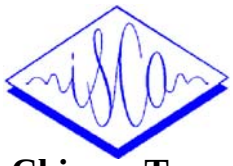

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Pitch Targets Anchor Chinese Tone and Intonation Patterns

Jinfu Ni & Hisashi Kawai

ATR Spoken Language Translation Research Laboratories, Japan
{jinfu.ni; hisashi.kawai}@atr.jp

Abstract

This paper presents a study on the role of pitch targets in the manifestation of Chinese tone and intonation. Pitch targets are particularly measured as F_0 (fundamental frequency) peaks and valleys over time. Analysis and perceptual experiments were conducted on 72 sentences, each with almost identical tone mapping, uttered two times by a female native in statements or questions. The tone and intonation patterns observed from the F_0 contours were quantitatively analyzed using a functional model and then re-synthesized using the model parameters predicted from the pitch targets measured. Two perceptual experiments were done. One rates the similarity between the re-synthesized tone and intonation patterns from the pitch targets and the original; the other tests human perception of tone and intonation when systematically varying the pitch targets of the final tone (Tone 2 and Tone 4) in two statements. Experimental results consistently indicate that the pitch targets are prominent for anchoring Chinese tone and intonation patterns; the exact shape of an F_0 contour is predictable, given the pitch targets.

1. Introduction

Pitch targets basically include *high* and *low*, which are commonly used to describe the intonation of accent languages, like English and Japanese [1]. Chinese is a tone language, and there exist four lexical tones, named Tones 1 to 4, and a neutral tone named Tone 0. If the range of a speaker's voice is divided into four equal intervals, marked by five points, 1 low, 2 half-low, 3 middle, 4 half-high, and 5 high, Tones 1 to 4 are represented by 55, 35, 214, 51, respectively [2]. Because both the actual intervals and the absolute pitch are relative to the individual voice and the mood at the moment of speaking, the term "pitch targets" used in this paper means F_0 (fundamental frequency) peaks and valleys over time. On the other hand, tone and intonation patterns focus particularly on the F_0 contours. The time scope for a tone pattern is limited to syllable size, while that for an intonation pattern may cover more than a syllable, and even the whole utterance.

There are many studies on the tone and intonation in the literature, e.g., [2][3][4][5]. Perception tests and instrumental analyses of the past have yielded a consensus that the F_0 contour of an utterance can multiply manifest lexical tones and intonation. However, no clear answer is available for some basic questions, at least in practice. For instance, do pitch targets suffice for the representation of Chinese tone and intonation? What are the essential features necessary for the synthesis of tone and intonation so as to achieve natural text-to-speech sounds? In this paper, we investigate this issue from the point-view of practice. The remainder of this paper is organized as follows. Section 2 describes the speech material and methodology. Section 3 presents experimental results, and Section 4 contains remarks.

2. Speech material and analysis method

2.1. Speech material

The speech data used here includes 72 Chinese sentences, which are almost all adopted from [3]. These sentences are divided into six groups, each containing 12 base sentences subdivided into three types. Each type includes 4 sentences of the same number of syllables and of the same grammatical structure characterized by the mapping of an identical tone onto the entire sentence, as listed in Table 1, where T1, T2, T3 and T4 indicate Tones 1, 2, 3, and 4, respectively. Type 1 comprises

Table 1: *List of the tone mapping used in the base sentences.*

Type 1	Type 2	Type 3
T1 T1 T1 T1	T1 T1 T1 T1 T1	T1 T1 T1 T1 T1 T1 T1 T1 T1
T3 T2 T2 T2	T2 T2 T2 T2 T2	T3 T2 T2 T2 T2 T2 T2 T2 T2
T3 T3 T3 T3	T3 T3 T3 T3 T3	T3 T3 T3 T3 T3 T3 T3 T3 T3
T4 T4 T4 T4	T4 T4 T4 T4 T4	T4 T4 T4 T4 T4 T4 T4 T4 T4

4 syllables with subject-verb (SV) structures; Type 2 has 5 syllables with subject-verb-object (SVO) structures; and Type 3 combines Types 1 and 2, i.e., Type 2 was added to Type 1 as its sentential object, yielding 9 syllables with SVO structures. These sentences are grouped into the following categories.

S: Types 1, 2, 3 in statements
U: Types 1, 2, 3 in lexically and grammatically unmarked yes -no questions (hereafter *unmarked questions*)
P: Types 1, 2, 3 in yes-no questions with interrogative particle *ma0* in sentence-final positions (*particle questions*)
N0: Type 2 in yes-no questions with *shi4-bu2-shi4* structures
N1: Type 2 in yes-no questions with *X-mei2-X* structures
N2: Type 2 in yes-no questions with *X-le0mei2-X* structures
Q0: Type 2 in alternative questions with *X-hai2shi4-Y* structures
Q1: Type 2 in questions with *shi4-X-hai2shi4-Y* structures
Q2: Type 2 in questions with *hai2shi4-X-hai2shi4-Y* structures
W0: Type 2 in why (*wei4she2me0*) questions
W1: Type 2 in when (*she2me0shi2hou4*) questions
W2: Type 2 in what (*she2me0*) questions

We recorded the 72 sentences two times in a sound-proofed room with a female speaker without expressive emotion.

2.2. A functional model of the F_0 contours

In this paper, we use a functional model [6] to represent the observed F_0 contours in a parametric form. According to the model, the voice register (a frequency register of utterances) of the speaker is first transposed to a so-called RONDO scale (similar to a log-scale). The RONDO-F_0 contour is then expressed in concatenative mountain-shaped patterns lined up in series at

10.21437/SpeechProsody.2004-22

the time axis. The F_0 contour $F_0(t)$ is given as follows.

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, \text{for } t \geq 0, \quad (1)$$

where

$$A(\lambda) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \lambda \geq 1, \quad (2)$$

and

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \text{Min}(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t) \ . \quad (3)$$

$\text{Min}(z_1, z_2)$ means the smaller one of both z_1 and z_2. Equations (1) and (2) jointly indicate the transposition of the voice register. Equation (3) expresses the RONDO-F_0 contour $\Lambda(t)$, where $\Lambda_{r_i}(t)$ and $\Lambda_{f_i}(t)$ indicate the rise and fall components of the ith mountain-shaped pattern, respectively. Particularly,

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t \leq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for t} \geq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

$$\text{where } D_{x_i}(t) = (1 + \frac{4.8t}{\Delta t_{x_i}})\, e^{-\frac{4.8t}{\Delta t_{x_i}}}, \text{for } t \geq 0. \quad (6)$$

Parameters ζ, λ_t and λ_b can be commonly fixed at 0.237, 1 and 2, respectively [6]. There are then two speaker-dependent but utterance-independent parameters in the frequency domain,

$[f_{0_b}, f_{0_t}]$: top and bottom frequencies of the voice register,

and five utterance-dependent but speaker-independent parameters in the RONDO-time space,

n : number of mountain-shaped patterns,
Δt_{x_i} : response time for the ith rise/fall component,
$\Delta\lambda_{x_i}$: amplitude of the ith rise/fall component, $x \in \{r, f\}$
(t_{p_i}, λ_{p_i}) : peak of the ith mountain-shaped pattern, $i = 1, ..., n$.

2.3. Methodology

The 144 observed F_0 contours were first automatically analyzed using the method in [7]. The F_0 peaks and valleys were then manually determined with a visual inspection of the F_0 contours, taking into account the underlying tones. The number of F_0 peaks for a tone was determined according to the tone modeling [7]; an F_0 valley was then determined using the contour between adjacent peaks. With model-generated F_0 contours, we re-synthesized these utterances for perceptual experiments using a tool called STRAIGHT [8]. Three analysis and perceptual experiments were conducted. Experiment 1 analyzes the effectiveness of the re-synthesis of the F_0 contours based on the F_0 peaks and valleys. Experiment 2 investigates the correlation of the interaction between tone and intonation with the change of F_0 peaks and valleys. Experiment 3 shows that the tone and intonation can vary with the change of pitch targets. Based on these experimental results, we then argue that pitch targets anchor the tone and intonation patterns.

3. Results

3.1. Re-synthesis of the tone and intonation patterns

Experiment 1 was conducted to investigate the effectiveness of the re-synthesis of the tone and intonation patterns from the F_0 peaks and valleys. Let (t_{v_i}, λ_{v_i}) denote the valley between the ith and the $i+1$th peaks. With the peaks given, the other model parameters necessary for generation of the F_0 contours are calculated as follows.

$$\Delta t_{r_i} = \max(0.05, t_{p_i} - t_{v_{i-1}}), \quad (7)$$
$$\Delta\lambda_{r_i} = \max(0.02, (\lambda_{v_{i-1}} - \lambda_{p_i}) \times 1.05), \quad (8)$$
$$\Delta t_{f_i} = \max(0.05, t_{v_{i+1}} - t_{p_i}), \quad (9)$$
$$\Delta\lambda_{r_i} = \max(0.02, (\lambda_{v_{i+1}} - \lambda_{p_i}) \times 1.05), \quad (10)$$

Table 2 shows the statistical results of the tone-related samples measured from the speech material, where μ_c and σ_c indicate the mean and variance of these model parameters manually checked (*checked parameters*), respectively; μ_p and σ_p indicate those predicted by the F_0 peaks and valleys (*prediction parameters*). The columns μ_e and σ_e list the mean and variance of the errors between the checked and prediction parameters.

Table 2: *Statistical results of the model parameters.*

	Count	μ_c	σ_c	μ_p	σ_p	μ_e	σ_e
Δt_r	366	0.140	0.047	0.122	0.043	0.022	0.022
$\Delta\lambda_r$	366	0.224	0.147	0.215	0.143	0.013	0.035
Δt_f	382	0.139	0.047	0.134	0.055	0.019	0.027
$\Delta\lambda_f$	382	0.196	0.129	0.188	0.122	0.007	0.015

In order to test the similarity between the re-synthesized tone and intonation patterns and the original, we performed a perceptual experiment with 288 stimulus pairs, including 144 re-synthesized utterances with the checked parameters and 144 utterances with the prediction parameters. The stimuli were presented to two natives through headphones in a silent room. After hearing a pair of stimuli, the listener rated the similarity of the tone and intonation between them with a 3-point scale, 0 (very different), 1 (similar), 2 (no difference). The listeners were allowed to hear the same stimuli several times before making a judgment. The average scores for the checked and prediction parameters were 1.93 and 1.89, respectively, and no "very different" sample occurred. The experimental result indicated that the pitch targets, i.e., the F_0 peaks and valleys over time, suffice to capture the nature of the tone and intonation patterns.

3.2. The interaction between tone and intonation

In Experiment 2, we examined the interaction between tone and intonation by analysis of the F_0 peaks and valleys for each of the 12 categories. The main results are described below.

Firstly, the F_0 contours of the utterances in questions were, more or less, moved upward entirely compared to those in statements. This result agrees with the finding in [3][4]. In the utterances with identical Tone 1 and Tone 4 mapping, their F_0 peaks and valleys were raised to a higher register than those with identical Tones 2 and 3 mapping. Figure 1 shows examples of the F_0 contours for the two sentences "hong2bi2tou2 mei2 quan2" and "guo4lu4ke4 zhao4 xiang4" uttered in statements, which means (a)"The Red Nose does not have power." and (b) "A passerby takes pictures.", respectively, and in unmarked questions, which means (c)" Does not the Red Nose have power?" and (d) "Does a passerby take pictures?", respectively.

Secondly, there exist sentence-final-tone dependent manners to manifest the intonation in both unmarked questions and particle questions. Let us group the final tones into two sets: {Tone 2, Tone 3} and {Tone 1, Tone 4}. In the former, the F_0 peak of the rise portion is raised to a high register, thus considerably widening the tone range. In the latter, however, both F_0 peaks and valleys (if any) are raised to a high register, thus narrowing the tone range and moving up the F_0-valley scale. This phenomenon can be clearly observed from the examples shown in Figure 1 (c) and (d).

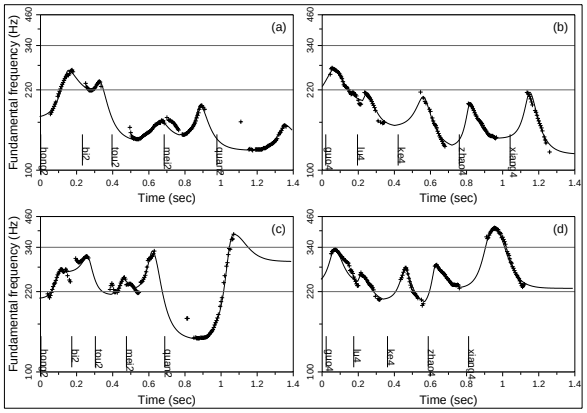

Figure 1: *Examples of F_0 contours of sentences uttered in statements and in unmarked questions. The "+" sequence indicates the observed F_0 contours, and the solid lines indicate the model-approximated contours with the checked parameters.*

Thirdly, there exists a rising-falling pattern in yes-no questions with the *X-not-X* structure, i.e., the categories N0, N1 and N2, and wh-questions, i.e., the categories W0, W1 and W2. The rising-falling pattern is basically manifested by arranging the F_0 peaks and valleys of the tone structure of the function words, like *shi4-bu2-shi4*. Figure 2 shows four examples: (a) "bao1shen1gong1 *shi4-bu2-shi4* ca1 che1?" (Does the indentured laborer clean the car, or not?); (b) "lao3shou3zhang3 *shi4-bu2-shi4* mai3 jiu3?" (Does the old senior officer buy sake, or not?); (c)"bao1shen1gong1 ca1-*mei2*-ca1 che1?"(Has the indentured laborer cleaned the car, or not?); and (d) "lao3shou3zhang3 mai3-*mei2*-mai3 jiu3?"(Has the old senior officer bought sake, or not?). It is clear from this figure that tones involved in function words were adjusted to fit the rising-falling pattern. If a tone was in conflict with it, the tone lost its basic shape and followed the trajectory of the rising-falling pattern, for instance, syllables ca1 and mei2 in Figure 2 (c).

Fourthly, a *transition pattern* was used for the function words *shi4* and *haishi4* in alternative questions (categories Q0, Q1 and Q2). The term *transition pattern* means that the tones in the function words take a rather narrow F_0 range and are located

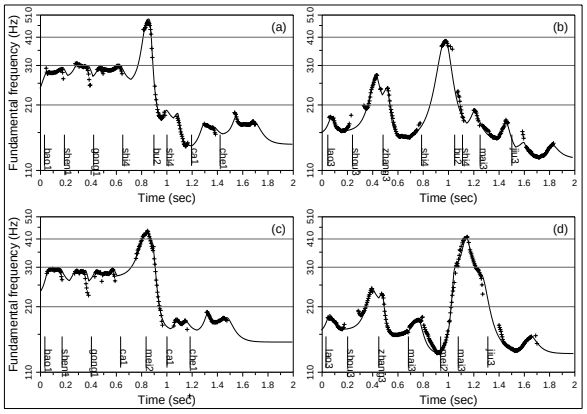

Figure 2: *Examples of F_0 contours of utterances in yes-no questions with the X-not-X structure.*

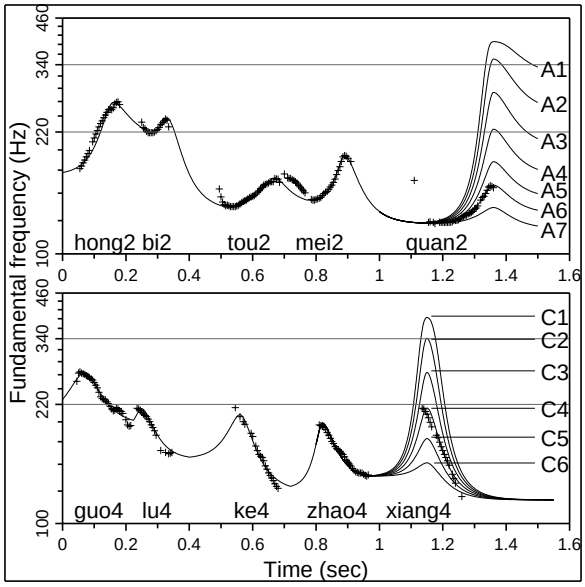

Figure 3: *The model-generated F_0 contours while systematically varying the F_0 peaks of the final tones.*

at a mid register. Instead, the focus phenomenon [5] is usually observable at the phrases around the function words. These observations clearly show that the interaction between tone and intonation can be well captured by the pitch targets.

3.3. Perception of artificial tone and intonation patterns

Experiment 3 investigated the perception of tone and intonation while systematically varying the pitch targets. We used the two utterances in statements shown in Figure 1 (a) and (b) as carrier utterances and changed the F_0 peaks and valleys of the final tones in two/three ways. The first way was to change the F_0 peaks (particularly, the model parameter λ_{p_i}) around the original with step size 0.1, but fix the F_0 valleys unchanged. Figure 3 displays the observed F_0 contours ("+" sequence) and these model-generated contours (solid lines, called *artificial tone and intonation patterns*). The second way was to change both the F_0 peaks and valleys by simply moving the tone up or down with the same step size 0.1. The model-generated F_0 contours are shown in Figure 4 marked by symbols B1 to B6 and D1 to D6. The third way, particularly for Tone 4, was to fix the F_0 peak but move the F_0 valley upward with step size 0.1 so as to raise its valley. The model-generated contours are shown in Figure 4 marked by E1 to E3. The two carrier utterances were re-synthesized with all of the model-generated F_0 contours.

We performed a perceptual test of these utterances with three natives. The stimuli were presented to a listener two times in a random order through headphones in a silent room. After hearing a stimulus, the listener answered three questions.
(1) Is the utterance a statement or a question?
(2) Is the last syllable emphasized, normal, or neutralized?
(3) Which tone did you hear at the last syllable?
The experimental results are summarized in Table 3, where "Que" and "Sta" indicate "question" and "statement", respectively; "Emp", "Nor", and "Wea" indicate "emphasis", "normal" and "weak stress", respectively.

Three observations can be made from this experiment.

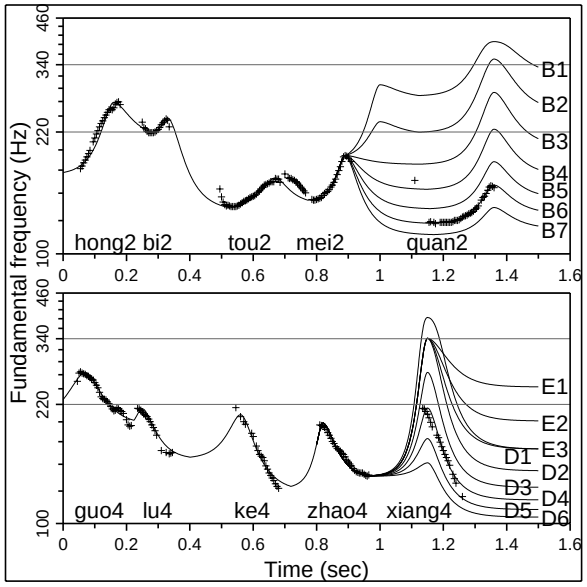

Figure 4: *The model-generated F_0 contours while systematically varying both F_0 peaks and valleys of the final tones.*

Firstly, F_0 contours can multiply manifest tone, stress and intonation [2][3]. When raising the F_0 peaks, the syllables consistently were perceived to be emphasized; when lowering the F_0 peaks, the syllables were perceived with weak stress. Secondly, it proves the final-tone dependent manifestation of question intonation. In the case of Tone 2, the higher the F_0 peak, the easier the utterance being judged as a question. In the case of Tone 4, only when the tone is located at a high register, the utterance may be perceived as a question. However, when the tone valley is located at a low register, all the listeners perceived the utterances as statements. The experimental result also indicated that the features of the final tone are not sufficient for discriminating questions from statements; the other features, like tempo, also give perceptual cues. Finally, the tone is determined by the F_0 peaks, valleys and their alignment with the syllable [2]. Tone 1 and Tone 3 were perceived under certain conditions in this experiment. In addition, the result that the F_0 contours marked by B1, B2 and E1 were perceived as Tone 1 provides an explanation of the phenomenon that Tone 1 can show a rise contour in a high register to meet the need of the manifestation of intonation, as shown in Figure 2 (c), but not miss its perception.

4. Remarks

Several analysis and perceptual experiments were conducted on a well-designed speech material for studying Chinese tone and intonation patterns. Experimental results indicated that the pitch targets play a prominent role in anchoring the tone and intonation patterns; the exact F_0 contours can be predicted from the F_0 peaks and valleys, for example, using the functional model. Based on this result, we assume that an observed F_0 contour can be skeletonized as a sequence of F_0 peaks and valleys without losing the primary linguistic and para-linguistic information it conveyed. This shall promote the application of prosodic information to speech information processing.

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

Table 3: *Result for the perception of the artificial intonation.*

Factor	Que(%)	Sta(%)	Emp(%)	Nor(%)	Wea(%)	T1(%)	T2(%)	T3(%)	T4(%)
A1	66.7	33.3	100				100		
A2	66.7	33.3	100				100		
A3	66.7	33.3	66.7	33.3			100		
A4	33.3	66.7	66.7	33.3			100		
A5	16.7	83.3	66.7	33.3			100		
A6		100	33.3	66.7			100		
A7		100		66.7	33.3		66.7	33.3	
B1	100		100			100			
B2	100		83.3	16.7		50	50		
B3	83.3	16.7	100				100		
B4	16.7	83.3	100				100		
B5		100	83.3	16.7			100		
B6		100	16.7	83.3			100		
B7		100	16.7	50	33.3		100		
C1		100	100						100
C2		100	100						100
C3		100	66.7	33.3					100
C4		100	33.3	66.7					100
C5		100	16.7	50	33.3				100
C6		100		66.7	33.3			33.3	66.7
D1	66.7	33.3	100						100
D2		100	100						100
D3		100	66.7	33.3					100
D4		100	33.3	66.7					100
D5		100	16.7	66.7	16.7				100
D6		100		50	50			16.6	83.3
E1	66.7	33.3	83.3	16.7		50			50
E2	66.7	33.3	100						100
E3	66.7	33.3	100						100

5. References

[1] Beckman, M. E. and Pierrehumbert, J. B. ,1986. Intonational Structure in Japanese and English, *Phonology Yearbook 3*, 255-309.

[2] Chao, Y. R., 1968. A Grammar of Spoken Chinese. Berkeley, CA. University of California Press.

[3] Shen, X. S., 1990. The Prosody of Mandarin Chinese. University of California Publications.

[4] Shen, J., 1994. Hàny*ǔ* y*ǔ*dià oh*é* y*ǔ*dià olèix*í*ng. *Zhōngguó yǔwén*, 3, 221-228.

[5] Xu, Y., 1999. Effects of Tone and Focus on the Formation and Alignment of F_0 Contours. *Journal of Phonetics*, 27, 55-105.

[6] Ni, J. and Hirose, K., 2000. Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences. *ISCSLP2000*. Beijing, 319-322.

[7] Ni, J. and Kawai, H., 2003. Tone Feature Extraction through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching. *ICASSP2003*. Vol. 1, 72-75.

[8] Kawahara, H., Ikuyo, M. K., Cheneigne, A., 1999. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. *Speech Communication*, 27, 187-207.