6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ANALYSIS AND SYNTHESIS OF THE FOUR TONES IN CONNECTED SPEECH OF THE STANNDARD CHINESE BASED ON A COMMAND-RESPONSE MODEL

Changfu Wang [1], *Hiroya Fujisaki* [2], *Sumio Ohno* [2] *and Tomohiro Kodama* [2]

[1] University of Science and Technology of China, Hefei, Anhui, China
[2] Science University of Tokyo, Noda, 278-8510 Japan

10.21437/Eurospeech.1999-376

ABSTRACT

The Chinese language is a typical tone language in which a syllable possesses several tone types and thus can represent different morphemes. While these tone types have rather clear manifestations in the fundamental frequency contour in isolated syllables, they vary considerably in connected speech due to the influences of such factors as tones of adjacent syllables, syntactic and pragmatic information of the whole utterance. This paper describes the results of analysis of F_0 contours of the Standard Chinese using a command-response model, and shows that systematic relationships exit between the timing of the tone commands and the "vowel plus coda" part of a syllable. The results are then used to derive rules for tone command generation in speech synthesis. The validity of the rules has been confirmed by the naturalness of prosody of synthetic speech.

1. INTRODUCTION

The Chinese language is a typical tone language in which a syllable possesses several tone types and thus can represent different morphemes. In Standard Chinese, there are four tone types; the first (high), second (rising), third (low), and fourth (falling), respectively denoted by T1, T2, T3, and T4, or simply by 1, 2, 3, and 4.

While these tone types have rather clear manifestations in the fundamental frequency contour (henceforth F_0 contour) in the case of isolated syllables, they vary considerably in connected speech due to the influences of such factors as tones of adjacent syllables, syntactic and pragmatic information of the whole utterance. The present paper describes some of the most recent results of our investigations to elucidate the nature of these influences, and to describe them in quantitative terms. Quantification of characteristics of an F_0 contour is accomplished using a command-response model, originally developed by Fujisaki and his coworkers for F_0 contours of Japanese [1, 2] and modified to apply to F_0 contours of the Standard Chinese [3, 4].

2. A MODEL FOR F_0 CONTOUR GENERATION OF THE STANDARD CHINESE

Figure 1 shows the model for F_0 contour generation of the Standard Chinese. The phrase commands generate the overall contour of an utterance, while the tone commands generate the local contours due to the presence of respective tones. While T1 and T3 are generated by a single tone command (positive in T1 and negative in T3), T2 and T4 are generated by a pair of tone commands (negative-positive in T2 and positive-negative in T4). These commands are applied to the respective control mechanisms which produce phrase and tone components. These mechanisms are assumed to be second-order linear systems. The phrase components and the tone components are added onto a constant value ($\ln F_b$), and their sum is then applied to the F_0 contour generation mechanism, which produces changes in the logarithm of F_0 in proportion to the combined phrase and tone components.

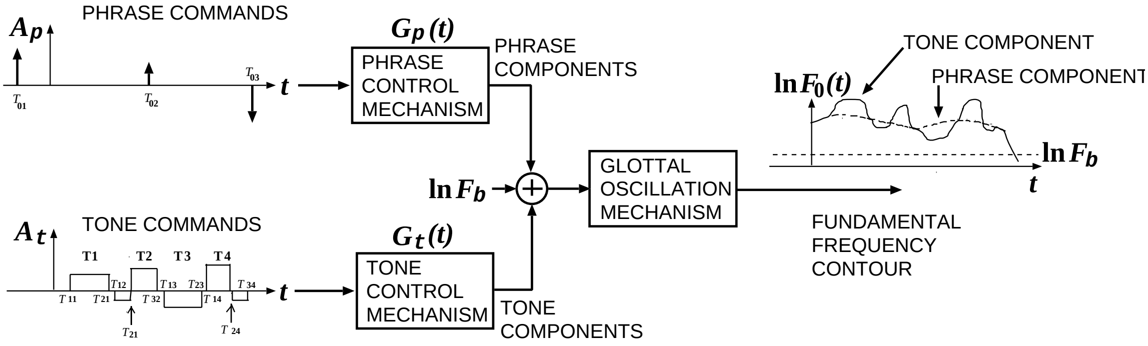

Fig. 1. A command-response model for F_0 contour generation of Standard Chinese.

Thus the F_0 contour as a function of time can be expressed by the following equations:

$$\begin{aligned}\ln F_0(t) \quad = \quad & \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) \\ & + \sum_{j=1}^{J} [A_{t1j} \{G_t(t - T_{1j}) - G_t(t - T_{2j})\} \\ & + A_{t2j} \{G_t(t - T_{2j}) - G_t(t - T_{3j})\}] \quad (1)\end{aligned}$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases} \qquad (2)$$

$$G_t(t) = \begin{cases} \min [1 - (1 + \beta t) \exp(-\beta t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \qquad (3)$$

where $G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_t(t)$ represents the step response function of the accent control mechanism.

The symbols in Eqs. (1) to (3) indicate

Fb	:	baseline value of fundamental frequency,
I	:	number of phrase commands,
J	:	number of tone command pairs,
Ap_i	:	magnitude of the ith phrase command,
A_{t1j}	:	amplitude of the first tone command in the jth command pair,
A_{t2j}	:	amplitude of the second tone command in the jth command pair,
T_{0i}	:	timing of the ith phrase command,
T_{1j}	:	onset of the first tone command in the jth command pair,
T_{2j}	:	end of the first tone command and onset of the secoond tone command in the jth command pair,
T_{3j}	:	end of the second tone command in the jth command pair,
α	:	natural angular frequency of the phrase control mechanism,
β	:	natural angular frequency of the tone control mechanism,
γ	:	relative ceiling level of tone components.

It is to be noted that the onset of the second tone command is constrained to coincide with the end of the first tone command within a tone command pair. Although Eq.(1) provides a pair of tone commands for every syllable, only one tone command is necessary for tones T1 and T3. Thus the amplitude of the second tone command is always set to zero in these tones.

3. ANALYSIS OF F_0 CONTOURS OF CHINESE UTTERANCES

3.1. Speech Material

The speech material for the present study was taken from a preliminary database of the spoken Standard Chinese collected at the University of Science and Techlonogy of China. It consists of recordings of more than 3000 isolated words uttered by three speakers (two males and one female) and a total of 480 sentences uttered by 12 speakers (6 males and 6 females). The speakers were staffs and students of USTC who were born in Beijing and are native speakers of the Standard Chinese.

3.2. Analysis Procedure

The speech signal was digitized at 10 kHz with 16 bit precision. The fundamental frequency was extracted at 10 msec intervals by the modified autocorrelation analysis of the LPC residual. For each utterance, the measured F_0 contour was aligned with the speech waveform whose syllable boundaries and onsets of vowels were marked by visual inspection of the waveform.

The validity of the proposed model can be tested by Analysis-by-Synthesis, i.e., by constructing the best approximation to an observed F_0 contour, and by examining the closeness of the approximation. The optimization is carried out by minimizing the mean squared error in the $\ln F_0(t)$ domain through a hill-climbing search in the space of model parameters. This allows one to decompose a given F_0 contour into its constituent components, and to estimate their underlying commands by deconvolution.

3.3. Experimental Results

Figure 2 shows an example of the Analysis-by-Synthesis of the F_0 contour of the utterance:

Mu4 ni2 hei1 buo2 lan3 hui4 bu2 kui4 shi4 dian4 zi3 wan4 hua1 tong3. (The Munich exposition is really an electronic kaleidoscope.)

The figure shows, from top to bottom, the speech waveform, the measured F_0 values (+symbols), the model-generated best approximation (solid line), the baseline frequency (dotted line), the phrase commands (impulses), and the accent commands (stepwise functions). The dashed lines indicate the contributions of phrase components, and the differences between the F_0 contour and the phrase components correspond to the accent components. Analysis of a number of Chinese utterances

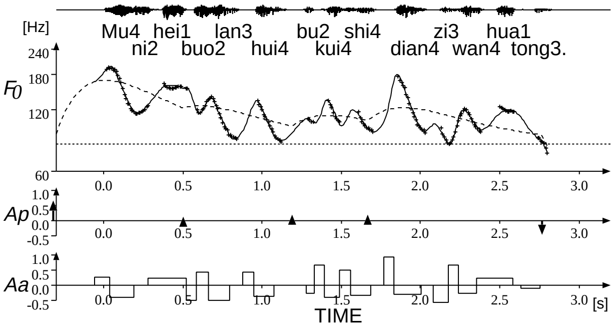

Fig. 2. An example of Analysis-by-Synthesis of an F_0 contour of the Standard Chinese.

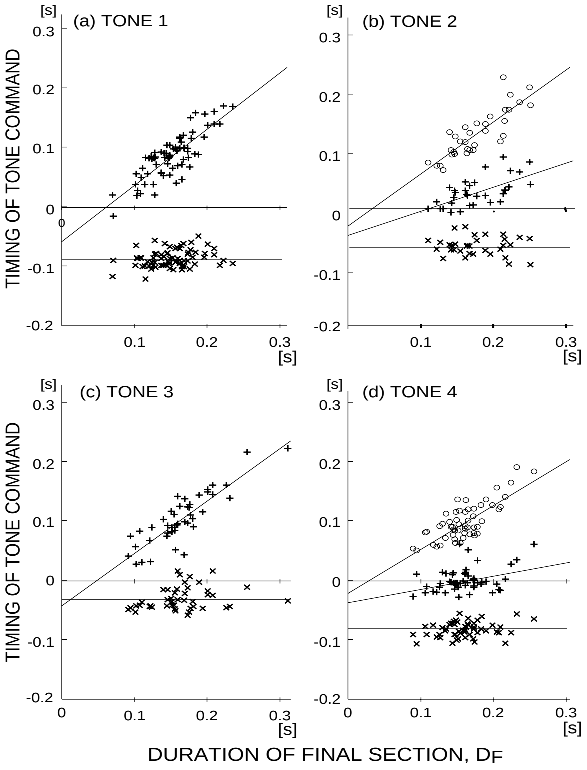

Fig. 3. Tone command timing relative to the onset and duration of final section of the corresponding syllable.

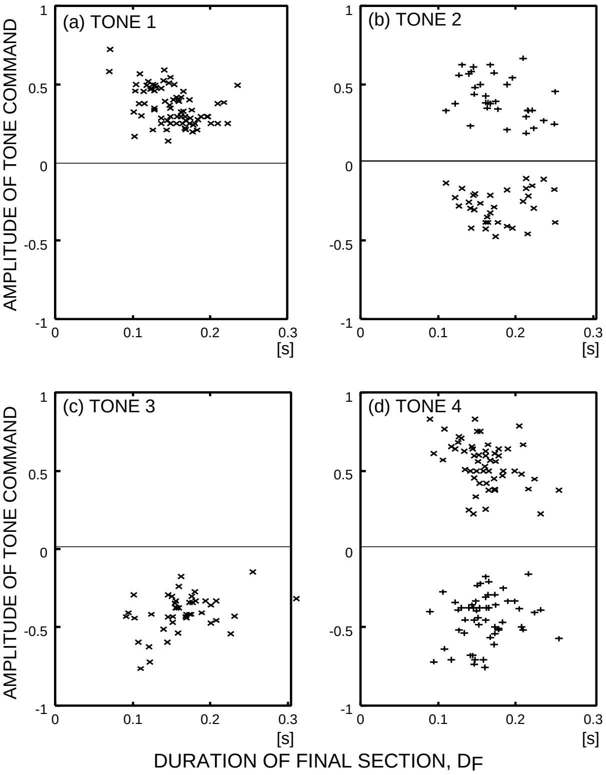

Fig. 4. Tone command amplitude versus duration of final section of the corresponding syllable.

has shown that the model can always generate very good approximations to the measured F_0 contours, if the timing and amplitude of the commands are optimized. These parameters represent the underlying linguistic information concerning the tones and intonation of a given utterance, and are useful both for the study of tone realization in connected speech and for speech synthesis by rule.

3.4. Timing and Amplitude of the Tone Commands

A syllable in spoken Chinese can be divided into two parts: the initial (i.e., the initial consonant) and the final (i.e., the rest of the syllable). Preliminary analysis showed that the timing of tone commands are closely correlated with the final part. Figures 3 (a)∼(d) show the timing of each tone command relative to the timing of final part of the corresponding syllable. The regression lines in each panel can be used as rules for determining tone command timing in speech synthesis.

Figures 4 (a)∼(d) show the amplitudes of tone commands relative to the duration of the final part of the corresponding syllable. Contrary to the data in Figs. 3 (a)∼(d), the correlation here is quite low, suggesting that the amplitudes of tone commands are determined mainly by factors other than syllable duration. The wide range of distribution of tone command amplitude suggests, however, that the amplitude may be quantized to several discrete levels without loss of naturalness in speech synthesis.

4. SYNTHESIS OF F_0 CONTOURS OF CHINESE UTTERANCES

Since the model can generate very close approximations to F_0 contours of natural utterances, it is clear that it can be quite useful in coding the F_0 contour information in terms of the command parameters. From the point of view of speech synthesis by rule, it is of interest to find out to what extent these parameters can be standardized and can be derived automatically from text. This section shows preliminary results toward synthesis by rule of F_0 contours of Chinese utterances using the command-response model.

In order to find out how such standardization will affect the quality of synthetic speech, the following rules were derived on the basis of measurements described in the preceding section. It should be noted here that the current rules are concerned only with the synthesis of F_0 contour, and seperate rules are necessary to determine syllable duration. In the present study, we use the syllable duration as found in natural utterances.

4.1. Rules for Tone and Phrase Commands

4.1.1. Rules for Tone Commands

• **Timing:** The timings of each tone command (T_{1j}, T_{2j}, and T_{3j}) are determined by the regression formulae for each tone type.

• **Amplitude:** Based on the results shown in Fig. 4 , the amplitudes of tone commands are standardized to three values for each tone type, as shown in Table 1.

4.1.2. Rules for Phrase Commands

Based on the results of a similar analysis of timing and magnitude of phrase commands, their values are standardized to four sets of values depending of whether the phrase command occurs at sentence-initial or sentence-medial positions, as shown in Table 2.

Table 1. Standardization of tone command amplitude.

Tone Command		Tone1	Tone2	Tone3	Tone4
First	Enhanced	0.45	−0.40	−0.65	0.70
	Normal	0.35	−0.25	−0.45	0.50
	Suppressed	0.25	−0.10	−0.25	0.30
Second	Enhanced		0.60		−0.75
	Normal		0.40		−0.50
	Suppressed		0.20		−0.25

Table 2. Standardization of timings and magnitudes of the phrase commands.

Position		Timing[s]*	Magnitude
Utterance-initial		−0.30	0.50
Utterance-medial	High	−0.25	0.35
	Medium	−0.20	0.25
	Low	−0.15	0.15

* The timing reference is the vowel onset of the initial syllable of the whole utterance.

4.2. Speech Synthesis and Evaluation

In order to test the validity of the current approach and to evaluate the effects of various rules on the synthesized speech, several versions of synthetic speech have been generated by LPC analysis-resynthesis and used for subjective evaluation of naturalness of prosody. They are LPC analysis-resynthesis with:

(a) the original F_0 contour
(b) synthetic F_0 contour with the minimum mean squared error
(c) synthetic F_0 contour in which only the tone commands are generated by rules
(d) synthetic F_0 contour in which only the phrase commands are generated by rules
(e) synthetic F_0 contour in which both the tone commands and phrase commands are generated by rules.

Although the difference between the original and the synthetic F_0 contours increases steadily as we go from (b) to (e), the result of a listerning test indicates that the subjective evaluation of naturalness of tone and intonation remains essentially unaffected by the successive introduction of rules.

5. SUMMARY AND CONCLUSIONS

This paper has described analysis and synthesis of F_0 contours of connected speech of the Standard Chinese using the command-response model developed by Fujisaki and his coworkers. In the first place, the validity of the model was confirmed by its ability of generating extremely close approximations to F_0 contours of all the utterances analyzed. The results of analysis then showed the existence of systematic relationships between the timing of the tone commands and the duration of the “final” part (i. e., “vowel plus coda part” of a syllable), indicating that the seemingly large variations in the shape of the local F_0 contour actually arise from these systematic relationships. On the other hand, no significant correlation has been found between the amplitude of the tone commands and the duration of the final part of a syllable. The wide range of distributions of the command amplitude, however, suggested the possibility of quantizing it to several levels without loss of naturalness of prosody, Finally, on the basis of these analysis results, rules have been derived for determining the timing and amplitude/magnitude of the tone/phrase commands in speech synthesis. The validity of these rules has been confirmed by the naturalness of prosody of synthetic utterances, judged by native speakers of the Standard Chinese. The current rules are concerned only with tone and phrase commands, whose timings are specified relative to the duration of the final part of each syllable. Work is in progress to develop rules for the duration of the initial and final parts of a syllable, as well as for the choice of levels of the tone and phrase commands.

REFERENCES

[1] Fujisaki, H. and Nagashima, S.: “A model for the synthesis of pitch contours of connected speech,” *Annual Report of the Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[2] Fujisaki, H. and Hirose, K.: “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” *J. Acoust. Soc. Jpn (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[3] Fujisaki, H., Hirose, K., Hallé, P. and Lei, H. T.: “Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese,” *Proceedings of 1990 International Conference on Spoken Language Processing*, vol. 2, pp. 841–844, 1990.

[4] Fujisaki, H., Hirose, K. and Lei, H. T.: “Prosody and syntax in spoken sentences of Standard Chinese,” *Proceedings of 1992 International Conference on Spoken Language Processing*, vol. 1, pp. 433–436, 1992.