Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 59–70
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model

Hiroya Fujisaki [a], Changfu Wang [b], Sumio Ohno [c], Wentao Gu [a,d,*]

[a] The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[b] University of Science and Technology of China, 96 Jinzai Road, Hefei, Anhui 230026, China
[c] Tokyo University of Technology, 1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan
[d] Shanghai Jiaotong University, 1954 Huashan Road, Shanghai 200030, China

Received 26 January 2005; received in revised form 5 May 2005; accepted 28 June 2005

Abstract

While the tonal characteristics of Chinese syllables have been qualitatively described in traditional phonetics, quantitative analysis requires a mathematical model. This paper presents such a model for the fundamental frequency contours of Standard Chinese, based on an extension of a model that has already been proved to be applicable to non-tone languages including Japanese, English, and others. The model allows one to interpret a given fundamental frequency contour in terms of tone commands and phrase commands, and to analyze various tonal phenomena in quantitative terms. The paper then describes the results of analysis of fundamental frequency contours of a number of utterances, revealing systematic relationships between the timing of the tone commands and the final of each syllable. The results are used to derive constraints for tone and phrase command generation in speech synthesis. The validity of the rules is confirmed by evaluating the naturalness of prosody of synthetic speech. The validity of introducing these constraints in speech synthesis of Standard Chinese is confirmed by perceptual tests on naturalness of prosody as well as on intelligibility of tones, using speech synthesized with and without these constraints.

Keywords: Standard Chinese; Tone; Fundamental frequency contour; Command–response model; Constraint; Perceptual test

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,

* Corresponding author. Tel.: +81 3 5841 6767; fax: +81 3 5841 6648.
E-mail addresses: fujisaki@alum.mit.edu (H. Fujisaki), ohno@cc.teu.ac.jp (S. Ohno), wtgu@gavo.t.u-tokyo.ac.jp (W. Gu).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.06.009

respectively denoted by Tone l, Tone 2, Tone 3, and Tone 4, or simply by Tl, T2, T3, and T4.

The primary distinctions between these four tone types are found in the contours of the voice fundamental frequency. The upper part of Fig. 1 illustrates one example of fundamental frequency contour for each of these four tone types for the monosyllabic words "yi". In traditional phonetics and phonology of Standard Chinese, these four tone types are respectively referred to be "high", "rising", "low" and "falling" tones.

While these tone types have rather clear manifestations in the fundamental frequency contour in the case of isolated syllables, they vary considerably in continuous speech due to the influences of such factors as tones of adjacent syllables, syntactic and pragmatic information of the whole utterance, and the overall speaking rate. The main factors causing these variations are tone coarticulation, tone enhancement/suppression, and phrasing, in addition to the well-known phenomenon of tone *sandhi*.

Although there have been numerous studies on tones and intonation of Standard Chinese, a fully quantitative representation of continuous fundamental frequency contour of Standard Chinese speech is a rather difficult problem. The existing approaches diverge widely in density of specified points, in number of parameters, in relationship between local tones and global phrase intonation, and in sophistication of mathematical representation. These approaches fall into two broad categories: model/rule-based approaches and data-driven approaches. The former category can further be subdivided into two groups: those with explicit representation of both global and local components, and those with explicit representation of only local components of the fundamental frequency contour. Here we give a brief overview on a few typical methods.

As far as the present authors are aware, a quantitative model for the fundamental frequency contours of Standard Chinese has first been proposed by Fujisaki and his coworkers. Namely, on the basis of earlier works on modeling the generation process of fundamental frequency contours of Common Japanese (Fujisaki and Nagashima, 1969; Fujisaki and Hirose, 1984), Fujisaki and his coworkers proposed a command–response model for fundamental frequency contours of Standard Chinese (Fujisaki et al., 1987, 1990, 1992). The original model for Common Japanese assumes the presence of two types of commands: the impulse-shaped phrase commands giving rise to phrase components for global intonation, and the pedestal-shaped accent commands giving rise to accent components for local undulation due to word accent. The contour of fundamental frequency, expressed in the logarithmic scale, is represented as the sum of these two types of components and a baseline component, which remains to be constant at least during an utterance. Instead of accent commands whose polarities are always positive in languages including Japanese and English, the model for Standard Chinese assumes tone commands whose polarities are either positive or negative. Each of the four tone types is associated with a specific pattern of tone commands, as shown in the lower part of Fig. 1.

The Soft Template Mark-Up Language (Stem-ML) proposed by Kochanski and Shih (2003) and applied to Standard Chinese (Shih and

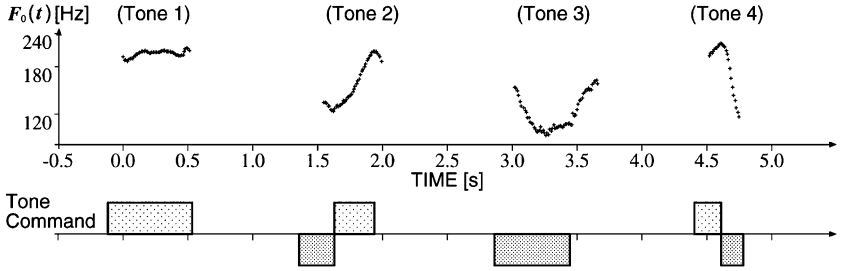

Fig. 1. The fundamental frequency contours and the underlying tone command patterns for the isolated words "yi" of four lexical tones. For example, "yi1" means one, "yi2" means maternal aunt, "yi3" means chair, and "yi4" means a hundred million.

Kochanski, 2000) is a tagging system which generates fundamental frequency contours from a set of mathematically defined mark-up tags, including both stress tags for local tone shapes and step and slope tags for global phrase curves. For each syllable, the resulting soft template is a compromise between articulation effort and communication accuracy. The balance is controlled by a *strength* parameter.

On the contrary, the intonation module for Standard Chinese in the Bell Labs TTS system (Shih and Sproat, 1996; van Santen et al., 1998) represents fundamental frequency contours by sparsely specifying a sequence of locally defined high (H) and low (L) tonal targets, based on early works by Pierrehumbert (1980) and Liberman and Pierrehumbert (1984). The identities of tonal targets are determined by a set of coarticulation rules. There is no explicit representation for global intonation, but the actual values of tonal targets are modulated by the downstep and declination effects.

The Parallel Encoding and Target Approximation (PENTA) model proposed by Xu (2004) also tries to model fundamental frequency contours of Standard Chinese in terms of local pitch targets (Xu et al., 1999). Multiple layers of communicative functions are encoded in parallel into melodic primitives, based on which fundamental frequency contour is implemented by successively approaching local pitch targets. Unlike bidirectional smoothing assumption in Stem-ML, PENTA model assumes asymmetrical contextual influences.

In addition to these model/rule-based approaches, a number of data-driven approaches have also been proposed for modeling fundamental frequency of Standard Chinese, for example, artificial neural network (Chen et al., 1998), CART (classification and regression tree), and other statistical methods (Chen et al., 1992; Chou et al., 1996; Yu et al., 2002). Recently, the HMM-based approach proposed by Tokuda et al. (1999) has also been applied for Standard Chinese (Ni et al., 2005). These data-driven approaches automatically derive the relationship between input linguistic labels and fundamental frequency values by use of machine learning algorithms, and require little linguistic knowledge. Although these data-driven approaches can produce quite natural fundamental frequency contours, they are short of generalization ability, especially when the training corpus is not large enough or when a new speaking style or a new speaker's voice need to be constructed.

The current study is a continuation of the work on the command–response model for Standard Chinese proposed by Fujisaki and his coworkers. From the point of view of speech synthesis, it is desirable if systematic relationships can be found among parameters of the model so that they can be expressed in the form of constraints in controlling the model parameters. The present paper aims at finding such relationships based on the analysis of fundamental frequency contours of a number of utterances. The validity of the approach is tested, first by checking the goodness of fit between observed and synthesized fundamental frequency contours obtained by introducing these relationships as constraints in the Analysis-by-Synthesis, and then by evaluating both the intelligibility of tones and the naturalness of prosody of synthetic speech generated under these constraints.

2. A model for generating the fundamental frequency contour of Standard Chinese

The command–response model represents the contour of the logarithm of fundamental frequency, i.e. $\ln F_0(t)$, as the sum of phrase components, accent/tone components, and a baseline level $\ln F_b$. For the rest of this paper, we shall call $\ln F_0(t)$ simply as the F_0 contour. Fig. 2 shows the model for the process of F_0 contour generation for Standard Chinese. The phrase commands generate the global contour of an utterance, while the tone commands generate the local contours due to the presence of tones. These commands are applied to the respective control mechanisms which produce phrase and tone components. These mechanisms are assumed to be critically-damped second-order linear systems. The command–response model not only generates close approximations to measured F_0 contours but also has physiological and physical bases (Fujisaki et al., 2000, 2004).

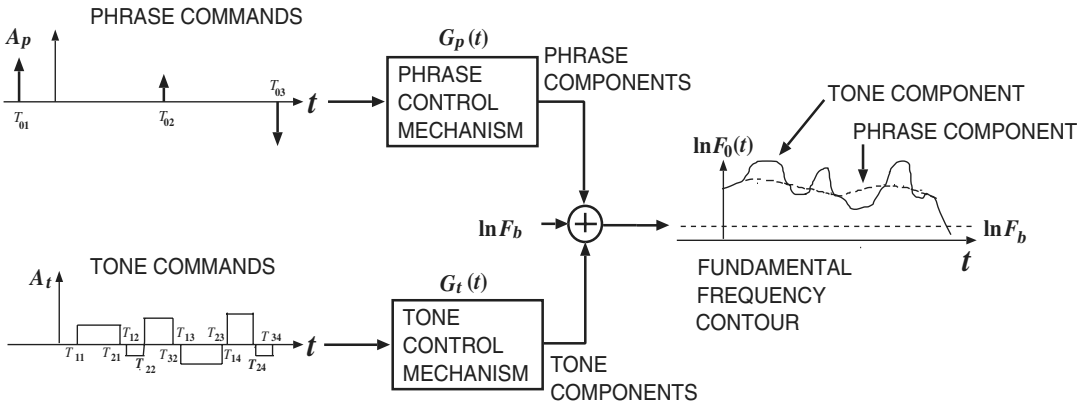

Fig. 2. A command–response model for the process of F_0 contour generation for Standard Chinese.

Unlike most non-tone languages, which have only positive accent commands, Standard Chinese requires both positive and negative tone commands (Fujisaki et al., 1987, 1990). While Tl and T3 are considered to possess a single tone command (positive in Tl and negative in T3), T2 and T4 are considered to have a pair of tone commands (negative–positive in T2 and positive–negative in T4). For the sake of uniformity of mathematical formulation, however, we allow that each tone type possesses a pair of tone commands, but assume that the second tone command has zero amplitude for T1 and T3.

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

$$\ln F_0(t) = \ln F_\mathrm{b} + \sum_{i=1}^{I} A_{pi} G_\mathrm{p}(t - T_{0i}) + \sum_{j=1}^{J} [A_{\mathrm{t}1j}\{G_\mathrm{t}(t - T_{1j}) - G_\mathrm{t}(t - T_{2j})\} + A_{\mathrm{t}2j}\{G_\mathrm{t}(t - T_{2j}) - G_\mathrm{t}(t - T_{3j})\}], \quad (1)$$

$$G_\mathrm{p}(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geqslant 0, \\ 0, & t < 0, \end{cases} \quad (2)$$

$$G_\mathrm{t}(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & t \geqslant 0, \\ 0, & t < 0, \end{cases} \quad (3)$$

where $G_\mathrm{p}(t)$ represents the impulse response function of the phrase control mechanism and $G_\mathrm{t}(t)$ represents the step response function of the tone control mechanism. The symbols in Eqs. (1)–(3) indicate

F_b	baseline value of fundamental frequency
I	number of phrase commands
J	number of syllables
$A_{\mathrm{p}i}$	magnitude of the ith phrase command
$A_{\mathrm{t}1j}$	amplitude of the first command in the jth syllable
$A_{\mathrm{t}2j}$	amplitude of the second command in the jth syllable
T_{0i}	timing of the ith phrase command
T_{1j}	onset of the first command in the jth syllable
T_{2j}	end of the first command (and onset of the second command if the second command exists) in the jth syllable
T_{3j}	end of the second command if the second command exists in the jth syllable
α	natural angular frequency of the phrase control mechanism, set empirically at 3/s
β	natural angular frequency of the tone control mechanism, set empirically at 20/s
γ	relative ceiling level of tone components, set empirically at 0.9

Exactly speaking, the natural angular frequency β of the tone control mechanism may not necessarily be constant, but may vary with the polarities of tone commands. On the basis of a preliminary study, however, we use the same value of β for

both positive and negative tone commands. Also, the end of the first tone command may not necessarily coincide with the onset of the second tone command for T2 and T4, but here again they are assumed to coincide for the sake of simplicity of model formulation.

This model incorporates the effects of tone coarticulation, tone enhancement/suppression and phrase intonation in an explicit way. Namely, tone coarticulation is automatically taken care of by the transfer characteristics of the tone control mechanism. Tone enhancement/suppression (i.e. word emphasis/de-emphasis) is implemented by magnifying/minifying the amplitude or lengthening/shortening the duration of tone command or both (Chen et al., 2004), while emphasis on a compound word group or a short phrase can also be implemented by an additional phrase command or an increased magnitude of phrase command. Finally, phrase intonation is explicitly represented by the phrase components.

3. Analysis of F_0 contours of utterances of Standard Chinese

3.1. Speech material

The speech material for the present study was collected at the University of Science and Technology of China. It consists of utterances read by 12 native speakers of Standard Chinese (6 male and 6 female) at various speaking rates. In this paper we only present the result of analysis based on 40 utterances recorded by one male speaker at his normal speaking rate, which turns out to be 4.8 syllables/s.

3.2. Analysis procedure

The speech signal was digitized at 10 kHz with 16 bit precision. The fundamental frequency was extracted at 10 ms 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 and the spectrogram.

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.

In the current study, Analysis-by-Synthesis is conducted manually based on the known information of tone types and syntactic structure. Basically, tone commands for each syllable comply with the inherent command patterns for the particular tone type (Fujisaki et al., 1987, 1990), while the occurrences of phrase commands usually coincide with major intonation groups which are largely consistent with various syntactic boundaries but do not always follow the same hierarchical structure of syntax (Fujisaki et al., 1992).

3.3. Analysis results

Analysis of a number of Chinese utterances 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.

Fig. 3 shows an example of the Analysis-by-Synthesis of the F_0 contour of the utterance:

Mu4 ni2 hei1 bo2 lan3 hui4 bu2 kui4 shi4 dian4 zi3 wan4 hua1 tong3.

(The Munich exposition is really an electronic kaleidoscope.)

Fig. 3 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 tone commands (stepwise functions). The dashed lines

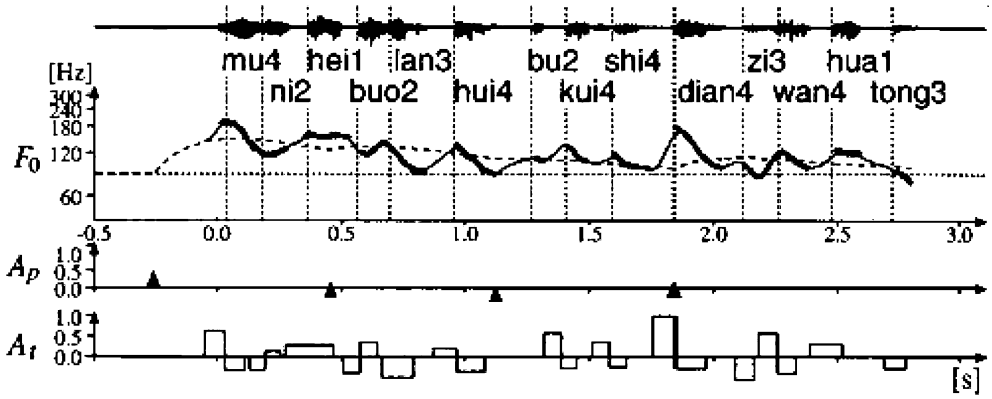

Fig. 3. An example of Analysis-by-Synthesis of the F_0 contour of an utterance of the sentence: "Mu4 ni2 hei1 bo2 lan3 hui4 bu2 kui4 shi4 dian4 zi3 wan4 hua1 tong3."

indicate the contributions of phrase components, and the differences between the F_0 contour and the phrase components correspond to the tone components. The RMS error between the observed and approximated $\ln F_0$ values within voiced intervals is 0.0038. This is equivalent to a very small percentage error of 0.38% in F_0.

In this utterance, the tone commands for each syllable mostly coincide with their inherent polarity patterns. Various contextual effects on tones can be observed. For example, tone articulation is distinctly shown by comparison between the F_0 patterns for "zi3" and "tong3", where the F_0 rising in the later part of "zi3" is due to the fact that the positive command for the following "wan4" has already risen before the end of "zi3". For another example, among those syllables of T4 in the utterance, "kui4" and "shi4" in the predicate generally shows smaller command amplitude or shorter command duration than others. This can be explained by the smaller degree of default emphasis on the predicate.

The occurrences of phrase commands in this utterance are consistent with the syntactic structure, though not strictly following the hierarchy of syntax. Four phrase commands occur before the noun "mu4 ni2 hei1," the noun "bo2 lan3 hui4," (these two nouns combine into the subject) the predicate "bu2 kui4 shi4," and the noun object "dian4 zi3 wan4 hua1 tong3," respectively. It is shown that the utterance-initial phrase command gives the highest amplitude, which is also observed in almost all other utterances. With the introduction of phrase components, it is quite easy to explain the different average F_0 levels for those syllables of T4 by their different positions in the respective prosodic phrases, for example, "mu4" is higher than "hui4", "kui4" is higher than "shi4", and "dian4" is higher than "wan4". Such an observed overall declination can be handled by the model very well.

3.4. Timing and amplitude of the tone commands

3.4.1. Tone command timing

According to traditional Chinese phonetics, a syllable in Standard Chinese can be divided into two parts: the initial (i.e., the syllable onset, composed of an initial consonant including null-initial) and the final (i.e., the rhyme of the syllable, composed of an optional semi-vowel, a nucleus and an optional coda). Since tone is a property of a syllable, the occurrence of tone commands should be closely related to segmental timing of the corresponding syllable. Our preliminary investigation indicated that the correlation is the highest if we adopt the onset of the final as the reference of segmental timing, as shown in Fig. 4.

Fig. 5(a)–(d) shows the timing of each tone command versus duration of the final of the corresponding syllable (Wang et al., 1999). The symbols ×, +, and ○ in these figures respectively indicate the measured values of T_{1j}, T_{2j}, and T_{3j} obtained by the above-mentioned analysis, while the straight lines

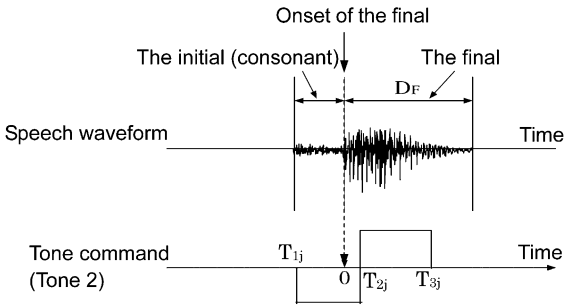

Fig. 4. Definition of tone command timing. (For the jth syllable, T_{1j}, T_{2j} and T_{3j} are defined relative to the onset of the final of the syllable.)

indicate approximate regression lines. These figures indicate that the timing of onset of tone command is hardly influenced by duration of the final, but the timing of offset is highly correlated with duration of the final and varies almost linearly with it. Thus they suggest that good approximations to actual F_0 contours will be obtained even if the timing of each tone command is constrained to follow the respective lines as shown in Fig. 5.

3.4.2. Tone command amplitude

In the first place, we investigate the relationship between tone command amplitude and duration of

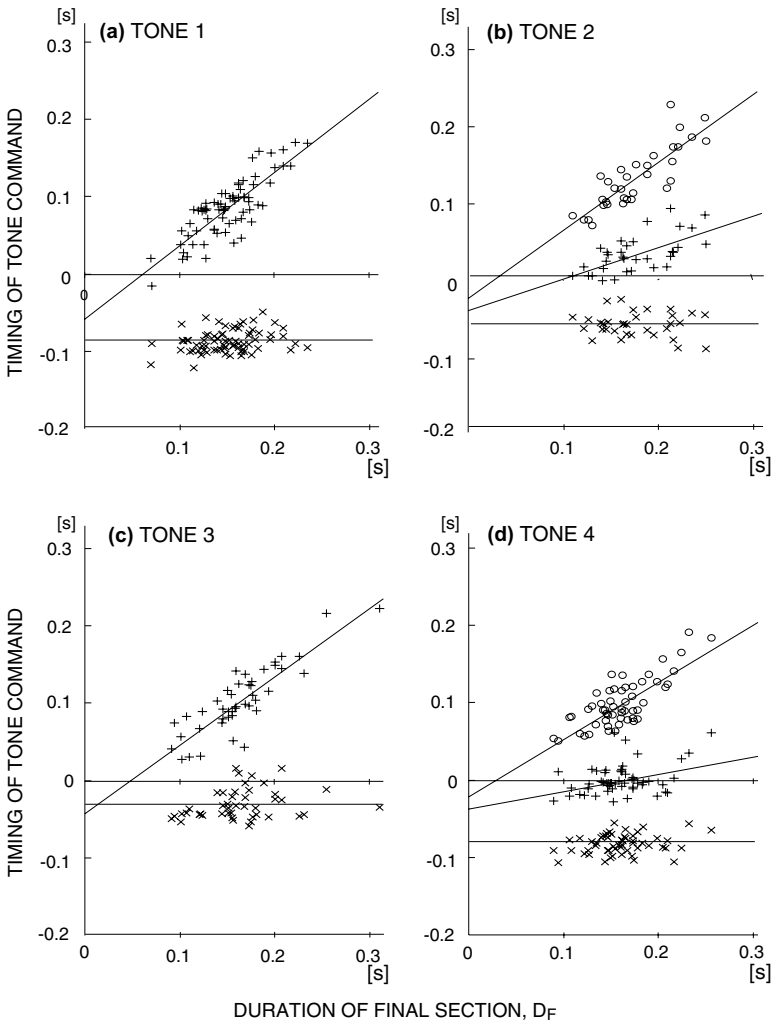

Fig. 5. Tone command timing relative to the onset and duration of final section of the corresponding syllable.

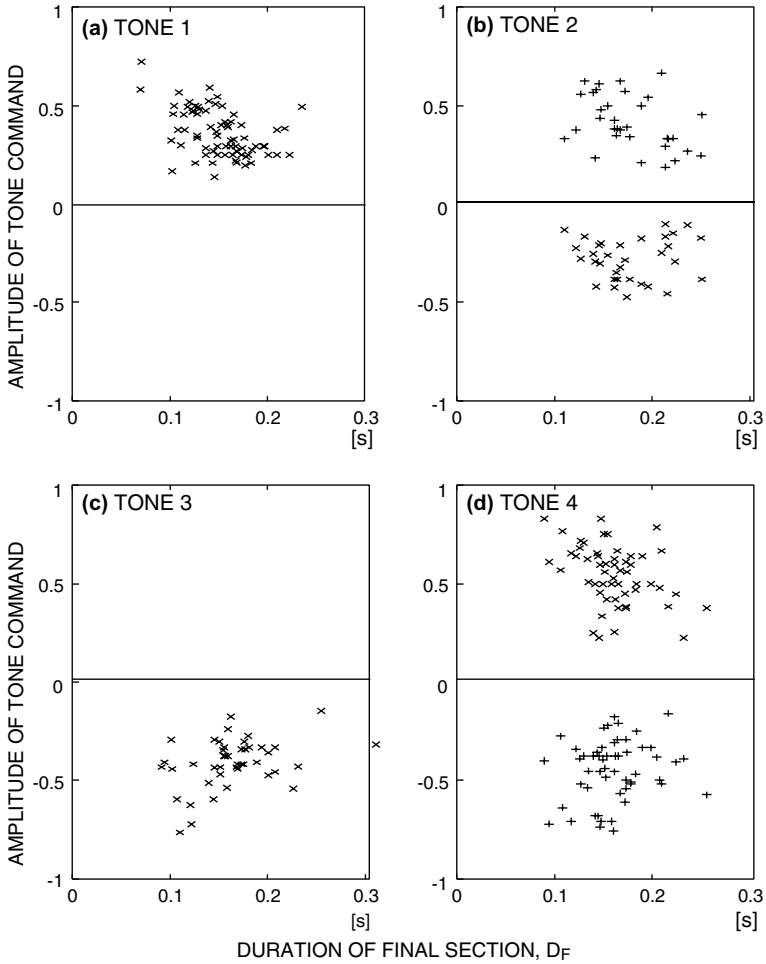

Fig. 6. Tone command amplitude versus duration of final section of the corresponding syllable.

the final. Fig. 6(a)–(d) shows the amplitudes of tone commands versus duration of the final of the corresponding syllable (Wang et al., 1999). Unlike the data in Fig. 5(a)–(d), the correlation here is quite low, suggesting that the amplitude of tone command is not affected by duration of the final but is determined by other factors. For example, the amplitude of tone command will increase when a narrow focus is placed on the word (Chen et al., 2004).

Secondly, we investigate the relationship between the absolute values of amplitudes of the two commands in T2 and T4 syllables. Fig. 7(a) and (b) shows the ratios of the absolute values of the amplitudes of the two tone commands versus the absolute value of the amplitude of the first tone command for T2 and T4, respectively. The straight lines indicate regression lines (Kodama et al., 1999). Although individual measured points show certain deviations from the regression lines, the results still suggest the possibility of introducing further constraints on the ratios of the absolute values of the amplitudes of the two tone commands for T2 and T4.

3.5. Comparison of results of Analysis-by-Synthesis with and without constraints

In order to find out the effects of constraints on tone command timing and on tone command

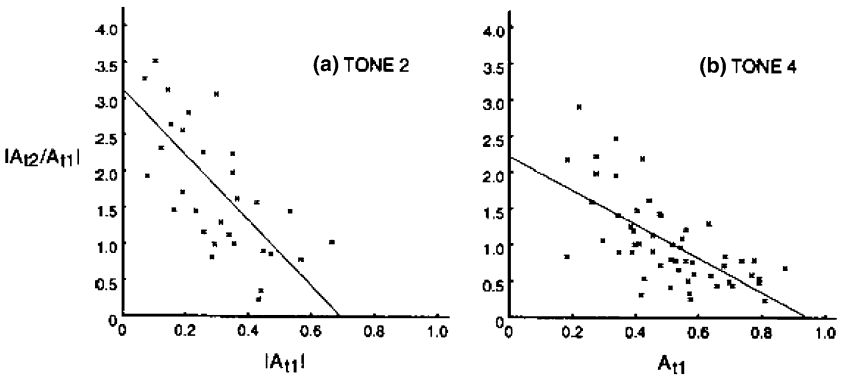

Fig. 7. Ratios of the absolute amplitudes of the two tone commands versus amplitude of the first tone command for Tone 2 and Tone 4.

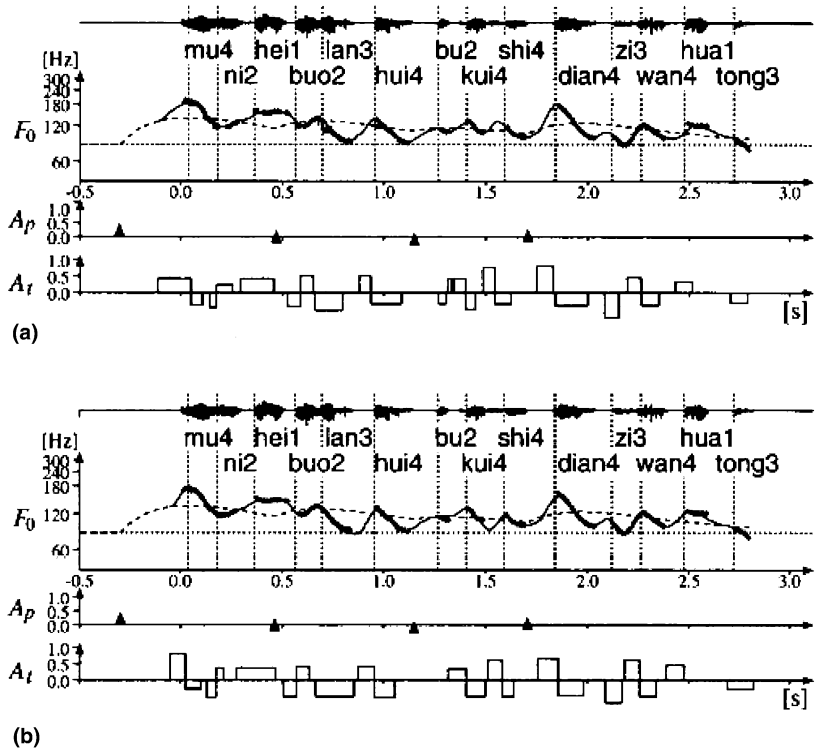

Fig. 8. Analysis-by-Synthesis of the F_0 contour of the utterance as given in Fig. 3 with different constraints. (a) With constraints on tone command timing. (b) With constraints both on timing and on amplitude ratio of the tone commands.

amplitude ratio, the F_0 contours were re-analyzed. Fig. 8(a) and (b) shows the results of Analysis-by-Synthesis of the F_0 contour of the same utterance as shown previously in Fig. 3 but obtained with different constraints: Panel (a) with constraints only on tone command timing, and Panel (b) with constraints both on timing and on amplitude ratio of the tone commands for T2 and T4 syllables.

Comparison of the two figures with Fig. 3 indicates that the model is capable of generating good approximations to the original F_0 contours even when the constraints on timing and amplitude ratio of tone commands are imposed.

4. Synthesis and perceptual evaluation of F_0 contours of utterances of Standard Chinese

4.1. Introduction of constraints for speech synthesis

Since the model can generate very close approximations to F_0 contours of natural utterances, it is expected to be useful in coding the F_0 contour information in terms of the command parameters. From the point of view of speech synthesis, it is of interest to find out to what extent these parameters can be standardized. This section describes results of our preliminary study toward standardization of parameters in the synthesis of F_0 contours of Standard Chinese utterances.

In order to find out how such standardization will affect the quality of synthetic speech, the following constraints are imposed on the timing and amplitude of tone commands on the basis of measurements described in the preceding section. As for tone command timing, we examine the effects of constraining it to follow the straight lines obtained experimentally as in Fig. 5. As for tone command amplitude, on the other hand, we examine the effects of constraining it to take one of three possible values (i.e. quantizing), while keeping the amplitude ratio of the pair of tone commands for T2 and T4 to fo'llow approximately the respective regression lines shown in Fig. 7. The purpose of quantization is to facilitate derivation of rules for mapping the input text onto a small number of levels instead of a continuum of tone command amplitudes in text-to-speech synthesis. The actual tone command amplitudes used for synthesis are shown in Table 1.

Although we have also conducted a similar study on the timing and magnitude of phrase commands, we will describe here only the results on tone commands, because of their importance in speech synthesis of Standard Chinese. For the following experiments, therefore, we keep the phrase commands as derived by analysis of natural utterances. It should also be noted here that the current study is concerned only with the synthesis of F_0 contours, while separate constraints are necessary for determination of syllable duration. In the present study, we use the syllable duration as found in natural utterances.

Table 1
Quantization of tone command amplitudes

Tone command	Tone 1	Tone 2	Tone 3	Tone 4
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		−0.55
Normal		0.50		−0.50
Suppressed		0.25		−0.45

4.2. Speech synthesis and subjective evaluation

In order to test the validity of the current approach and to evaluate the effects of various constraints on the quality of synthetic speech, the following five versions of synthetic speech were generated from the original utterance shown in Fig. 3 by the method of LPC analysis-resynthesis with:

(a) the original F_0 contour,
(b) model-based F_0 contour with the minimum mean squared error,
(c) model-based F_0 contour in which only timing of tone commands are constrained,
(d) model-based F_0 contour in which only amplitudes of tone commands are constrained (and quantized),
(e) model-based F_0 contour to which both the constraints in (c) and (d) are applied.

Two subjective evaluation tests were conducted using these synthetic speech. The subjects were six native speakers of Standard Chinese (four males and two females).

The first test (Test 1) was to evaluate the overall naturalness of prosody of each utterance, while the second test (Test 2) was to evaluate the intelligibil-

Table 2
Result of Test 1—mean scores for the overall naturalness of prosody of the synthetic speech

Stimulus	Subject						Average
	S1	S2	S3	S4	S5	S6	
(a)	10	9.8	10	10	8.1	10	9.65
(b)	9.0	8.7	10	10	8.1	9.4	9.12
(c)	8.3	7.6	8.5	9.5	7.6	9.1	8.43
(d)	8.5	8.6	9.4	9.5	8.0	9.1	8.85
(e)	8.5	7.4	8.5	9.5	7.8	9.1	8.47

ity of tones within each utterance. The method of constant stimuli was adopted as the test paradigm. In each test, the five versions of synthetic speech were repeated ten times to produce 50 synthetic utterances, which were then randomized and presented sequentially to each subject for listening and immediate evaluation. A 10-point scale was used for scoring: 10 corresponding to 'perfect', while 1 corresponding to 'extremely poor.' Training sessions were given for each subject in order to familiarize him/her with the test procedure until he/she can give the scores consistently.

For Test 1, the mean scores by each subject are presented in Table 2. Although in general the scores tend to decrease slightly when going from (a) to (b) and then to (c)–(e), the degradations are quite small, and even negligible for some subjects (S4 and S5). This result indicates that the naturalness of F_0 contours remains essentially unaffected by successive introduction of the constraints. For Test 2, all the six subjects give the full score (i.e. 10) to all the 50 utterances, indicating that tone identities are not affected by the above constraints at all. Thus, the results of these two subjective evaluation tests have confirmed the validity of introduction of these systematic constraints in speech synthesis.

5. Conclusions

This paper has described analysis and synthesis of F_0 contours of continuous speech of 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 utterances. The results of analysis then showed the existence of systematic relationships between the timing of tone commands and duration of the final of a syllable, indicating that the seemingly large variations in the shape of the local F_0 contour for each tone can be explained as the consequence of these systematic relationships. The results also indicated that the ratio of the amplitudes of the two tone commands for T2 as well as for T4 varies systematically with the amplitude of the first tone command. These findings suggested that these systematic relationships can be used to constrain the timing and amplitude of tone commands. Analysis-by-Synthesis of F_0 contours indicated that the model can generate good approximations to actual F_0 contours even when these constraints are imposed. On the other hand, no significant correlation has been found between the amplitude of tone command and the duration of the final of a syllable. The wide ranges of distributions of the amplitudes of tone commands, suggested the possibility of quantizing them to relatively small numbers of levels without causing appreciable degradation of quality of synthetic speech. On the basis of these analysis results, constraints have been derived for determining the timing and amplitude of tone commands in speech synthesis. The validity of these constraints has been confirmed by the high intelligibility of tones and naturalness of prosody of synthetic speech, judged by six native speakers of Standard Chinese.

References

Chen, S.H., Chang, S., Lee, S.M., 1992. A statistical model based fundamental frequency synthesizer for Mandarin speech. Journal of the Acoustical Society of America 92 (1), 114–120.

Chen, S.H., Hwang, S.H., Wang, Y.R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. IEEE Transaction on Speech and Audio Processing 6 (3), 226–239.

Chen, G.P., Hu, Y., Wang, R.H., Mixdorff, H., 2004. Quantitative analysis and synthesis of focus in Mandarin. In: Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tone Languages, Beijing, China, pp. 25–28.

Chou, F.C., Tseng, C.Y., Lee, L.S., 1996. Automatic generation of prosody structure for high quality Mandarin speech synthesis. In: Proceedings of ICSLP 1996, Philadelphia, USA, pp. 1624–1627.

Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustic Society of Japan 5 (4), 233–242.

Fujisaki, H., Nagashima, S., 1969. 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.

Fujisaki, H., Hallé, P., Lei, H., 1987. Application of F_0 contour command–response model to Chinese tones. Acoust. Soc. Jpn Autumn Meeting, Japan, pp. 197–198.

Fujisaki, H., Hirose, K., Hallé, P., Lei, H., 1990. Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese. In: Proceedings of ICSLP 1990, Kobe, Japan, pp. 841–844.

Fujisaki, H., Hirose, K., Lei, H., 1992. Prosody and syntax in spoken sentences of Standard Chinese. In: Proceedings of ICSLP 1992, Banff, Canada, pp. 433–436.

Fujisaki, H., Tomana, R., Narusawa, S., Ohno, S., Wang, C., 2000. Physiological mechanisms for fundamental frequency control in Standard Chinese. In: Proceedings of ICSLP 2000, Beijing, China, pp. 9–12.

Fujisaki, H., Ohno, S., Gu, W.T., 2004. Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command–response model for generation of their F_0 contours. In: Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tone Languages, Beijing, China, pp. 61–64.

Kochanski, G., Shih, C., 2003. Prosody modeling with soft templates. Speech Communication 39 (3–4), 311–352.

Kodama, T., Ohno, S., Fujisaki, H., Wang, C., 1999. Analysis and synthesis of fundamental frequency contours of speech of the Standard Chinese based on the command–response model. In: Proceedings of 1999 Japan-China Symposium on Advanced Information Technology, Tokyo, Japan, pp. 31–38.

Liberman, M., Pierrehumbert, J., 1984. Intonational invariance under changes in pitch range and length. In: Aronoff, M., Oehrle, R. (Eds.), Language Sound Structure. MIT Press, Cambridge, MA, pp. 157–233.

Ni, J.F., Kawai, H., Toda, T., Tokuda, K., Nishizawa, N., 2005. A Chinese text-to-speech system at ATR. In: Proceedings of 2005 Spring Meeting of ASJ, Tokyo, Japan, pp. 287–288.

Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Ph.D. dissertation, MIT, Cambridge, MA.

Shih, C., Kochanski, G.P., 2000. Chinese tone modeling with Stem-ML. In: Proceedings of ICSLP 2000, Beijing, China, Vol. 2, pp. 67–70.

Shih, C., Sproat, R., 1996. Issues in text-to-speech conversion for Mandarin. Computational Linguistics and Chinese Language Processing 1 (1), 37–86.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In: Proceedings of ICASSP 1999, Phoenix, USA, pp. 229–232.

van Santen, J.P.H., Shih, C., Mobius, B., 1998. Intonation. In: Sproat, R. (Ed.), Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Dordrecht, pp. 141–190.

Wang, C.F., Fujisaki, H., Ohno, S., Kodama, T., 1999. Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command–response model. In: Proceedings of EUROSPEECH 1999, Budapest, Hungary, pp. 1655–1658.

Xu, Y., 2004. Transmitting tone and intonation simultaneously—the parallel encoding and target approximation (PENTA) model. In: International Symposium on Tonal Aspects of Languages—with Emphasis on Tone Languages 2004, Beijing, China, pp. 215–220.

Xu, C.X., Xu, Y., Luo, L.S., 1999. A pitch target approximation model for F_0 contours in Mandarin. In: Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, pp. 2359–2362.

Yu, M.S., Pan, N.H., Wu, M.J., 2002. A statistical model with hierarchical structure for predicting prosody in a Mandarin text-to-speech system. In: Proceedings of ISCSLP 2002, Taipei.