A SYNTHESIS-ORIENTED MODEL OF PHRASAL PITCH MOVEMENTS IN STANDARD CHINESE

Jinfu Ni, Goh Kawai and Keikichi Hirose

Department of Information and Communication Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
njf@gavo.t.u-tokyo.ac.jp , goh@kawai.com , hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper proposes a computable, tone-driven method to model phrasal pitch movements in standard Chinese by (1) formulating physical constraints for phonetic control mechanisms, (2) defining four phrasal tones quantified by model parameters for generating phrasal tunes, and (3) forming phrasal pitch movements by mapping lexical tones onto phrasal tunes. Several experiments confirm the method's validity. The proposed method is an effective component technology for analyzing, synthesizing and understanding spoken Chinese intonation.

1. INTRODUCTION

The problem addressed in the paper is modeling pitch movements of standard Chinese phrases. Pitch movements are observed as tonal entities, which are the actual values of a complex combination of lexical tones, stress and intonation. Chinese is a tone language with four lexical tones: 1st, 2nd, 3rd and 4th tones having high-level, high-rising, low-digging and high-falling contours, plus a neutral tone 0. Grasping the interaction among lexical tones, stress and intonation is crucial for constructing intonation models suited for high-quality text-to-speech conversion systems. A number of proposals [1-6] on analysis or synthesis of intonation contours shed light on the processes involved with Chinese intonation.

A superpositional model [7] was proposed to simulate the generation process of F0 contours of Chinese sentences based on a Japanese F0 contour generation model [1], where the F0 contour was analyzed as a combination of tone and phrase components. Although the model approximates some F0 contours well [7] [8], not all contours can be decomposed into tone and phrase components. A major factor in tonal languages is that the large undulations of lexical tones obscure the phrase components. Assuming that lexical, phrasal and boundary tones determine the intonation of the utterance suggested in [9], this paper proposes an alternative method to model phrasal pitch movements. In the remainder of this paper, section 2 outlines the method, section 3 presents three experiments and analyses, and section 4 discusses how phrasal pitch movements are formed when tones are given.

2. OUTLINE OF THE METHOD

Pitch is perceived as speech melody correlating with changes in the fundamental frequency (F0) of the speech signal reflecting changes in the rate of vocal cord vibrations. The intonation contour of an utterance is regarded as an ensemble of tonal entities over time on the logarithmic scale.

2.1. Phonetic control mechanism formulation

In earlier work [10], two formulations were proposed to originally simulate the control mechanisms for implementing of Chinese lexical tones:

$$\lambda_e(t) = \lambda_0 + \Delta\lambda(1 - (1 - \gamma t)e^{-\gamma t}),\ t \geq 0 \quad (1)$$

$$\lambda_n(t) = \lambda_0 + \Delta\lambda(1 + \sigma t)e^{-\sigma t},\ \ t \geq 0, \quad (2)$$

$\lambda(t)$ is the response of a critically-damped second-order linear system to certain excitations. $\lambda_e(t)$ results from an external input as the excitation at $t = 0$, and $\lambda_n(t)$ results from the system's pre-stored initial value. γ and σ are the natural angular frequencies of the corresponding systems. λ_0 and $\Delta\lambda$ are newly introduced to formulate the initial and final states of the system responses.

It is known that the implementation of a tone is constrained by the vocal-cord system. Approximating this physical system as a second-order system allows us to express the system's frequency responses to certain external excitations as:

$$f(t) = \rho * \{\frac{1}{\sqrt{(1-\eta_0\lambda(t))^2 + 4\zeta^2\eta_0\lambda(t)}} - 1.0\},\ t \geq 0 \quad (3)$$

where ρ is a coefficient. ζ is the damping ratio of the system, and the coefficient η_0 depends on ζ. In its original formulation, $\lambda(t)$ is the ratio of the external excited frequency to the natural frequency of the system over time. Here $\lambda(t)$ is assumed as quantified tonal excitation over time. $\lambda_e(t)$ simulates active tonal excitation, where γ is in inverse proportion to the excitation period [8]. $\lambda_n(t)$ simulates natural decay with time constant σ. Eqs. (1) through (3) describe the control mechanism for generating tonal entities.

2.2. Modeling phrasal tunes

Chinese intonation contours are determined by a sequence of lexical, phrasal and boundary tones. Phrasal and boundary tones form *phrasal tunes.* Lexical tones are mapped on their phrasal tunes. As there are no common definitions of Chinese phrasal tones, we define phrasal tones within the context of synthesis-oriented modeling of phrasal pitch movements.

Inspired by previous work in other languages (e.g., [9] [1] [3] [4]) and based on our analysis of over 700 read-speech utterances taken from a Chinese speech database, we propose four phrasal tones (two point tones and two contour tones) to transcribe Chinese phrasal tunes. These phrasal tones are quantified by *target values* based on Eqs.(1) through (3). Table 1 defines the four phrasal tones and shows their types and symbols. A target value specified by λ_0 is assigned to a point tone L or H. The target value places the tone on the speaker's pitch scale. The contour tones H-L and H+L are defined by two target values specified by λ_0 and $\lambda_0 + \Delta\lambda$. λ_0 determines the start-point of the contour, and $\Delta\lambda$ shows the relative location of the end-point. H+L and H-L differ in that H+L may be associated with downstep triggered by the 3rd or 4th lexical tones under certain circumstances.

Phrasal tone name	Type	Symbol	λ_0	$\Delta\lambda$
Low	Point tone	L	$0.0 \sim 0.6$	n/a
High	Point tone	H	$0.8 \sim 1.6$	n/a
Phrase-medial fall	Contour tone	H-L	$1.0 \sim 1.6$	$0.4 \sim 1.0$
Phrase-final fall	Contour tone	H+L	$1.0 \sim 1.6$	$0.5 \sim \infty$

Table 1. Definition of the phrasal tones categorized by certain target values based on Eqs.(1) to (3). Target values are specified by the parameters λ_0 and $\Delta\lambda$.

Two boundary tones from [9] are used. The phrase-initial tones %L and %H determine the location of a phrasal tune. The phrase-final tones L% and H% affect target values of adjacent phrasal tones.

In phonetic terms, Eq.(1) represents how actively excited tones are concatenated, as in L to H, H to H, H to H-L, H to H+L, H-L to H-L, and H-L to H+L. Eq.(2) represents the contour tones H-L and H+L, assuming these tones decay with time constant σ starting at an initial target value (determined by λ_0) and falling towards a final target value (determined by $\lambda_0 + \Delta\lambda$). Eq.(1) or (2) are not necessarily associated with particular tone sequences. Actual values of tones and their concatenations are chiefly determined by tonal target values and physical constraints represented by Eq.(3).

The Chinese phrasal tunes we observed consist of the following combination of phrasal and boundary tones:

$$\left\{ \begin{array}{c} \%\mathrm{L} \\ \%\mathrm{H} \end{array} \right\} [\mathrm{L}]\ \mathrm{H}\ (\mathrm{H}, \mathrm{H\text{-}L})^{*} \left\{ \begin{array}{c} \mathrm{H} \\ \mathrm{H{+}L} \end{array} \right\} \left\{ \begin{array}{c} \mathrm{L}\% \\ \mathrm{H}\% \end{array} \right\} \quad (4)$$

where $\{ \begin{array}{c} \mathrm{x} \\ \mathrm{y} \end{array} \}$ means either x or y , $[x]$ means x is optional, $(x, y)^*$ means zero or any number of x and y occurring in any order. The apparent values of a phrasal tune are called *anchor lines* . Anchor lines are positioned relative to the *key* (a value with respect to the register of the speaker). The key is determined by initial boundary tones, which in turn are controlled by higher-level information. Fig.1 illustrates typical anchor lines. In the figure, the segment *a-b* is the transition from the point tone L to the point tone H, segment *b-c* shows the actual values of the contour tone H-L, and segment *d-e-f* shows those of the contour tone H+L. At point *b* , the point tone H and the contour tone H-L overlap. The figure shows the anchor lines (the segments *a-b-c* as long as *d-e-f*) for the phrasal tune %L L H H-L H+L %L. If the phrasal tune were %L L H H+L %L, its anchor line would be the segment *a-b-d-e-f* (the segment *b-c* should is replaced by the segment *b-d*).

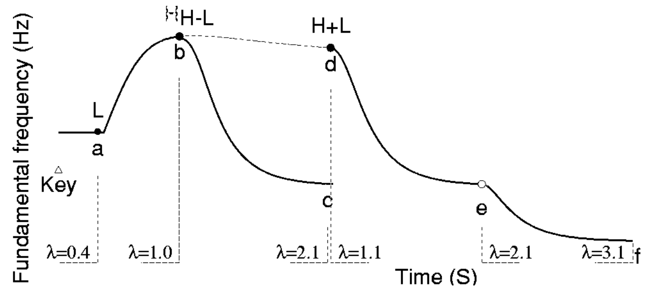

Fig. 1. Example of anchor lines and their underlying phrasal tones. Solid lines are anchor lines. Filled circles indicate underlying phrasal tones. The empty circle is a potential downstep occurring in H+L. The triangle is the key, which determines where the anchor lines is placed with respect to the register of the speaker. λ values for points *a* to *f* are listed at the bottom of the figure. $\zeta = 0.24$, $\rho = 0.85$ and $\sigma = 12.5$ on the logarithmic scale.

Phrasal tunes are determined by a sequence of point, contour tones as long as boundary tones. These tones are realized using several formulations of phonetic control mechanisms. Declination is given by the target values of underlying tones. Baselines or toplines are not needed.

3. EXPERIMENTS AND RESULTS

Experiments were run (a) to see how Eqs.(1) to (4) capture typical phrasal phenomena, and (b) to generate phrasal pitch movements when lexical, phrasal and boundary tones are given. F0 contours of utterances read by of 19 native speakers were quantitatively analyzed on the logarithmic scale with ζ fixed as 0.24 and ρ ranging between 0.8 to 0.9. The utterances were taken from a standard Chinese database (dataset 1), and recorded prompts for an information service system (dataset 2). Following subsections report experiments and results for task (a). Task (b) is discussed in setion 4.

3.1. Experiment 1

Speech samples taken from dataset 1 were 37 context-free, random digit strings recorded by an announcer and a female professional speaker (WL). Results show that context-free phrasal tunes were H tones on the 1st or 2nd syllables plus an H+L on the last or penultimate syllables depending on lexical tones. No downstep occurred clearly.

3.2. Experiment 2

We quantitatively analyzed the F0 contours of 20 utterances from dataset 1 (2 topics, 10 utterances each, recorded in conversational style by WL) and 10 utterances from dataset 2 (recorded by a female announcer FL). In the examples shown in Figs.2 and 3, the symbol "+" denotes the measured F0, and the solid lines are the anchor lines. The values for each phrasal tone and the decay time constant σ are listed at the bottom of each figure. The target values of the start and downstep points of the H+L tones are separately shown. These two examples show that the phrasal pitch movements can be explained by phrasal tunes, and that lexical tones trace or weave around the anchor lines. When lexical tones move off the anchor line, they depart from the anchor line and return to it. The examples show how lexical tones occasionally obscure anchor lines. The more obscured the anchor line is, the stronger the lexical tone's prominence becomes. For instance, the syllables *"da4" "lou2"* " in Fig.2 and content word *"wei1 er3 tong1 xing4"*in Fig.3 are prominent as required by the discourse context. Eqs (1) through (4) predict a wide variety of phrasal phenomena, including tonal entities, declination and downdrift.

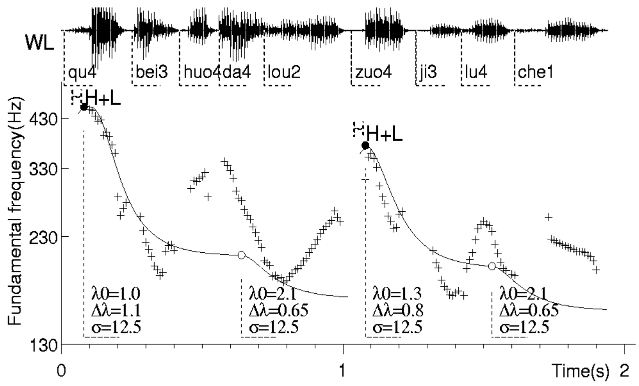

Fig. 2. Example of speaker WL's utterance "dao4 bei3 huo4 da4 lou2, zuo4 ji3 lu4 che1?"/ Which number bus (should I) take to (the) department?/. Both phrasal tunes are transcribed as %H H H+L H%.

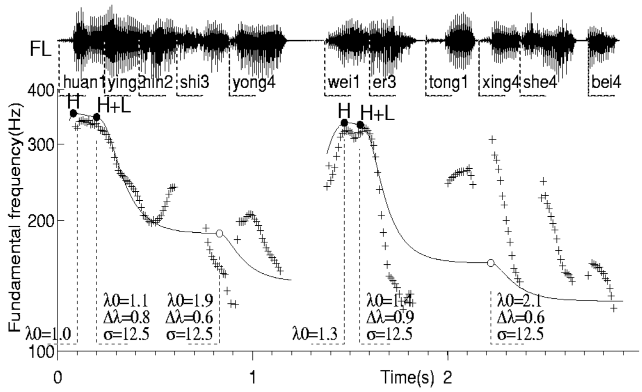

Fig. 3. Example an utterance fragment "nin2 hao3, huan1 ying2 shi3 yong4 wei1 er3 tong1 xing4 she2 bei4."/ Hello, welcome to *WEIER* communication devices./ uttered by FL. Both phrasal tunes are transcribed as %H H H+L H% and %H H H+L L%.

3.3. Experiment 3

The robustness of the proposed method was tested on data from 8 male and 8 female Chinese college students who read the declarative sentence "mei3 guo2 qi1 huo4 shi4 chang3 de0 hui4 yuan2 xi2 wei4, ge4 ren2 dou1 ke3 yi3 hua1 qian2 mei3 dao4."/Anyone can buy in the U.S. commodities exchange./. Experimental results confirm the validity of the method. Fig.5 shows results of (a) a female and (b) a male speaker. Results indicate that speakers tend to choose particular phrasal tunes to convey the speaker's notion of how the sentence should be expressed. Some phrase-final tunes ended with an H tone rather than an H+L tone.

3.4. Summary of the results

The results confirmed the validity of the proposed method. Major findings include:

(1) Eq.(4) predicts most phrasal tunes.

(2) There are clear, fixed relationships between lexical tones and anchor lines.

(3) The decaying time constant σ can be constant for each speaker. In the analyzed data, σ ranged from 9.0 to 12.5, somewhat correlating to speaker and phrasal tune differences.

(4) At most a single downstep occurs in these utterances, usually near the medial period of the syllable whose lexical tone triggers the downstep.

(5) Anchor lines predict phrasal pitch movements and prominence.

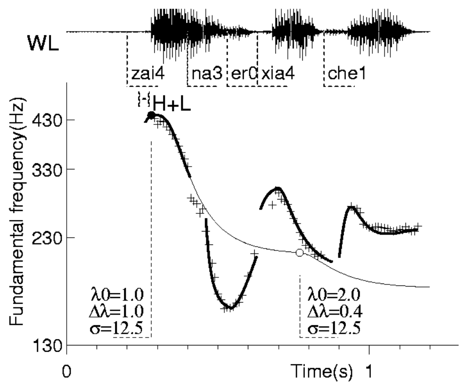

Fig. 4. Example of utterance "zai(4)na(3)er(0)xia(4) che(4)?" /Where (should I) get off?/ illustrating how lexical tones are mapped onto anchor lines and yield phrasal pitch movements.

4. FORMATION OF PHRASAL PITCH MOVEMENTS

Phrasal pitch movements are formulated by the following steps:

step 1: generate anchor lines according to phrasal and boundary tones; and
step 2: map each lexical tone onto the anchor lines according to the lexical tone's context and the sentence's discourse structure.

We experiment by resynthesizing the pitch movements of WL's utterances. Fig.4 shows an example. The thin line is the anchor line of the phrasal tune %H H H+L H%, and the thick lines are the synthesized lexical tone using Eqs. (1) to (3). The first lexical tone shares part of the anchor line. The lexical tones of the 2nd, 4th and 5th syllables depart from the anchor line but remain anchored to the line.

5. CONCLUSION

The paper proposed a synthesis-oriented model for Chinese phrasal pitch movements. Phrasal pitch movements are formed by mapping lexical tones onto phrasal tunes. Phrasal tunes consist of two boundary tones and four phrasal tones. Phrasal tones are determined by formulas that reflect the phonetic control mechanisms of lexical and phrasal tone generation. Experiment results confirm the validity of the proposed method. The next step is to formulate synthesis rules.

REFERENCES

[1] H.Fujisaki and K.Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn(E), Vol.5, No.4, pp.233-242 (1984).

[2] J. Pierrehumbert, "Synthesizing intonation," J. Acoust. Soc. Am.,Vol.70,No.4,pp.985-995 (1981).

[3] N. Thorsen, "Sentence intonation in Danish," Proc. of the XIII International Congress of Linguists, Tokyo, pp.47-56, (1982)

[4] G. Bruce, "Models of intonation: from the Lund horizon," Proc. of ESCA Workshop on intonation: theory, models and applications, pp.11-18, Greece,(1997)

[5] E. Garding, "A comparative study of intonation," Proc. of the XIII International Congress of Linguists, Tokyo, pp.85-94, (1982)

[6] J. 'tHart, R.Collier and A.Cohen, "An perceptual study of intonation: an experimental-phonetic approach to speech melody," Cambridge University Press, 1990.

[7] J. Ni, R. Wang and K. Hirose, "A quantitative model for generating sentence F0 contours of spoken Chinese," Proc. of CJSLP'97, pp.103-110, China, (1997).

[8] J. Ni, R. Wang and K. Hirose, "Quantitative analysis and formulation of tone concatenation in Chinese F0 contours," Proc. of EUROSPEECH97, pp.195-198, Greece, (1997).

[9] M.Beckman and J.Pierrehumbert, "Intonational structure in Japanese and English," Phonology Yearbook 3, pp.255-309 (1986).

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

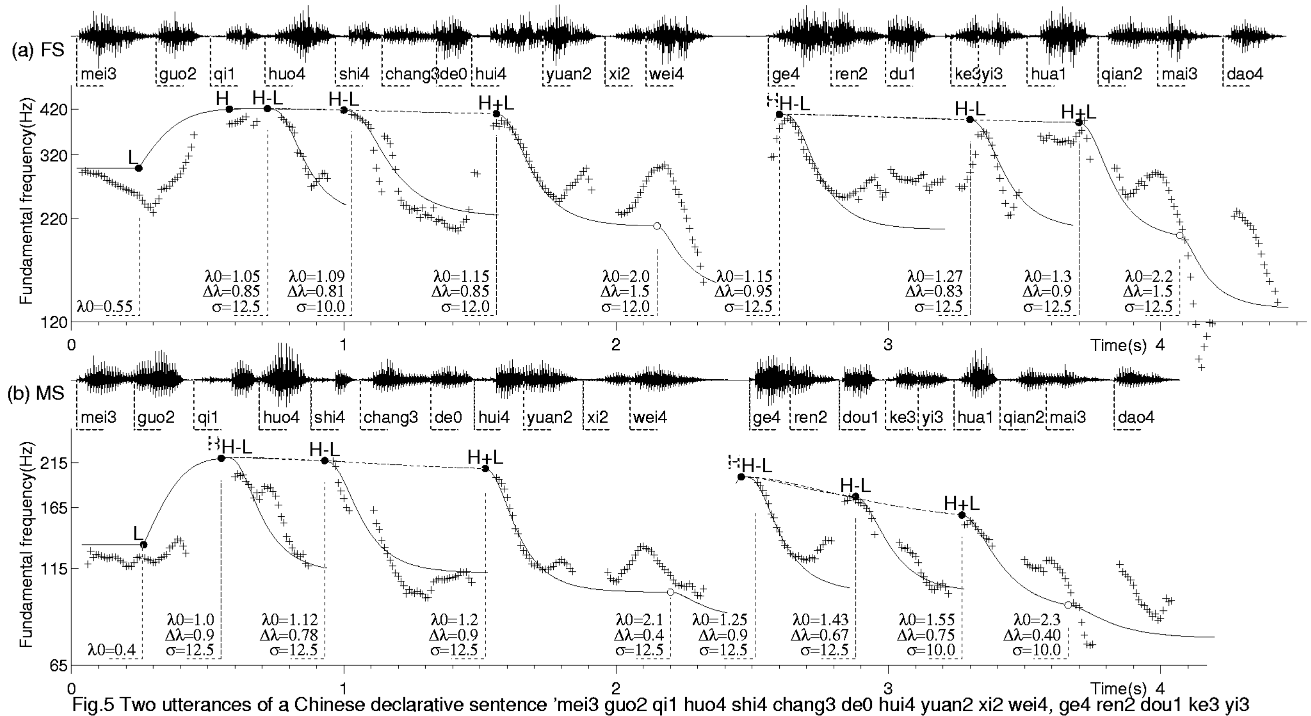

Fig.5 Two utterances of a Chinese declarative sentence 'mei3 guo2 qi1 huo4 shi4 chang3 de0 hui4 yuan2 xi2 wei4, ge4 ren2 dou1 ke3 yi3 hua1 qian2 mei3 dao4.'/Anyone can buy commodities exchange from the U.S./, where (a)for a female and (b)for a male speaker.