A FUNCTIONAL MODEL FOR GENERATION OF THE LOCAL COMPONENTS OF F0 CONTOURS IN CHINESE

Jinfu Ni , Renhua Wang and Deyu Xia

University of Science & Technology of China
P.O. Box 4, Hefei, AnHui, 230027, The People's Republic of China

ABSTRACT

In this paper, a new functional model is introduced, which is designed to simulate the control mechanism for the generation of the local component of F0 contour. With the model two kinds of motor command are defined to control the model parameters to generate two basic rise-fall feature patterns, on the logarithmic scale of fundamental frequency the local component of a F0 contour is approximated by the algebra sum of these patterns. The experimental results in analyzing and synthesizing Chinese F0 contours indicate that the real F0 contours can always be approximated very closely by the model, and a close correlation exists between the model parameters and the structure of rise-fall pattern.

1. INTRODUCTION

The fundamental frequency of the voice (henceforth the F0) plays an important role in the transmission of linguistic information concerning lexical tone, sentence structure and the discourse structure. In order to construct prosodic rules for synthesizing high-quality speech, the F0 contour of speech should be quantitatively analyzed with respect to the linguistic information with the aid of a model. Actually, In the previous work [1,2,3,4], various models have already been designed either to analyze and interpret observed F0 contours or to generate F0 contour for speech synthesis by rule, especially, the idea of pitch being the response of a mechanical system to excitation commands has been proven to be successful in the model proposed by professor H.Fujisaki and his coworkers in the analysis and synthesis of natural F0 contours of Japanese and even English [1,2], which contributes a valuable insight into the structure of intonation. However, prosodic features mostly differ form language to language, almost all the models for generating F0 contour, more or less, depend on the prosodic features of native language. It is true that, untill now, no existed model can be adopted successfully for the analysis and synthesis of Chinese F0 contours.

Chinese is a typical tone language, in which the degree of prominence at different syllabic position of a word, tone and intonation are all signaled primarily in terms of F0 contour. With focus on the structure of F0 contour, however, like other languages, the F0 contour of a Chinese utterance can be characterized by the presence of local rise-fall component superposed on a declining baseline,too. In the study of F0 contours of Chinese words, we postulate the existence of two cardinal rise-fall patterns, rise-fall component and rise-high-fall component, which forms the rich local characteristics of F0 contour. In this paper, we introduce a functional model to simulate the control mechanism for the generation process of the rise-fall patterns and a command-pattern proposed for Chinese tonal F0 pattern, and then show some experimental results to examine for the validity of the model for generating Chinese F0 contours.

2. MODELING AT PHYSICAL LEVEL

The characteristics of F0 contour is determined by the laryngeal F0 mechanisms, which mainly consist of the rotation mechanism of the thyroid cartilage and the cricoid cartilage. The rise-fall pattern, as the cardinal component in F0 contour, is the inevitable consequence of such mechanisms. It is well known that the principal biomechanical factor in the regulation of F0 is the tension of the vocal cord, it is also believed that vocal cord tension is regulated by the vocal cord length which is determined by the angle of the circothyroid joint [6]. During the rotation of the cricothyroid joint, because of such laryngeal muscles as thyroarytenid's active motion, the equivalent stiffness of the vocal cord has both linear and nonlinear components. It is expected that the nonlinear stiffness should play an important role in the process of producing the transition characteristics of F0 raising or lowering. To make good use of the knowledge to establish a quantitative model, we quote and set up the following hypotheses:

Hypothesis 1: *The logarithm of fundamental frequency varies linearly with the strain, i.e. elongation, of the vocal cord* [2].

Hypothesis 2: *The dynamic behavior of the vocal cord strain can be described as the response characteristics of the forced vibration of a micro-vibration system under the excitation of certain external force with the form of* $A_F \sin(\omega_t t + \varphi)$.

Hypothesis 3: *(1) The influence of the nonlinear component of the vocal cord stiffness on the dynamic behavior of the vocal cord strain can be regarded as the influence of the external force on the response characteristics of the forced vibration system. (2) Furthermore, the tension change of the vocal cord caused only by the nonlinear stiffness component can be simulated by the output response of two critically-damped second-order linear systems each corresponding to F0 raising and lowering process respectively, then it is converted into the angular frequency,* ω_t, *as that of the external force to influence the response characteristics of the forced vibration.*

Based on the hypotheses, a functional model expressed by Eq.(1) is deduced from three mechanical models [6].

$$G_{RP}(t,T_d)=\frac{A_m}{1-2\zeta\delta\sqrt{1-\zeta^2}}*\left(\frac{1}{\sqrt{(1-\eta\lambda(t))^2+4\zeta^2\eta\lambda(t)}}-\delta\right], \quad (1)$$

$$\lambda(t)=\begin{cases}1-(1-\gamma t)e^{-\gamma t} & T_d>t>=0, T_d>=\gamma^{-1};\\ \left(1+\sigma(t-T_d)\right)e^{-\sigma(t-T_d)}, & t>=T_d,\end{cases}$$

where:

A_m: a magnitude parameter;

ζ : ratio of damp;

η : a coefficient, $0<\eta<=1.0$;

γ : the natural angular frequency of a critically-damped second-order linear system which controls F0 raising transition;

T_d: the continuation time of F0 raising and keeping transition;

σ : the natural angular frequency of a critically-damped second-order linear system which controls the F0 lowering transition;

δ : the factor of negative correlation which decides the falling of rise-fall pattern down to a global baseline.

Among these parameters, the parameters ζ and η, viz. ration of damp and a coefficient, are regarded as system constant, generally are set equal to 0.1 and 0.96 respectively. A_m controls the magnitude of the rise-fall pattern, the parameters γ , σ determine the transition characteristics of the pattern. The parameter δ mainly serves as Chinese tone 3, which has a lower pitch feature, the more emphasis a tone 3 takes, the lower region its pitch goes, while δ usually keeps 1.0 for other tones. Fig.1 shows the response curve of the model varies with parameters A_m, γ , σ, in which ζ =0.1, η=0.96, δ =1.0.

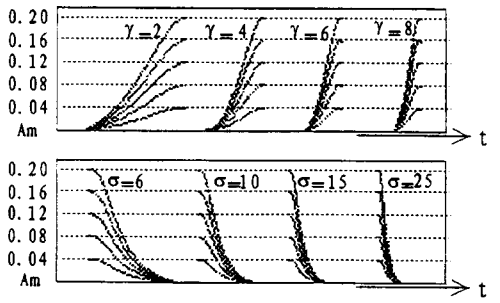

Fig.1 The diagram of the response curve of the model varying with model parameters $A_m-\gamma$ on top, and with model parameters $A_m-\sigma$ at the bottom.

3. MODELING AT LINGUISTIC LEVEL

With the functional model, two commands are defined by a set of parameters $\{Tm_0, A_m, \gamma, Tm_1, \sigma, \delta\}$, namely RF command for rise-fall pattern, and RHF command for rise-high-fall pattern, where, Tm_0 is the onset of a command, A_m, magnitude of a command, γ , σ are the time constants of a command, Tm_1 is the beginning time of the F0 falling transition in rise-fall pattern. we distinguish RF command in which $T_{m1}-T_{m0}=\gamma^{-1}$ from RHF command in which $T_{m1}-T_{m0}>\gamma^{-1}$ only by the relationship between $T_{m1}-T_{m0}$ and γ^{-1}. Fig.2 shows the generated patterns and the corresponding command symbols. It can be seen from the figure that the parameter Tm_1 comes later in RHF command than in RF command, consequently, the characteristics posture will remain when the rising transition reaches the maximum value, such reflects the physiological property that the vocal cord tension reaches a certain extent and keeps for a period.

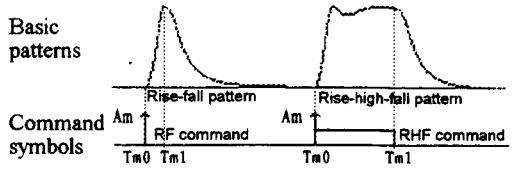

Fig.2 The basic patterns and corresponding command.

Next, discuss the application of the rise-fall pattern to Chinese tone F0 pattern. In standard Chinese, there phonologically exist four lexical tones, denoted by tone 1, tone 2, tone 3 and tone 4, and a neutral tone denoted by tone 0. It is well known that F0 contour for each lexical tones has its own basic pattern, and the basic pattern is almost identical across syllables when the syllables are uttered in isolation. In connected speech, although the actual value of tones is changed, the tone modification or tone sandhi is triggered in certain tonal environments, and even its tone disappears when a syllable is completely unstressed, yet Y.R.Chao (1933) assumed that one of the following four features was simultaneously added to tones to form the resultant sentence melodies: (1) general raised level of pitch, (2) general lowered level of pitch; (3) pitch range widened; and (4) pitch range narrowed, and the instrumental analysis has also confirmed that each lexical tone is capable of keeping its own basic pattern in a normal stress environment. As a result of above consideration, a stable command-pattern should be associated to each lexical tone. A schematic rule is illustrated in Fig.3, particularly, one RHF command for tone 1, two RF commands for tone 2 and tone 3, and one RF command for tone 4, tone 0 is assumed to be associated with no command at all. We call such single or pairs of RF, RHF command as "tone command". The command-pattern definition conforms to the actual physiological observation given by Sagart (1986) of what about the action of the circothyroid muscle (CT) and stornohyoid muscle (SH) related to the generation of Chinese tones[7]. Take the case of tone 2 as an example: first there exists a weak CT motion to raise the vocal cord tension to generate the initial part of the contour of tone 2, which is imitated by RF1 command with small magnitude in the schematic rule, and then a strong CT motion causes rapid change in the vocal cord tension to swiftly raise F0 for the second part of that, in the command-pattern, a RF2 command with large magnitude and appropriate time constant γ is used to simulate such a process. Although only four tone commands are defined, we can describe various variation of tone pattern and tone sandhi by setting appropriate command parameters.

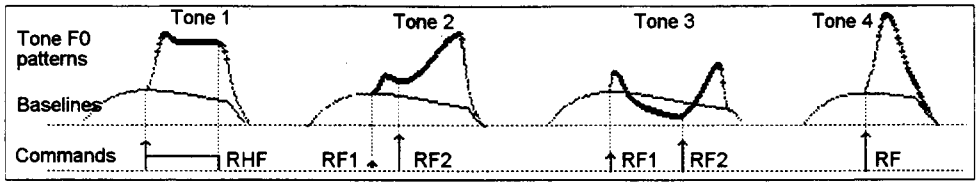

Fig.3 The block diagram of the command-pattern for Chinese tone.

On the basis of the results and the idea of Fujisaki's F0 model [2], a funcational model is constructed to express the entire generation process of a Chinese F0 contour, in which two functional model are used to simulate the control mechanism of the vocal cord vibration, one is the function mentioned above, $G_{RF}(t)$, which generates tone component with the excitation of tone command, the other is the phrase control mechanism cited from Fujisaki's F0 model [2] for generating phrase component with the excitation of phrase command to form a global baseline. On the logarithmic scale, an F0 contour is approximated by the algebra sum of both tone component and phrase component along with a asymptotic value of fundamental frequency Fb.

4. ANALYTICAL EXPERIMENTS

4.1. Speech Material

Two kinds of text were selected for the present study, text 1 was 63 monosyllables made of the voiced consonants of "m, n, l, y, w, r" and the monovowels of "a, i, e, u, v, o", text 2 was 100 sentences selected from textbook 《Chinese 300 sentences》 which was initially designed for foreigner to learn oral Chinese. These sentences cover main Chinese sentence patterns, various tones and moods, and context relationship in the spoken Chinese. Text 1 was uttered by four professional announcers (2 females, WL and FL, 2 males, ZR and GY), and text 2 was recorded by the female speaker WL. All the speech materials were recorded in read speech form of standard Chinese at a sound-isolated recording room.

4.2. Method of Analysis

The analysis of F0 contour is conducted in two steps: (1) extraction of the F0 contour of an utterance; (2) extraction of the model parameters for a given F0 contour. Particularly, F0 is converted from the fundamental period measured by detecting the maximum peak of the short-time autocorrelation function of the prediction residual obtained from a linear predictive analysis of speech with the frame length 20msec, overlapped by 10 msec intervals, and the error in extracting about F0 is corrected by hand. The model parameters are then estimated by minimized the mean squared error between the measured F0 contour and that of the model on the logarithmic scale. The minimization process uses the method of iteration with a initial estimate. For simplification, in the case of monosyllables, the same initial phrase command with magnitud 0.4 and the asymptotic value Fb (119Hz for female's and 65Hz for male's) are set to above speech materials. In the case of sentences it is more complex, generally speaking, when an utterance contains more than one breath group, or prosodic phrase, place the largest phrase command at the initial phrase, then successively decrease its magnitude for each subsequent phrase, unless required otherwise by the discourse condition, the Fb for all sentences is set equal to 119Hz.

4.3. Characteristics Of The Model

By the technique of Analysis-by-Synthesis, all the F0 contours are decomposed into phrase component(s) and tone component(s), (i.e. rise-fall patterns), and the parameters, $Tm_0, A_m, \gamma, Tm_1, \sigma, \delta$, of their underlying command are estimated. the results clearly indicate that the functional model could always reproduce a given F0 contour very well, and a good correlation exist between the model parameters and the structure of F0 pattern, the following are details.

First focus on the characteristics of the model parameters resulted from the analysis of the monosyllabic materials. The magnitudes of RF or RHF commands are rather stable and directly link to tone contours relative to the phrase components except that of RF1 in tone 2 command, which is believed to be affected by the microprosody of initial consonant. According to Eq.(1), the parameters, γ and σ, should determine the characteristics of the F0 raising and lowering transition respectively, and γ^{-1} give the exact time during which F0 goes from valley to peak. On the other hand, because the F0 contour is the output of the vocal cord vibration, which is a physical system, the F0 raising trace from valley point to peak point should depend on its duration. Just as expectation, the result in analysis of the main rising component of the F0 contours of the monosyllables with tone 2 or tone 3 clearly indicates that there exists a comparatively stirct linear relationship between the system response time γ^{-1} and the measured transition duration, D, from valley to peak, and it is independent on individual speaker. Although the F0 raising duration D varies across a certain range with the factors of speaker, toneme and prominence, the result show us a very simple rule, $\gamma = D^{-1}$, to control F0 raising transition very closely if the transition duration D is given. Also, we analyze the relationship between parameter σ and duration relative to F0 lowering, D, the result shows that σ slightly depends on D, and four quantization levels, namely 6/s, 10/s, 20/s and 30/s, can make σ work well.

Then take a look at the accuracy given by the model. The results in analyzing and synthesizing the F0 contours with the functional model indicate that the model could always approximate a given F0 contour with a high precision and the mean squared error of each utterance (a monosyllable or a sentence) is less than 0.006, for example, statistics of that of 63

monosyllables produced by WL, according to tone type, are respectively 0.001674 for tone 1, 0.001434 for tone 2, 0.000514 for tone 3 and 0.001891 for tone 4. Fig. 4 shows an example of the result of Analysis-by-Synthesis the Chinese utterance "zhe(4) feng(1)xin(4) chao(1) zhong(4) liang(3) ke(4), yao(4) tie(1) yi(1) kuai(4) si(4) mao(2) de(0) you(2) piao(4). (The letter is 2 grams overweight, (please) stick on 1.4 yuan worth of stamp)" with mean squared error 0.000896. Where, from top to bottom, it shows the waveform, the F0 contour and its decomposition, and the underlying commands, RF, RHF command and phrase command, the symbol "+" stands for the measured F0 (displayed every two points). Furthermore, to show the error distribution clearly, the statistical analysis is conducted on the error on linear fundamental frequency scale, namely, 97.7% of the point error is less than 20hz, 88.0% of one is less than 10hz, and 62.6% of one is less than 5hz, in which the match error under 10hz is mainly caused by the relevant fluctuation of the actual F0 contour. Fig.5 gives the error distribution bar for both the 63 monosyllables and the example sentence.

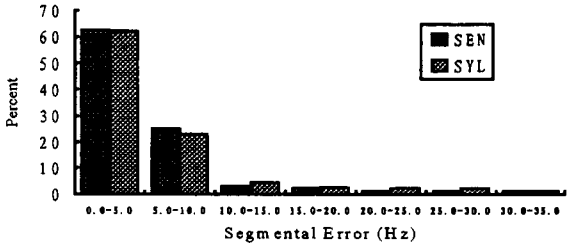

Fig.5 The distribution bars of the error of the 63 monosyllables and the example sentence.

5. CONCLUSION

A functional model for the control mechanism of generating rise-fall patterns in F0 contour is introduced. The results in analysis of speech sample of Chinese monosyllables and sentences indicate the validity of the proposed functional model for generation of local F0 contour in Chinese and summarization of prosodic rules.

6. REFERENCES

1. Fujisaki, H. & Hirose,K. "Analysis Of Voice Fundamental Frequency Contours For Declarative Sentences Of Japanese", *Journal of the Acoustical Society of Japan,(E)5,233-242,1985.*
2. Fujisaki,H."A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", *Ann.Bull.RILP, 21, 165-175,1987.*
3. Pierrehumbert,J., "Synthesizing Intonation", *JASA, 70 (4), 985-995,* Oct. 1981.
4. Fant,G. & Kruchenbery,A., "Notes On Stress And Word Accent In Swedish", *Proceeding of international symposium on prosody,19-36,*18,sept,1994.Yokohama,Japan.
5. Honda,K. "Laryngeal And Extra-Laryngeal Mechanism Of F0 Control", *Aug., 31, 1994.*
6. Ni, J.F. and Wang, R.H. "Modeling The Control Mechanism For Generating The Rise-Fall Pattern In F0 Contour", *Published in ACTA ACOUSTIC,1996.*(in Chinese)
7. Wu,Z.J. and Lin, M.C. 《Shi Yan Yu Yin Xue Gai Yao》, *High Education Press Of China,1989.*

The work is supported by National Natural Science Fund (No. 19304010).

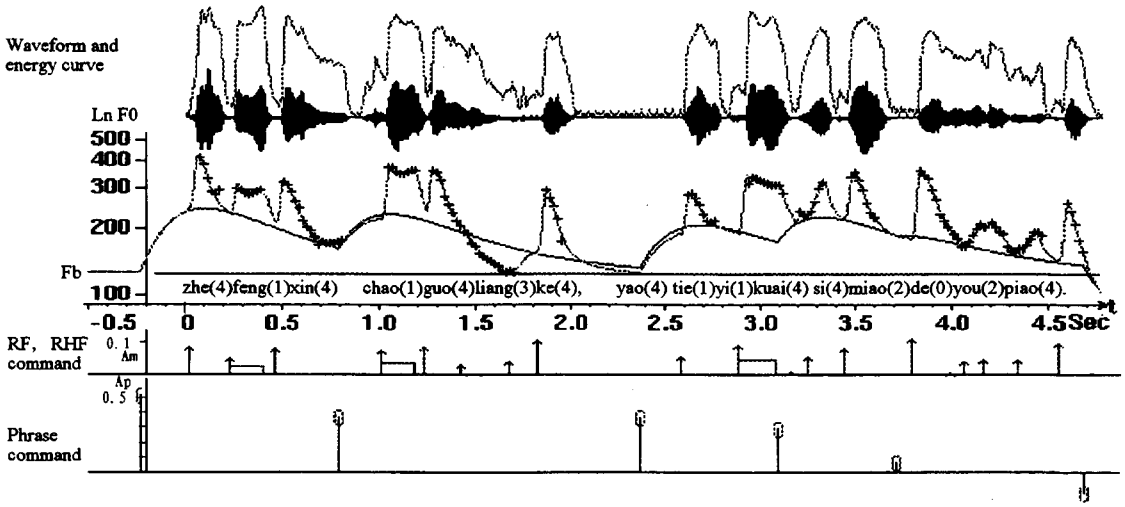

Fig.4 An example of analysis and synthesis of Chinese utterance "zhe(4) feng(1) xin(4) chao(1) zhong(4) liang(3) ke(4), yao(4) tie (1) yi(1) kuai(4) si(4) mao(2) de(0) you(2) piao(4).(The letter is 2 grams overweight, (please) stick on 1.4 yuan worth of stamp)". The figure illustrates the optimum decomposition of the F0 contour into the phrase components and rise-fall patterns and also shows the underlying commands for these components.