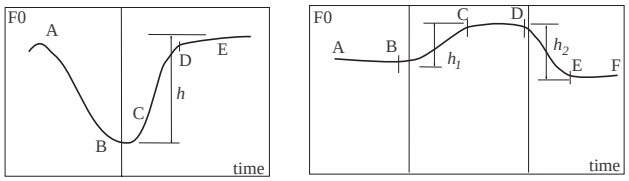

Figure 3: *Illustration of high-level intonation functions on tonal F0 contours.*

syllable is usually free from coarticulation effect from the first syllable [7]. Given a continuum of Tone 4 and Tone 1, "H" targets in Tone 1 are usually rather lower than that of the "H" target in the preceding Tone 4 mainly due to the carryover lowering effect from the "L" offset target of Tone 4 when the two tones are in one word. But when there exists a prosodic boundary (such as a phrase boundary) between the two tones, the carryover lowering effect may be stopped and the "H" targets of Tone 1 are raised to a higher position, as illustrated in (a) of Fig. 3. This often results in a substantial rising transition F0 locus "CD" to the "H" targets of Tone 1. However, based on the Tone-nucleus model, the rising "CD" F0 locus is an articulatory transition, and the tone information is mainly carried by the tone nucleus "DE" whose shape still conforms to the underlying pitch targets. Furthermore, if the tone nuclei "AB" and "DE" of the two tones can be detected, then the F0 range reset h of the Tone 1 should be an important cue for detecting the prosodic phrasing boundary.

Sentential focus may also bring about substantial F0 variations to syllable F0 contours. According to [3], focus is related with three distinct pitch ranges: expanded range in non-sentence-final focused words, suppressed in post-focus words, and neutral in all other words. The (b) of Fig. 3 illustrates this idea through a continuum of three Tone 1s with a focus on the second Tone 1. We can see that the focus lead to a substantial rising and a falling F0 loci in the early portions of F0 contours of the second and the third syllable respectively. These kinds of F0 loci are regarded as articulatory transitions according to the Tone-nucleus model, whereas the segments "AB", "CD" and "EF" are the tone nuclei of the three Tone 1s. Compared with the whole syllable F0 contours, the tone nuclei show more consistent patterns with the underlying pitch targets. Also, if the three tone nuclei can be detected, the range differences represented by h_1 and h_2 are estimated as the gaps between the preceding tone offsets and succeeding tone onsets of neighboring tone nuclei. These range differences may serve as acoustic cues for focus detection, which is important for interpreting the speaker's intention for developing spontaneous dialogue systems.

Statistical distributional analyses showed that there are rather stable distributions with respect to the location and energy features of tone nuclei [6]. A proposed tone nucleus detection method has achieved a correct detection rate of 97.5% for a speaker dependent continuous task [6]. The panel (c) of Figure 4 showed the detected tone nuclei of the original F0 contours in the panel (b).

3. Anchoring-based F0 Normalization

Tone nucleus model helps to get rid of the influences from articulatory transition loci for building tonal models. However, in continuous speech, both the F0 heights and slopes of tone nuclei may vary significantly from the standard tonal F0 patterns, even making the underlying tonalities difficult to be identified from the surface F0 contours. For the example in (b) of Fig. 4, 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. On the other hand, perception experiments showed human beings are able to perceive the purported underlying lexical tones with high consistency despite of the substantial F0 variations, provided the tonal context [4]. This indicates that there exist other discriminating cues in the tonal context besides the F0 height and F0 slope coefficients.

Aiming at finding a more efficient feature for discriminating the lexical tones, we adopted the psycho-acoustic perception findings [10, 11] to make the following anchoring hypothesis [8, 9] :

- 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.
- There should be a timing allocation mechanism for the competition effects [11].

Based on this hypothesis, a lexical tone in continuous speech can also be acoustically characterized using relative pitch values, besides using the flat, rising, dipping or lowering F0 patterns. An *H* pitch target should usually have *positive* relative pitch values, whereas an *L* target have *negative* relative values. A statistical distributional study [9] have been made on gender balanced data of 20 speakers, using the speaker based normalized z features.

$$z = \log F0_{norm} = \frac{\log F0 - \overline{\log F0}}{\hat{\sigma}_{\log F0}} \qquad (1)$$

For each tone, two z values are collected:

- z_0 for tone onset,
- z_1 for the tone offset.

And four anchoring features are calculated.

- $z_i^L = z_i - z_1$ of the preceding tone: indicate the left-to-right anchoring effect, $i = 0, 1$.
- $z_i^R = z_i - z_0$ of the succeeding tone: indicate the right-to-left anchoring effect, $i = 0, 1$.

Table 2: *Tone based group mean values of the collected features.*

Tone	z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
Tone 1	.912	.874	.913	.876	.496	.458
Tone 2	-.384	.386	-.475	.295	-.866	-.009
Tone 3	-.322	-1.306	-.686	-1.670	-.433	-1.418
Tone 4	.596	-.8108	.850	-.557	.916	-.491

Table 3: *Tone onset based group statistics.*

Tone onset		z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
mean value	L	-.353	-.460	-.580	-.688	-.650	-.757
	H	.754	0.003	.882	.159	.706	-.002
F(1,78)		477.7	6.2	652.5	17.3	283.1	26.6
P		.000	.015	.000	.000	.000	.000

Table 4: *Tone offset based group statistics.*

Tone offset		z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
mean value	L	.137	-1.059	0.008	-1.114	.242	-.954
	H	.264	.630	.219	.586	-.185	.181
F(1,78)		.886	486.6	.62	213.4	6.6	115.7
P		.349	.000	.434	.000	.012	.000

Table 2 listed the mean z values. Table 3 is ANOVA analyses with respect the onset target (L/H) while pitch values of offset are collapased, i.e., the L group includes samples from tone 2 and 3, while H group includes tone 1 and 4. Table 4 is ANOVA analyses with respect to the offset target. Conclusions include:

- The H is usually associated with positive pitch values, and L with negative ones.
- The statistics showed that it is very evidential that the proposed anchoring features are discriminative for the H and L targets, either on the onset or offset.
- The discriminating efficiencies of the six features for H and L onset targets can be given based on the F ratio of between-group variance and within-group variance:

$$z_0^L > z_0 > z_0^R > z_1^R > z_1^L > z_1$$

- The discriminating efficiencies of the features for the H and L offset targets can be ordered as:

$$z_1 > z_1^L > z_1^R > z_0^R > z_0 > z_0^L$$

Intuitively, anchoring-based tone discrimination could provide consistent accounts for the fact that human perception showed consistent discrimination of the tones undergoing substantial F0 variations such as downstep lowering and contextual assimilation [12]. For the purpose of tone recognition, we proposed to include two additional normalized F0 features to build tonal acoustic models. 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. 4 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. The normalized F0 contours $\log F0'$ and $\log F0''$ can be combined with the normal F0 features to do tone recognition.

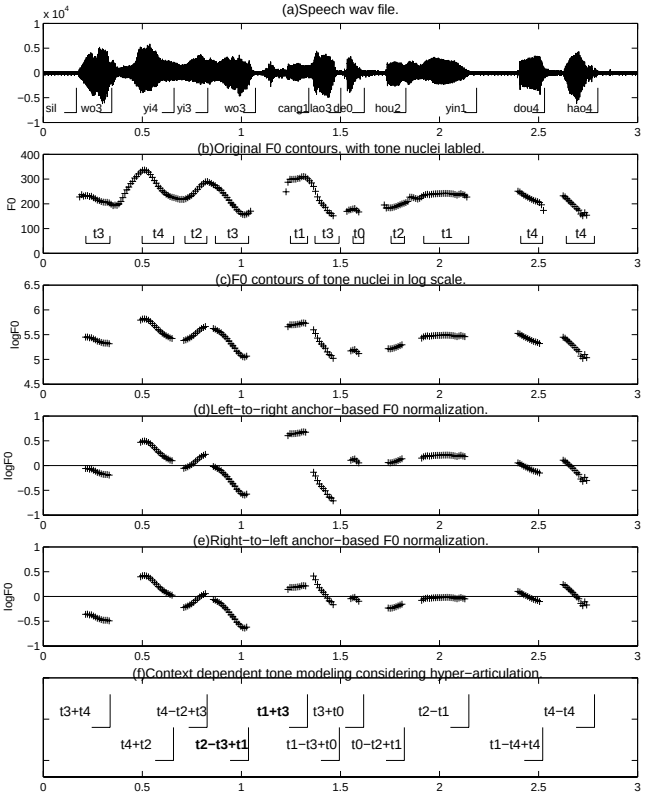

Figure 4: *Illustration of sentential F0 processing under the proposed multi-level framework.*

4. Hypo- And Hyper-articulation F0 Model

The third method is proposed to model the interplay of tonality, contextual tone and high-level prosodic events such as foci and phrasing structures. The method regards the sentential F0 contours as a continuum of two kinds of tonal coarticulation, after [13]:

- 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.

Table 5: *Defined Hypo-articulation patterns in our training of tonal acoustic models. "A" stands for assimilatory effect, while "D" for dissimilatory effect.*

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 5 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. 4, 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. 4, the onset of the Tone 1 was not lowered, thus hyper-articulated.

In our first practical approach [7] to realize the above modeling technique to build tonal models, we used tri-tone context dependent (CD) models to model the hypo-articulation F0 variations, and mono-tone, bi-tone to model the hyper-articulation F0 variations. For the example in Figure 4, 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 the panel (f).

5. Tone Recognition Experimental Results

A series of tone recognition experiments have been carried out on one female speaker data [5, 7, 8], using continuous density tonal HMMs. The standard feature vector has $\log F0$, frame energy and their 1st, 2nd order time derivatives. 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 [**?**]. 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.

Table 6: *Average correct rates for the four basic and the neutral tones in all nine tone recognition experiments.*

Tonal HMMs	Recognition correct rates (%)		
	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 7: *Correct rates for the four basic lexical tones in four representative 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 6 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 7 give detailed performances for the four basic lexical tones in four representative experiments. The whole framework proposed 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.

6. Conclusion

This paper presents out *Tone Nucleus* based 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.

7. References

[1] Fujisaki H., 1997, Prosody, Models, and Spontaneous Speech, In *Computing Prosody: computational models for processing spontaneous speech*, Sagisaka Y. and et al (ed.). New York: Springer, 27-42.

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

[3] Xu Y., 1999, Effects of tone and focus on the formation and alignment of F0 contours, *Journal of Phonetics*, 27(1), 55-105.

[4] Xu Y., 1994, Production and perception of coarticulated tones, J.A.S.A. (4), 1994, 2240-2253.

[5] Hirose K., Zhang J.-S., 1999, Tone recognition of Chinese continuous speech using tone critical segments, In *Proc. of Eurospeech*, Budapest, Hungary, pp.879-882.

[6] Zhang J.-S.,Hirose K., 2004, Tone nucleus modeling for Chinese lexical tone recognition, Speech Communication, Paper in press.

[7] Zhang J.-S., Kawanami H., 1999, Modeling carry-over and anticipation effects for Chinese tone recognition, In*Proc. of Eurospeech*, Budapest, Hungary, 747-750.

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

[9] Zhang J.-S., Nakamura S., Hirose K., 2000, "Discriminating Chinese lexical tones by anchoring F0 features", In*Proc. of ICSLP*, Beijing, Vol. II, 87-90.

[10] Tsumura T. and et al, 1973, Auditory detection of frequency transition, J.A.S.A. Vol. 53, No. 1, 17-25.

[11] Shigeno S., Fujisaki H., 1979, Effect of a preceding anchor upon the categorical judgment of speech and non-speech stimuli, *Japanese Psychological Research*, Vol. 21, No. 4, 165-173.

[12] Zhang J.-S., Nakamura S., Hirose K., 2004, Tonal contextual F0 variations and anchoring based discrimination, *Speech Prosody*, Nara, Japan.

[13] Lindblom B., 1990, Explaining phonetic variation: a sketch of the H&H theory, In *Speech Production and Speech Modeling*, Kluwer Academic Publishers, 403-439.