A Concatenative-Tone Model with its Parameters' Extraction

Gao Peng Chen, Yu Hu, Yi Jian Wu and Ren Hua Wang

iFlytek Speech Lab
University of Science and Technology of China
{gpchen;jadefox}@ustc.edu; jasonwu@mail.ustc.edu.cn rhw@ustc.edu.cn

Abstract

This paper presents a novel method to describe concatenative-tone in Mandarin with parameters of Fujisaki model. The method is based on an essential assumption that when applying Fujisaki model on Mandarin, the F0 contour mostly depends on how different tone types joint,, We can illustrate the concatenative-tone by tone command, which is a combination of accent commands. The patterns of tone concatenation can be represented by different tone commands. A set of equations are designed to predict the tone commands of natural speech with prosodic information such as tone types and word boundary types, etc as their parameters. Typical categories of tone commands for one-syllable and double-syllable words have been successfully obtained from F0 contours by solving these equations.. Approach of equation solution is also given in detail in this paper.

1. Introduction

Mandarin is a tonal language including four basic tone types: tone 1, tone 2, tone 3, tone 4 and a unique tone zero tone. The famous Chinese phonetician, Prof. Zhao Yuanren, considered in [7], "The pitch movement of Tone language is composed of three elements: 1. a syllable's tone, 2. the variety of tone in continuous utterance, 3. the pitch movement caused by mood or attitude." The last one is regarded as the intonation of the sentence. Besides, Prof. Zhao proposed in [8] that, "Tone and intonation can occur together and their superposition is the pitch movement." "Figuratively tone and intonation can be regarded as wavelet and sea. The sea is waving on which there are wavelets, so tone and intonation are paratactic." Tone shape and range effected by mood would be flexible like attaching to shirr.

Fujisaki model is originally designed for Japanese. The fundamental idea is that the pitch movement is made up of two components, accent component and phrase component, which respectively represent the pitch accent and the declination of phrase contour (Figure 1). The illustrating expressions are given in the following.

$$\log_e F0(t) = \log_e Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} Aa_j [Ga(t - T_{1j}) - Ga(t - T_{2j})] \quad (1)$$

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases} \quad (2)$$

$$Ga(t) = \begin{cases} \min\{1 - (1+\beta)\exp(-\beta t), \gamma\}, & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases} \quad (3)$$

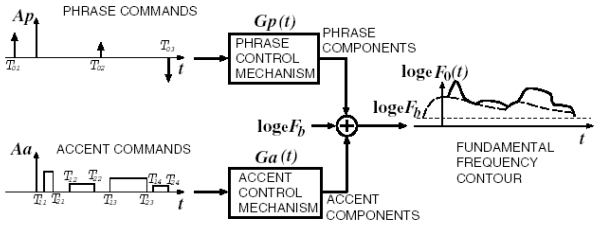


Figure 1: *Fujisaki model illustration.*

Fujisaki Model has been adapted for several languages including Korean, Greek, Spanish, and Polish etc. Prof. Mixdorff successfully applied Fujisaki model on German [2]. Although Fujisaki model is not adequate for tonal languages, its main idea concords with that of Prof. Zhao on Mandarin pitch structure. Some research has been done to quantify parameters on Mandarin [4]. Through experiments we found that parameters of Fujisaki model could somehow approximate the pitch movement of Mandarin tone. It proves to be another way to study the Mandarin tone. This paper presents how we improve the performance of Fujisaki Model on Mandarin.

2. Patterns of tones on Mandarin

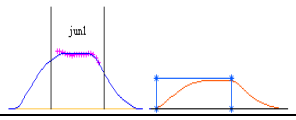
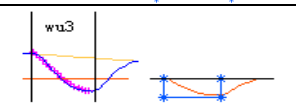
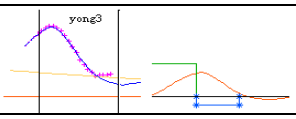
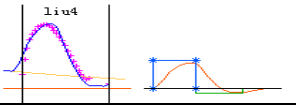
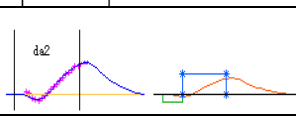
Based on the previous research we find that in Mandarin words with the same concatenative-tone have the similar pitch movement. As words of one or two syllables are in a great measure in Mandarin, we focus our research on those word units. Tone command, which is defined to be the combination of several accent commands, is proposed in the Fujisaki model to approximate the four basic tone types or the comcatenative-tone patterns, thus the observed F0 contours of utterances are the integration of actions of intonation and tone types. Therefore the contour is composed of two parts: phrase component and tone component. In practice we use phrase commands to represent phrase component and tone commands to represent tone component. The phrase component is the indication of the pitch declination in utterances. Our research centers on the tone component which lies on the phrase component.

2.1. Obtain the patterns by clustering

The contour of one-syllable word can be parameterized and illustrated as in Table 1. Most of tone 3 syllables pertain to low pitch pattern while few to high pitch pattern. The high pitch pattern begins with a high pitch and looks like the contour of tone 4 syllable. The F0 of fully uttered tone 3 syllables falls and rises immediately, but in spontaneous speech the next syllable comes directly after the falling part and supersedes the rising part. Double-syllable words will form stable concatenative-tone patterns because the pitch

pattern of two syllable words are connected tightly and interacts each other. Part of the concatenative-tone patterns are listed in Table 2. These patterns are obtained by F0 contour clustering, command fixing and pattern grouping. Firstly, cluster the pitch contours of all the two syllable words in our speech corpus. Here we minimize the variance of each category to less than 6Hz, so that all samples in a category can be represented by the cluster center sample. Secondly, fix the tone command types with reference to one syllable patterns. Thirdly, combine those cluster centers which are identical in their command structure. Finally we come across dozens of patterns and part of which are listed in Table 2. Due to the internal relations between these patterns and the concatenative-tone of two syllable words, we call them tone patterns. Experiments show that these tone patterns are representative of the F0 contours of two word syllables. These patterns are parameterized to simulate and rearrange the pitch movement.

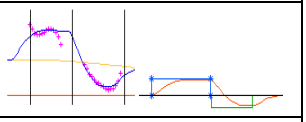
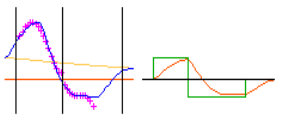
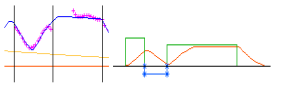
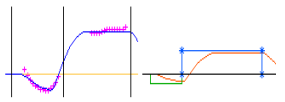
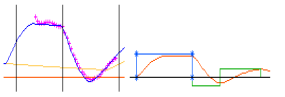
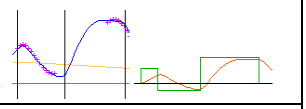
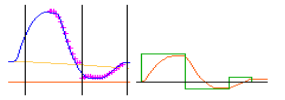
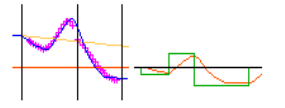
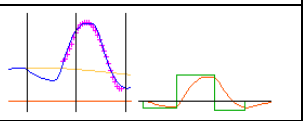
Table 1: *One-syllable tone patterns*

Tone Commands		Tone	Illustration
1 Acc	+	1	
	−	3 (low)	
2 Acc	+−	3 (high)	
	+−	4	
	−+	2	

2.2. Validation of the patterns

To describe the tone pitch with tone commands is to predefine a settled template shape for each tone and concatenative-tone., and adjust the parameters to change the template shape in order to fit the original syllable's pitch. Large numbers of experiments show that most of words' pitch contours can be approached by the tone patterns. Due to the complexity and variability of natural language and speech, the pitch contours of some words don't agree with these tone patterns. We have tried to use the common tone patterns (as listed in Table 1) to imitate the pitch of these disobedient words and resynthesize by PSOLA. The synthesized wave sounds natural and is acceptable. So it is confirmed that the tone patterns are also applicable to them. Here we have ignored the micro-prosody because it is beyond the describing ability of Fujisaki model and it imposes little influence on F0 contour.

Table 2: *Concatenative-tone Pattern examples.*

Tone commands		Concatenative-Tone	Illustration
2 Acc	+−	13	
		43	
	−+	21	
		31	
3 Acc	+−+	12	
		41	
		42	
	−+−	23	
		24	

3. Solution of Parameters

The primary analysis is based on the assumption that the pitch contour of a sentence is the integration of actions of tone types and intonation, that is, the pitch contour of syllables can be discomposed into two parts: phrase component and tone component, while it mainly depends on their tone types. There is a weak interaction from neighboring syllables. The tone array of a sentence determines the general contour. Therefore, we can split the F0 of a sentence and extract the parameters as the following.

3.1. Estimation and Regulation of F0

Calculate the F0 contour from wave file and correct the error. Normalize the F0 contour of each syllable by the patterns illustrated in Table 1, such as tone 1 pattern is high flat, tone 2 riseing, tone 3 low or high falling, tone 4 falling, etc. The RMSE between normalized and original F0 is 4.2Hz. When resynthesized by PSOLA, they sound natural and lossless. Then interpolate the unvoiced portion with quadratic spline, smooth the continuous contour with a median filter, sample

it per 10ms. And finally calculate the natural logarithm of each point to get a log F0.

3.2. Solution of phrase commands

Input the continuous log F0 into a high pass filter. The output is approximate tone component. The filtered part corresponds to the phrase component and Fb. Subtract Fb (use the minimum or previously assigned value) from it then get the phrase component. Calculate its derivative and extremums. Draw out all the sections of declination between two poles. A broad declination corresponds to a phrase component. T_{0i} and Ap_i are obtained through nonlinear regression applied on equation (2) (α =2 is fixed). Calibrate the phrase commands with prosodic boundary information acquired from text parser. The distance of two phrase commands is required to be longer than 1.2s. Almost each phrase boundary corresponds to a phrase command. There is only one phrase command within a phrase word.

3.3. Solution of tone commands

Compute the following expression:

$$\log_e F0(t) - \log_e Fb - \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) \qquad (4)$$

The result is the continuous tone components. Match the contour to the concatenative-tone patterns word by word to solve the tone commands. Each tone command is composed of a series of positive or negative accent commands. With respect to each accent command, a set of parameters $Aa, T_1, T_2, \beta, \gamma$ needs to be solved. γ is normally fixed as 0.9. T_1 , T_2 depend on the time of inflexion and voiced segmentation. The inflexion can be easily determined through those tone pitch patterns described above. β denotes the gradient of rise or fall. It's assigned a value of 30 for a tone 1 syllable in the front of a word, and 20 otherwise. Thus the one-accent tone command is solved. As to the multi-accent tone commands, an accent command only acts on the next one and is "＋－" or "－＋". For the "＋－" pattern (Figure 2), the curve of the two parts can be expressed by equation (5) and (6). The aim is to solve the group of nonlinear equations to get $Aa_1, \beta_1, Aa_2, \beta_2$. β_1 and β_2 are initialized as above-mentioned value. The temporal position of each command is fixed by the time of inflexion,, voice beginning and voice end. Solve the equation (5) first and its root, that is the value of ?? is applied to equation (6). Therefore all the parameters would be calculated. The "－＋" pattern is similar to "＋－". The multi-accent pattern is considered as the combination of several "－＋" and "＋－". Calculate the parameters one by one from the first accent.

$$f_1(t) = Aa_1 Ga(t - T_b) \qquad (5)$$

$$f_2(t) = Aa_1[Ga(t - T_b) - Ga(t - T_m)] - Aa_2 Ga(t - T_e) \qquad (6)$$

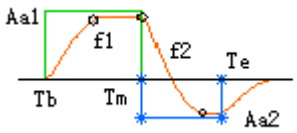


Figure 2: *Tone command with positive and negative extremums.*

4. Modification of Parameters

4.1. Reasons causing errors

The new pitch contours reconstructed by the extracted parameters of commands are close to the original contours. though there is still some discrepancy between them. We modify these parameters for the sake of better results. The main reasons that cause errors are:

(a) The interaction between words is neglected as to separately extract parameters of each word;
(b) Solution of the equations are based on an approximate hypothesis;
(c) The real pitch contour does not match the tone patterns sometimes.

The errors caused by (a) and (b) can be corrected by mathematical methods. The errors caused by (c) are acceptable and permissible under our hypothesis. It is validated in section 2.2. So the RMSE of the reconstructed pitch contour is only required to be within a certain range.

4.2. Parameters adjustment to fit the contour

We use the back propagation algorithm to correct the parameters. First scan and check all the phrase commands and tone commands. Delete the redundant phrase commands and add missing phrase commands near the syntax boundary. Find the tone commands that overlap each other, then combine wide commands and delete the narrow. For each tone command, adjust the parameters in a small range supervised by the feedback error. Where affects the curve's range, the gradient depends on β. Since we have considered that the tone patterns are reasonable, the onset and offset time, and of tone commands keep changeless. Do the modification repeatedly until the feedback error is unchanged or less than the threshold value. Finally we can get a contour very close to the original contour (e.g. Figure 3).

The whole process of extraction is done for a corpus with 14000 sentences. The statistical result shows that the contours rebuilt have a high mean correlation of 0.94 with the original contours.

4.3. Check and selection for Prediction

The processing is successful and used to analyze our corpus database. However there exist some unpredictable exceptions because of the data with noises. So after the processing is done, a program will check the match degree of the original and the reconstructed contours. First the sentences whose RMSEs are less than 10Hz and correlations are more than 0.85 are picked out and labelled as 'R'. Secondly, the words having less RMSE than 10Hz are selected from the 'R' sentences and labelled as 'RR'. The 'R' sentences are about 85% of all the sentences. The

'RR' words are about 82% of all the words. The sentences or words unlabelled should be corrected manually before used.

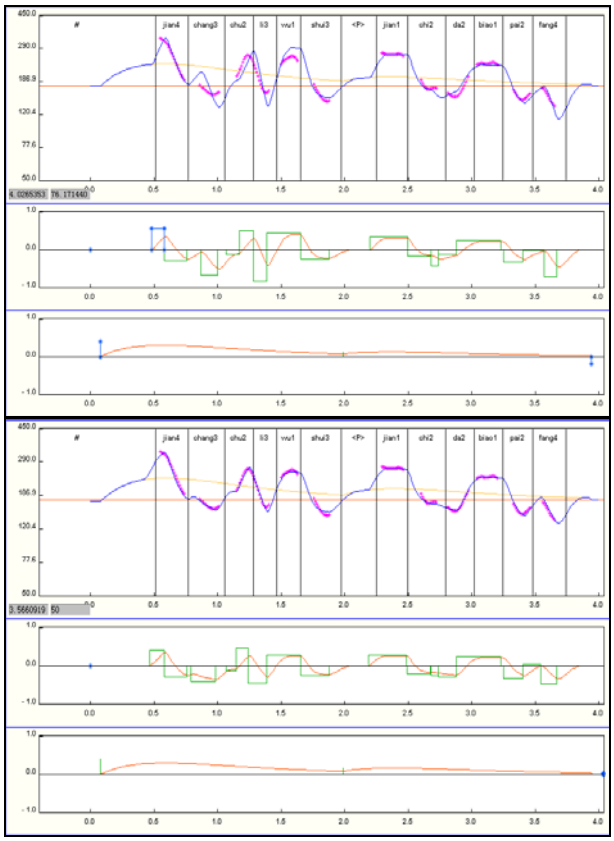


Figure 3: *The upper is after solving the equation. The lower is after correcting.*

5. Conclusions

In conclusion the concatenative-tone patterns can represent the actual words' pitch contours and hold their characteristics when approaching and reconstructing. It is convenient to rise, fall or rotate the pitch by changing the parameters of tone commands so that it is much helpful to study the rules that pitch contours vary with the circumstance. We can design the three-syllable words' patterns, and they would be processed as same as the double-syllable words. Zero tone has not steady pattern, it is smoothed by the fore-and-aft contours. Because of using the information such as tone type, word segmentation, and phrase boundary to extract parameters, the prediction of model will be possible and straightforward in the aftertime. The prediction will be implemented hierarchically. First intonation phrase boundary got from text parser will predict the phrase command. Then use tone and lexical information to predict tone patterns word by word on the phase. Later we'll go on with this work to study the variety of commands in different context and realize the pitch prediction in our TTS system.

6. References

[1] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoustic. Soc. Jpn (E), vol. 5, no. 4, pp. 233–242, 1984.

[2] H. Mixdorff, "Intonation Patterns of German Model-based quantitative Analysis and Synthesis of F0 contours." Ph.D. thesis.

[3] H. Mixdorff, "A novel approach to the fully automatic extraction of fujisaki model parameters," Proc. ICASSP 2000. Istanbul, vol. 3, pp. 1281–1284 (2000).

[4] Jin-Fu Ni, Ren-Hua Wang, "Modeling the Control Mechanism for Generating the Rise-Fall Pattern in F0 Contour". ACTA ACOUSTIC,vol 21,No.6,1996.

[5] Jin-Fu Ni, Ren-Hua Wang and Keikichi Hirose, "Quantitative analysis and formulation of tone concatenation in Chinese F0 contours," Proceedings 5th European Conference on Speech Communication and Technology, Rhodes, Vol.1, MAB.6, pp.195-198 (1997-9).

[6] Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose and Hiroya Fujisaki, "Automatic Extraction Of Model Parameters From Fundamental Frequency Contours Of English Utterances", ICSLP 2002.

[7] Yuan Ren Zhao, "The Tone and Intonation of Chinese", 1992.

[8] Yuan Ren Zhao, "Problems of Language", Commercial Press of China, 1980.