Quantitative Analysis and Formulation of Tone Concatenation in Chinese F0 Contours

Jin-Fu Ni, Ren-Hua Wang*, Keikichi Hirose***
*Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei,P.R.China, 230027
**Department of Information and Communication Engineering,
School of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, 113, Japan.
jfn@eeis.ustc.edu.cn rhw@ustc.edu.cn hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

With the aim of constructing a set of prosodic rules enabling to generate high-quality synthetic speech of Chinese, tone concatenation features were investigated for Chinese words. Using a superpositional model developed for Chinese F0 contours, quantitative analyses were conducted on 124 Chinese multi-syllable words to find out features on their F0 contours, especially the ones related to tone concatenation. A set of rules were then introduced for the control of model parameters to generate F0 contours of connected tones using the model. Comparison between F0 contours of natural utterances and those rule-generated for 340 words with various tone combinations showed the validity of the proposed rules.

1. INTRODUCTION

Prosody of spoken Chinese has two major factors, i.e. tone and intonation, both of which are manifested by the contour of voice fundamental frequency (henceforth F0 contour). Chinese is a typical tone language, where four lexical tones exist for a syllable: namely, tone 1 characterized by a high-flat F0 contour, tone 2 characterized by a rising contour, tone 3 characterized by a low-dip contour, and tone 4 characterized by a falling contour from high F0. It is well known that certain Chinese syllables are distinguished from others only by the direction and the range of F0 change (henceforth, called this kind of tonal feature as tone shape), and difference in the range is also known to be one of the most dominant cues in perceiving prominence of a word. Although, in connected speech, tone shapes still maintain their original shapes of isolated utterances in some extent, they suffer from large deformation due to tone concatenation, known as tone sandhi. Therefore, it is of great importance for approaching high-quality in Chinese synthetic speech to clarify the relationship between features appearing on F0 contours due to tone concatenation and their underlying linguistic factors. Several efforts have already been spent for this purpose[1,2], and qualitative analyses have yielded some results, especially some important Chinese tone sandhi rules[1]. However, since they were mostly based on rude models on F0 contours, the existed quantitative formulation and realization method of tone concatenation were still far from meeting the need of high-quality speech synthesis. Different from these work, in the current research, a quantitative model for the generation of Chinese F0 contours was used for the analysis and synthesis of F0 contours. This model is based on the superpositional model (known as Fujisaki model [3]), and is modified to fit to the Chinese cases by the authors[4]. Consequently, more complex tone concatenation phenomena can be explained and realized with simpler rules as compared to the previous work [1, 2].

2. MODEL FOR CHINESE F0 CONTOURS

The F0 contour of an utterance generally shows local humps superposed on a smoothly decaying baseline[3]. These humps may differ in their height and also vary in their shape at rising or/and falling portions. When a long pause exists in an utterance, the decaying baseline will be interrupted before the pause and will resume after. In the framework of the quantitative model[2], a logarithmic F0 contour can be represented as the sum of tone components and phrase components; tone components assumed as responses of a second-order system modified from the original one[3] to stepwise tone commands[5], and phrase components assumed as those of an original second-order linear system to impulse-like phrase commands. An explanation for these features of the model is illustrated in Fig. 1. The mathematical expression for the model F0 contour of an utterance is given by,

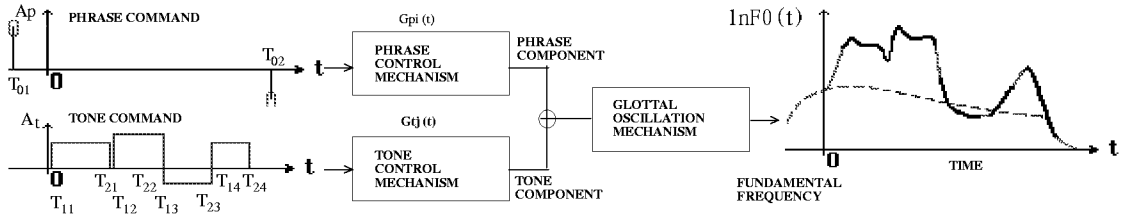

Fig.1 Schematic illustration of the quantitative model for the generation process of Chinese F0 contours.

$$\ln F0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{tj} G_{tj}(t - T_{1j}, T_{2j} - T_{1j}), \quad (1)$$

where $G_{pi}(t) = \alpha_i^2 \, t \, e^{-\alpha_i t}, \quad t >= 0,$ (2)

and $Gtj(t, D_j) = \frac{1}{1 - 2\zeta\delta\sqrt{1-\zeta^2}} * \left(\frac{1}{\sqrt{(1-\eta\lambda_j(t))^2 + 4\zeta^2\eta\lambda_j(t)}} - \delta \right)$, $\lambda_j(t) = \begin{cases} 1 - (1 - \gamma_j t) e^{-\gamma_j t} & D_j > t >= 0, D_j >= \gamma_j^{-1}; \\ (1 + \sigma_j (t - D_j)) e^{-\sigma_j (t - D_j)}, & t >= D_j, \end{cases}$ (3)

respectively indicate the impulse response function for phrase control mechanism and the stepwise response function for tone control mechanism. In our model, an extended version of the original superpositional model[3], a second-order system Gtj(t) is adopted for the tone control mechanism instead of a second-order linear system Gaj(t) adopted in the accent control mechanism in the original model. Although Chinese sentence F0 contours can be well represented in the framework of the original model[6], the above extension was conducted to obtain a better and effective modeling to the observed F0 contours. Output of the system Gtj(t) for a stepwise tone command shows one of rise-fall patterns indicated in Fig. 2 depending on the distance between the onset T11 and the offset T21 of the command. If the distance coincides with γ^{-1}, the output takes left-hand side pattern with a single peak (P type). On the other hand, if it is larger than γ^{-1}, the output takes the right-hand side pattern with a plateau-like shape (F type).

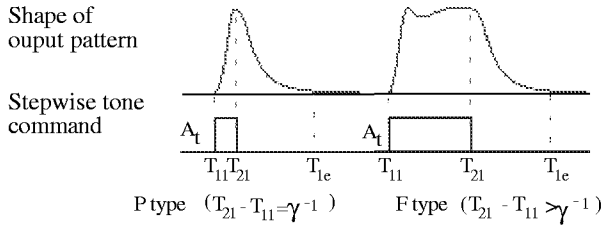

Fig.2 Illustration of two types of rise-fall patterns generated by tone control mechanism and corresponding stepwise tone commands. Shapes of the rising and falling parts of F0 contour are mainly dominated by parameters γ and σ, respectively, and end timing of the falling part is defined by T1e.

3. SPEECH SAMPLES AND METHOD OF ANALYSIS

Speech samples used for the current research include 124 Chinese (20 di-, 32 tri-, 72 quad-syllable) words as analysis material and 340 (20 di-, 64 tri-, and 256 quad-syllable) words as test material. These 340 words of test material are different from 124 words selected for the analysis and they cover all kinds of Chinese tone combinations. Both materials were uttered by a female announcer twice in standard Chinese, with neutral manner and at a normal speech rate. The speech samples recorded on a DAT were down sampled to 8kHz (16 bit), and their F0 contours were extracted by a method based on a short-term autocorrelation analysis of LPC residual and peak detection. After correcting extraction errors manually, model fitting was conducted by the scheme of Analysis-by-Synthesis. Initial values of model parameters necessary to start the AbS process were decided manually using a graphic editing tool. As for the phrase command parameters (onset and magnitude) and Fmin, they were decided so that the model-generated phrase component can well represent the assumed baseline of the observed contour. As for the tone command parameters, the onset and offset were roughly set to the F0 valley and the initial point of F0 fall, respectively. The values of model parameters are then optimized by an iterative process for minimizing the mean square error between measured and model-generated F0 contours. During the iterative process, α, η, ζ and δ were fixed to constant values, namely, α=3.0, η=0.96, ζ=0.1, and δ = 1.0 (for rising and level portions of tone components) and δ=1.1 (for falling portions of tone components).

4. ANALYSIS AND FORMULATION OF TONE CONCATENATION FEATURE

Analysis results indicated that a rather small set of parameters might be enough to represent the features of Chinese F0 contours. Moreover, they imply that we can fix the phrase component to that with Ap=0.4 for all the samples for analysis, this means that acoustic features related to tone concatenation can be attributed to tone components. Figure 3 shows the result of analysis for two examples with their waveforms at the top of panels. Symbols '+' indicate the measured F0 (displayed in every two points for the sake of visibility), while the solid line and the dashed line indicate the model-generated F0 contour and its phrase components, respectively. Also, the underlying tone and phrase commands are also illustrated at the bottom of panels. From the figure, it can be seen that the F type or P type pattern may be shared by several tones and a tone shape may cross both adjoining tone components. In the following sub-sections, only results related to tone components will be discussed.

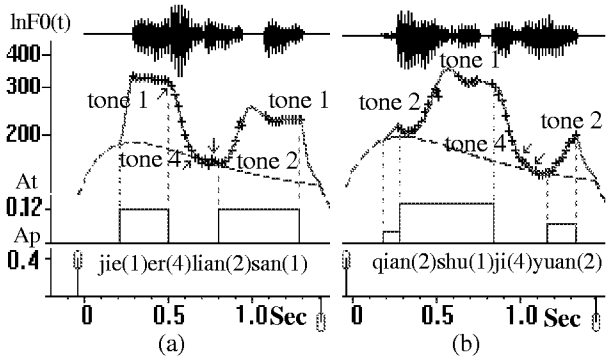

Fig. 3 Two examples of F0 contour analysis of Chinese words. Each F0 contour is well decomposed into phrase components (Ap=0.4) and tone components related to tone features closely. Where these arrows help to show initial or end point of a tone.

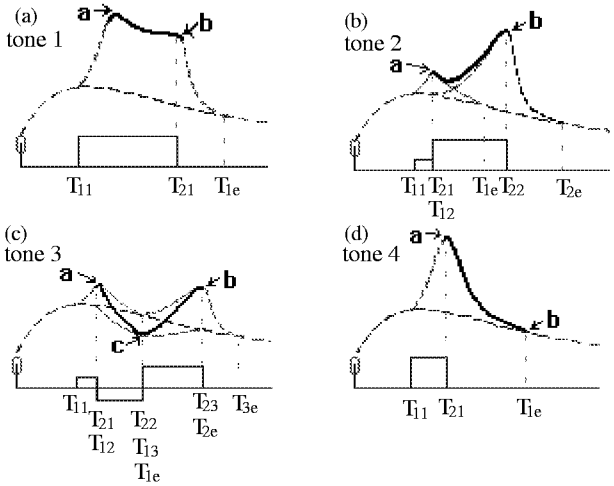

Fig. 4 Tone commands (bottom of each panel) and resulting F0 contours for four basic tones of Chinese. Each panel schematically illustrates how a tone shape (indicated by the thick line) is generated from tone commands. Beginning and end of a tone shape marked by "a" and "b" will be connected with beginning and end of a syllabic final vocalic part (*Yunmu*), respectively, while symbol "c" indicates the bottom of tone 3, which can roughly be defined at halfway of *Yunmu* segment.

4.1. Representing tone shapes by tone components

Results of the current analysis together with those formerly obtained imply that Chinese tones can well be represented only by the P type and F type patterns shown in Fig. 2. Figure 4 schematically shows how four basic tones can be represented using these patterns: tone 1 by an F-type pattern, tone 4 by a P type pattern, and tone 2 and 3 by concatenated P type patterns. Concretely, as shown at the bottom of each panel, tones 1 and 4 are generated by a single stepwise command, while two and three commands are necessary for tone 2 and tone 3, respectively. In the case of tone 3, the middle command should have negative amplitude. Several temporal constrains are applied for the command onsets/ offsets: T12=T21 located at beginning of corresponding tone shape (marked by "a") for tones 2, 3, T1e=T22=T13 located at the bottom (marked by "c") and T2e=T23 located at the end of corresponding tone shape (marked by "b") for tone 3. Roughly speaking, for a tone F0 contour, its shape is decided by the parameters γ and σ, while its height is determined by the amplitude(s) of tone command(s).

4.2. Parameters γ and σ

As mentioned already, rising and falling portions of a tone component are characterized by the model parameters γ and σ. Here, we try to relate them with duration of rising portion of F0 contour and that of falling portion. As for γ, the following equation comes from equation (3):

$$\gamma = 1/D_r \qquad (4)$$

where D_r is defined as duration of rising portion of a tone component. On the other hand, as for σ, the following equation was developed based on the results of F0 contour analysis shown in Fig.5.

$$\sigma = \mathrm{Min}[40.0, 0.6/((D_f - 0.09) + 6.0] \qquad (5)$$

where Df is defined as the duration of falling portion of a tone component, i.e. Df =T1e-T21. These results imply that a tone shape of syllable can be determined roughly only by the duration of syllabic *Yunnu*, which mostly covers the voiced part of the syllable.

Fig.5 Relationship between σ and F0 falling duration Df. It can be approximated by a hyperbolic function indicated by solid (and dotted) line. The symbol notified as "T4Sx" indicates the relationship for the observed portion of tone 4 in xth syllable position. The symbol notified as "others" indicates that for tone 3 falling portion and other observable falling portions which follow tone 1 or tone 2 in any syllable position.

4.3. Tone command amplitudes

Height of a tone component is mostly determined by the tone command amplitude At, as mentioned already. Figure 6 shows the distributions of tone command amplitudes for the speech material of current analysis. The results are shown separately depending on the tone types (marked by "Tx," where x standing for tone 1, 2, 3 or 4) and the position of the syllable (indicated by "Sy," where y standing for the 1st, 2nd, 3rd or the last position "e"). The results indicate that, although all the words were uttered in neutral manner (without intended prominence by the speaker), there exists a large variation in At of each group. When a syllable locates in a word other than at the final position, the tone command amplitude seems not to depend on the syllable position. When a syllable locates at the final position, the At decreases sharply as also shown in other languages. It seems that the variation is mainly caused by the tone coarticulation especially when tone 2 and tone 4 lie in the middle position of a phrase component. Preliminary rules for At will be presented in next subsection, further analysis is on the way to clarify the point for summarizing the rules in detail.

Fig. 6 Distribution of tone command amplitudes.

4.4. Realizing tone concatenation

Generally speaking, two phenomena can be observable in F0 contours of connected speech. One is the tone concatenation. F0 contours of adjoining syllables join together to form a smooth F0 contour. For instance, the rising and falling portions of F type tone component (for tone 1) in Fig.3(b) are shared by tone 2 and tone 4, respectively. The other one is that the height of a tone component is influenced by the heights of its neighboring syllables' and also by the focal condition. Taking these into consideration, the following algorithm was developed to generate tone commands automatically for an input word when tone types and key points (marked by "a", "b" and "c" in Fig. 4) of constituent syllables are given:

(1) Initiation: Generate tone command for each syllable by the rules shown in Fig.4. Then, calculate γ and σ according to Eq.(4) and Eq.(5). For other model parameters, use default values. For instance, default values for command amplitudes are as follows: for tone 1, At=0.115; for tone 2, At=0.085 in the last syllable position, otherwise At=0.105; for tone 3, At=0.035 and At=-0.035; for tone 4, At=0.135 in the first syllable position, At=0.105 in the last position, and At=0.125 for other cases. In addition, At=0.03 for the small positive command of tones 2 and 3.

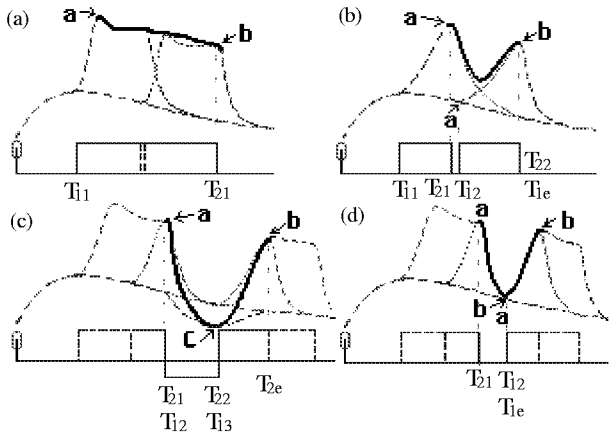

Fig.7 Rules for tone command merging due to tone concatenation. In each panel, the thick line indicates merging portion, dished lines show original P type or F type patterns. Symbols "a", "b", and "c" indicate the key points of syllables.

(2) Concatenation: As shown in Fig.7, tone commands are merged into one new command in several cases. Panels (a) and (c) respectively indicate the case of two tone 1 syllables, and the case of a tone 1, 2 or 4 syllable followed by a tone 3 syllable which further followed by a tone 1, 2 or 4 syllable. Both panels (b) and (d) show the typical cases of a tone 4 syllable followed by a tone 2 or tone 4 syllable, difference in both resultant patterns mainly results from whether T1e=T22 or T1e=T12. Of course, the dished lines on the left side of symbol "a" and right side of symbol "b" may also act as a tone contour. Temporal relationships among onsets and offsets of tone commands, end timing of a tone component T1e or T2e, and key points of tone shapes are indicated in the figure. These concatenation rules are repeatedly applied to longer words, and total number of resulting stepwise tone commands is minimum.

(3) Calculation of merged command parameters: T11, T21, γ and σ of each merged command are taken from those of original commands. In several cases, γ and σ need to be recalculated due to temporal constraints shown in Fig. 7. As for At, At of one of the original commands (before merging) is adopted according to the following priority: tone 4, tone 1, tone 2,or tone 3.

5. EXPERIMENTS OF TONE CONTOUR GENERATION USING DEVELOPED RULES

Validity of the developed rules is tested by comparing the model-generated F0 contours with the measured ones for the test material consisting of 340 Chinese words. The rules require the Key point timings, which were given manually for the current experiments. Phrase command parameters and Fmin are adopted from the result of the AbS analysis of the natural utterance. Results indicated that the F0 contours synthesized by the rules can approximate the measured F0 contours very well. Figure 4 shows four examples, where "+" symbols stand for the measured F0's, solid lines indicate model-generated F0 contours by rule. Therefore, the proposed rules are effective for generating F0 contour of connected tones without intended prominence.

6. CONCLUSION

As the first step to realize naturally sounding prosody in synthetic continuous speech, a number of F0 contours were analyzed for multi-syllable words and rules were constructed for F0 contour generation. Acoustic features related to tone concatenation were first investigated using a quantitative model of F0 contour generation. Then based on the results, the rules were constructed for the generation of F0 contours at the default case (without apparent prominence). Validity of the developed rules was indicated by the experiments on F0 contour generation for the test samples (open condition experiments).

This work was partly supported by National Natural Science Foundation of China.

REFERENCES

[1]Z.J. Wu, "The Basic Tone-sandhi patterns in Standard Chinese Intonation," Essays on Linguistics, pp 54-73 (1989)

[2]L. S. Lee, C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System", IEEE on Speech & Audio,Vol.1,No.3, 1993.

[3]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.

[4]J. F. Ni, R.H. Wang, K. Hirose and D.Y. Xia "A Quantitative Model for Generating Sentence F0 Contours of Spoken Chinese", Proc.of CJSLP'97, pp 41-49, 1997.

[5]J. F. Ni and R.H. Wang, "Modeling the Control Mechanism for Generating the Rise-Fall Pattern in F0 Contours", ACTA ACOUSTIC, vol 21,No.6, pp863-871, 1996. (In Chinese).

[6]K.Hirose, H.T.Lei & H.Fujisaki,"Analysis and Formulation of Prosodic Features of Speech in Standard Chinese Based on a Model of Generating Fundamental Frequency Contours", J. Acoust. Soc. Jpn, vol.50, no.3, pp.177-187, 1994.

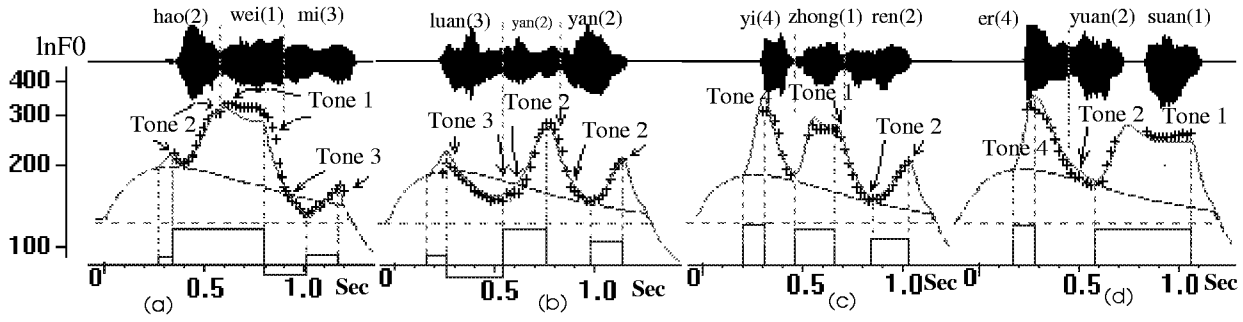

Fig.8 Comparison of rule-generated F0 contours and measured ones.