A Method for Automatic Tone Command Parameter Extraction for the Model of F_0 Contour Generation for Mandarin

Wentao Gu[1, 2]*, Keikichi Hirose*[1] *and Hiroya Fujisaki*[1]

[1]University of Tokyo, Japan
[2]Shanghai Jiaotong University, China
{wtgu; hirose}@gavo.t.u-tokyo.ac.jp fujisaki@alum.mit.edu

Abstract

The model for the process of F_0 contour generation, first proposed by Fujisaki and his coworkers, has been successfully applied to Mandarin, which is a typical tone language with a distinct feature that both positive and negative tone commands are required. However, the inverse problem, *viz.*, automatic derivation of the model parameters from an observed F_0 contour, is more difficult for Mandarin than for those non-tone languages, because the polarity of tone commands cannot be inferred directly from the F_0 contour itself. In this paper, an efficient method is proposed to solve the problem by using the information on syllable timing and tone labels. With the same framework as that proposed for Japanese and English, the method presented here for Mandarin is focused on the first-order estimation of tone command parameters. A set of intra-syllable and inter-syllable rules are constructed to recognize the tone command patterns within each syllable. The experiment shows that the method works effectively and gives results comparable to those obtained by manual analysis.

1. Introduction

In text-to-speech synthesis, accurate modeling of F_0 contours is critically important for the intelligibility and naturalness of synthetic speech. The model for the process of F_0 contour generation (henceforth the F_0 contour model), first proposed by Fujisaki and his coworkers [1], has been successfully applied to many languages including Mandarin, which is a typical tone language [2, 3]. The model is based on the formulation of the underlying physiological and physical mechanisms, which distinguishes it from all other approaches for intonation modeling. It can generate very close approximations to observed F_0 contours from a relatively small number of linguistically meaningful parameters.

While it is straightforward to generate an F_0 contour from a set of model parameters, the inverse problem, *viz.*, derivation of command parameters from a given F_0 contour, cannot be solved analytically. This inverse problem, however, is of great importance both for prosody analysis and for construction of prosody corpora for speech synthesis purpose. A successive approximation method with multi-stage first-order estimation has been proposed for both Japanese and English [4, 5]. With a reliable first-order estimation, the method ensures convergence to the globally optimum solution.

In the previous work [6] we have presented a modified method for Mandarin, with the focus on the first-order estimation of tone commands. The major problem for Mandarin is how to deal with both positive and negative tone commands, which have been shown to be necessary for modeling the F_0 contours of Mandarin [2]. In this paper we will present a more updated method, with a revised set of rules and the introduction of back-tracing correction process.

2. The F_0 contour model for Mandarin

The F_0 contour model is a command-response model that describes F_0 contours in the logarithmic scale as the sum of phrase components, accent components (or tone components for tone languages) and a baseline level $\ln F_b$. The model diagram for Mandarin [2] is shown in Fig. 1, where the phrase commands (impulses) produce phrase components through the phrase control mechanism, giving the global shape of the F_0 contour, while the tone commands (pedestals) generate tone components through the tone control mechanism, characterizing the local F_0 changes. Both mechanisms are assumed to be critically-damped second-order linear systems.

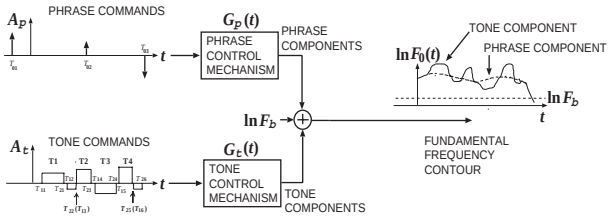


Figure 1: *The F_0 contour model for Mandarin.*

The model consists of the following parameters: A_{pi} and T_{0i} denote the magnitude and time of the ith phrase command respectively, while A_{tj}, T_{1j} and T_{2j} denote the amplitude, onset time and offset time of the jth tone command respectively. The constants α, β and γ are set at their respective default values 3.0 (1/s), 20.0 (1/s) and 0.9 respectively in the current study.

Unlike most non-tone languages, Mandarin requires both positive and negative tone commands. In Mandarin there are four lexical tones and a neutral tone: T1 (high tone), T2 (rising tone), T3 (low tone), T4 (falling tone) and T0 (neutral tone). These tones are attached to each syllable. As shown in Figure 1, T1 to T4 are assumed to correspond to their respective *tone command patterns* (*intrinsic patterns*): T1 (positive), T2 (negative followed by positive), T3 (negative) and T4 (positive followed by negative). For T2 and T4, the offset of the 1st tone command is assumed to coincide with the onset of the 2nd tone command. The command pattern for T0 is assumed to depend on the context and usually have reduced amplitudes.

3. Parameter extraction method

As mentioned above, the model parameters can only be derived by successive approximation (*i.e.*, Analysis-by-Synthesis) with a good initial estimation. In [4], a multi-stage approximation method was proposed to derive the first-order

estimation by approximating an observed F_0 contour with a set of piecewise 3rd-order polynomials which are continuous and differentiable everywhere. The method was also successfully extended to English in [5]. For Mandarin, we use the same framework, but modify the second stage, namely first-order estimation of tone commands (instead of accent commands).

The framework of the automatic parameter extraction method is as follows [4, 6], within which we will focus our discussion on the first-order estimation of tone commands:

(1) Pre-processing of an observed F_0 contour.
(2) First-order estimation of tone command parameters.

If we neglect the effect of phrase components which is much more gradual than that of tone components, the *positive maxima* (*p*-maxima) and *negative minima* (*n*-minima) of the first derivative (hence both are *inflection points*) of the smoothed F_0 contour should correspond to the onsets and offsets of tone commands with a constant delay of $1/\beta$.

(3) First-order estimation of phrase command parameters by a left-to-right successive detection from the residual contour.
(4) Optimization of parameters by Analysis-by-Synthesis.

4. First-order estimation of tone command parameters

4.1. Problem analysis

Unlike Japanese and English, Mandarin involves both positive and negative tone commands, which makes correct detection of tone commands much more difficult, because the polarity of tone commands cannot be inferred directly from the F_0 contour itself.

The task is difficult for Mandarin for two reasons: (1) there is no way to determine from the F_0 contour itself whether a *p*-maximum (or *n*-minimum) of the first derivative corresponds to the onset of a positive (or negative) command or the offset of a negative (or positive) command or both, (2) The *p*-maxima and *n*-minima of the first derivative do not necessarily occur in pairs.

To overcome this difficulty, it is necessary to use some additional information. Here we utilize syllable timing and tone labels available in Mandarin speech corpus, by which the tone command pattern within each syllable can be recognized on the basis of a certain set of rules. After tone command pattern recognition, it becomes straightforward to obtain the first-order estimation of tone command parameters.

4.2. Tone command pattern recognition

As assumed in the F_0 contour model for Mandarin, an intrinsic tone command pattern should be associated with the syllable of each lexical tone. In practice, however, the command pattern within a syllable undergoes many variations as discussed in [6]. Therefore, a set of rules needs to be constructed for tone command pattern recognition.

The basic requirement is that a series of tone commands can be generated from the inflection points of the F_0 contour. As stated above, each inflection point corresponds to either onset or offset of a tone command as shown in Figure 1, or in other words, to either ascent or descent edge of a pedestal. The starting and ending points for each edge of pedestals can be reasonably defined in three levels: 1 (positive), -1 (negative) and 0 (baseline). Therefore the recognition task is to decide the starting and ending levels of the pedestal edge corresponding to each inflection point according to the polarity of derivative. For example, for a *p*-maximum, we need to find out whether it is an onset of a positive command (0→1) or an offset of a negative command (-1→0) or both (-1→1).

4.2.1. *Intra-syllable command pattern recognition*

Firstly, intra-syllable tone command pattern recognition is carried out, based on the criterion of best match with the intrinsic command patterns, by the aid of some heuristics. Figure 2 shows some most frequently (*not* all) observed patterns associated with different series of inflection points. Each column in the figure depicts the command patterns associated with four different tonal syllables when a specific series of inflection points (+: *p*-maximum, −: *n*-minimum) are observed within the syllable. In the figure, the thin solid lines depict the recognized onsets or offsets of tone commands, while the dotted lines present a reference of the intrinsic tone command patterns. Patterns recognized as identical to the intrinsic tone command patterns are depicted by the red lines.

It is observed that some cases are associated with multiple candidates. For such cases, the command patterns are determined on the basis of timing [6].

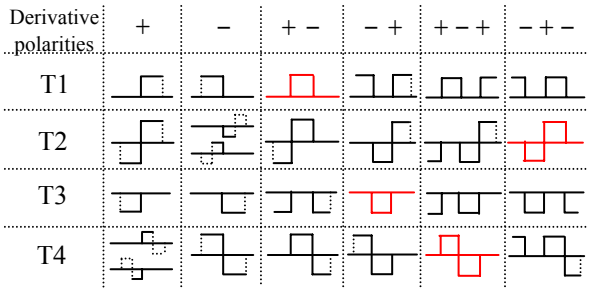


Figure 2: *Tone command patterns associated with different inflection points for each tonal syllable.*

4.2.2. *Inter-syllable command pattern refinement*

After intra-syllable tone command pattern recognition, the onset and offset of each tone command are still not necessarily uniquely determined. First, the ending level of the last inflection point in the current syllable may not necessarily match the starting level of the first inflection point in the next syllable. Second, the onset of utterance-initial tone command or the offset of utterance-final tone command may not be observed from the F_0 contour. Third, no command patterns have been specified for T0 syllables (within which both levels of all inflection points are set temporarily as 0).

Therefore, we apply the following set of inter-syllable rules to further refine the tone command patterns:

(1) Process for utterance-initial/final tone commands

(1.a) If the utterance-initial syllable is T1 within which no *p*-maximum is detected, a *dummy point* is added as the onset of positive tone command ahead of the initial T1 syllable.

(1.b) If the starting level of utterance-initial inflection point or the ending level of utterance-final inflection point is not 0, a *dummy point* should be added (with a constant time shift) before the initial one or after the final one as the onset of initial tone command or the offset of final tone command.

(2) Connect the ending level of the last inflection point in the current syllable and the starting level of the first inflection point following the current syllable.

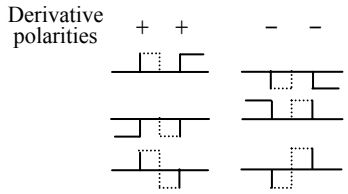


Figure 3: *Insertion of dummy points between two inflection points of the same polarity.*

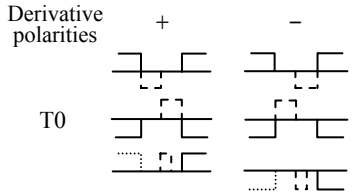


Figure 4: *Some cases for specifying tone command patterns in T0 syllables.*

(2.a) If both levels are 0 and the two inflection points are opposite in polarity, when there is another syllable of T1 (or T3) between them (no inflection points detected within this syllable), change both levels into 1 (or -1). This process is introduced to avoid missing the tone commands for syllables within which no inflection points are detected.

(2.b) If only one level is 0 and the two inflection points are opposite in polarity, simply change the level 0 to coincide with the other.

(2.c) Otherwise, if the two levels differ and the two inflection points are of the same polarity, as illustrated in Fig. 3, a *dummy point* needs to be inserted (as depicted by the dotted lines) in the middle of the two inflection points. This process is adopted to recover the offset of the current tone command (as shown in the 1st row) or the onset of the next tone command (the 2nd row) or both (the 3rd row).

(3) Determine the tone command patterns in T0 syllables.

If both the starting and ending levels of an inflection point remain 0, a set of heuristic rules will be used to determine the levels based on the context. For instance, as shown in the first two rows of patterns in Fig. 4, when the preceding and following tone commands are of the same polarity (depicted by the solid lines), the current inflection point (as indicated in the middle of the two tone commands) will be connected with the neighbors to compose a new tone command of the opposite polarity, as depicted by the dashed line.

However, there is an exception as shown in the bottom row of patterns in Fig. 4, where the offset of the preceding tone command is not an inflection point but a dummy point as depicted by the dotted lines. In such a case, a tone command will be composed with the current inflection point as the onset, while a *dummy point* will be inserted afterwards as the offset.

The purpose of introducing these specific processes in the above two cases is to produce reasonable and natural connections between inflection points.

4.3. Estimation of timing/amplitudes of tone commands

4.3.1. Basic formulation

After command pattern recognition based on inflection points, a series of tone commands need to be generated. The onset and offset time of these tone commands can be easily determined, with only a constant $1/\beta$ ahead of the corresponding inflection points.

The amplitudes of tone commands can be derived from the derivatives of $\ln F_0(t)$ at the inflection points. Since for Mandarin the *p*-maxima and *n*-minima of the derivative do not necessarily occur in pairs, the amplitudes of tone commands should be determined in connection with their neighbors. In the proposed method, we set the amplitude of the current tone command A_{tj} as follows:

(a) If T_{1j} corresponds to an inflection point,

$$A_{tj} = \begin{cases} F_0'(T_{1j}) \cdot e/\beta, & \text{if } s(T_{1j}) = 0 \text{ and } e(T_{2j}) \neq 0, \\ [F_0'(T_{1j}) - F_0'(T_{2j})] \cdot e/2\beta, & \text{if } s(T_{1j}) = 0 \text{ and } e(T_{2j}) = 0, \\ F_0'(T_{1j}) \cdot e/\beta + A_{t,j-1}, & \text{if } s(T_{1j}) \neq 0, \end{cases} \quad (1)$$

$$A_{tj} = \text{sgn}(F_0'(T_{1j})) \cdot A_{t\min}, \quad \text{if } \text{sgn}(A_{tj}) \neq \text{sgn}(F_0'(T_{1j})), \quad (2)$$

(b) If T_{1j} corresponds to an inserted dummy point,

$$A_{tj} = \begin{cases} -F_0'(T_{2j}) \cdot e/\beta, & \text{if } e(T_{2j}) = 0, \\ -F_0'(T_{2j}) \cdot e/2\beta, & \text{if } e(T_{2j}) \neq 0. \end{cases} \quad (6)$$

Here $F_0'(t)$ denotes the first derivative of $\ln F_0(t)$, while $s(t)$ and $e(t)$ denote the starting and ending level of the inflection point at t respectively. $A_{t\min}$ represents a threshold value 0.08 set as the minimum absolute amplitude of tone commands, and sgn(•) represents the sign function.

The equation (5) is introduced because the derivatives at a consecutive set of inflection points will not sum up exactly to 0 due to various noises. In the case where an obvious error occurs, *viz.*, the polarity of tone command reverses (only possible when $s(T_{1j}) \neq 0$), the correction is adopted, and a back-tracing process is conducted to re-examine other alternatives for the preceding tone command patterns.

4.3.2. Back-tracing correction

An underlying assumption in the above processes is that each tone command consists of both an onset starting from 0 and an offset ending to 0, which correspond to inflection points of the opposite polarity respectively. In other words, it is assumed that two consecutive inflection points of the same polarity cannot delimit a tone command. That is why an extra dummy point was inserted there as indicated in Fig. 3.

However, there are some cases where the onset and offset of a tone command can correspond to inflection points of the same polarity. This happens when the offset of one tone command overlaps with the onset of the next tone command of the same polarity. Fig. 5 shows all the four possible cases, among which we find that the 1st and 3rd cases are frequently related to estimation errors and hence need correction.

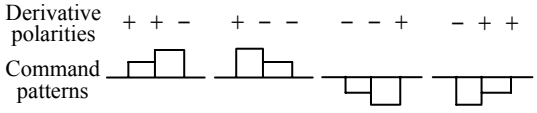


Figure 5: *Connection of two consecutive tone commands of the same polarity.*

In these two cases, the process introduced in Fig. 3 is not appropriate, which will apparently cause the succeeding tone components to shift up or shift down. This is highly responsible for the above mentioned polarity reversion error. Therefore the following process is used to correct the errors.

(1) When the polarity reversion errors occur consecutively or when the amplitude ratios between neighboring positive

and negative commands exceed a threshold consecutively, the extracted tone commands are considered to be unnaturally biased. In such a case, we trace back to search the nearest dummy point in the preceding vicinity.

(2) If a dummy point is found whose context may match the 1st or 3rd category shown in Fig. 5, restore the tone command pattern as it should be, and re-estimate the amplitudes of tone commands from there on.

(3) If a dummy point not corresponding to these two categories is found, adjust the amplitude of the tone command composed of the dummy point to eliminate the bias, and do re-estimation from there on.

(4) Otherwise, if no dummy point is found nearby, adjust the amplitudes of the current tone commands directly.

5. Experiment

The speech material used in the current study is the same as that used by Wang et al. in [3], which consists of 80 utterances read by two male native Mandarin speakers. The F_0 values are extracted by a modified autocorrelation analysis of the LPC residual signal. In the current experiment the initial value of the baseline F_b is set at 80 Hz.

The performance of the first-order estimation of tone command parameters is evaluated by comparison with the results of manual extraction given in [3], and is expressed in terms of miss and false alarm rates [5].

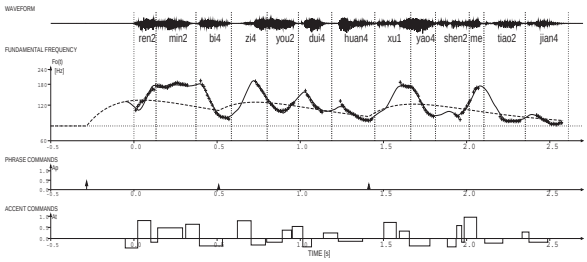
An automatically extracted tone command is regarded to be correct when both the onset and offset are within 0.15s distance from the corresponding manually extracted ones. Besides, when consecutive tone commands of the same polarity in the result of manual extraction are found to be merged in the result of automatic extraction, it is not considered as an error since such a merger reflects the effect of tonal coarticulation.

In the current experiment, the miss and false alarm rates for tone commands are 8.0% and 13.1% respectively. Considering that this is the result before Analysis-of-Synthesis, it is quite promising, since successive approximation is expected to further reduce the errors.

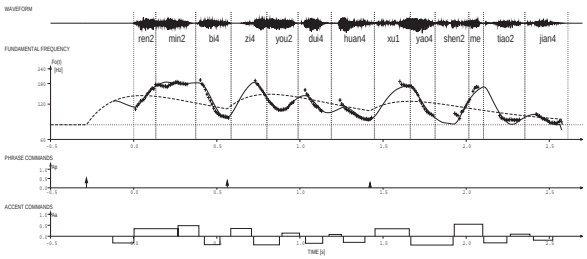
Although the current study is focused on the first-order estimation of tone commands while the first-order phrase commands estimation has not been optimized, the preliminary approximation still shows good performance. The average approximation error, defined as the average of root mean square (r.m.s.) error between observed and approximated $\ln F_0$ within voiced intervals for all the utterances, is equal to 0.071. This is equivalent to approximately 7.4% for the r.m.s. value of the relative error in F_0.

An example result for Mandarin utterance is shown in Fig. 6, where Fig. 6(a) gives the manual extraction result, while Fig. 6(b) gives the automatic extraction result after Analysis-by-Synthesis. The crossed symbols depict the observed F_0 values, while the solid lines and dashed lines depict the approximated F_0 contours and the contribution of phrase components respectively.

It is shown that the automatically approximated F_0 matches the observed F_0 quite well, though slightly inferior to the manual approximation. It can be noted in Fig. 6(b) that the descent offset of the 2nd tone command overlaps with the ascent onset of the 3rd tone command but gives good approximation, which confirms the validity of the back-tracing correction process.



(a) Manual extraction result.



(b) Automatic extraction result.

Figure 6: *An example of the F_0 contour model parameter extraction for a Mandarin utterance.*

6. Conclusions

A method was proposed for automatic extraction of tone command parameters of the F_0 contour model for Mandarin. Adopting the same framework as that developed for Japanese and English, the current study was focused on the first-order estimation of tone commands having both positive and negative polarities. The information on syllable timing and tone labels was employed, and a set of rules was constructed to help recognize the tone command patterns. The results of preliminary experiments showed that the method worked effectively. The proposed method can also be easily generalized to all other tone languages, with a language-specific set of rules for tone command pattern recognition.

7. References

[1] Fujisaki H.; Nagashima, S., 1969. A model for synthesis of pitch contours of connected speech. *Annual Report, Engg. Res. Inst., University of Tokyo*, 28, 53-60.

[2] Fujisaki, H.; Hallé, P.; Lei, H., 1987. Application of F_0 contour command-response model to Chinese tones. *Reports of Autumn Meeting, Acoust. Soc. Jpn*. 197-198.

[3] Wang, C.; Fujisaki, H.; et al., 1999. Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command-response model. *Proc. Eurospeech'99*. Budapest: Hungary, 1655-1658.

[4] Narusawa, S.; Minematsu, N.; Hirose, K.; Fujisaki, H., 2002. A method for automatic extraction of parameters of the fundamental frequency contour generation model. *J. Inf. Process. Soc. Jpn* 43(7), 2155-2168.

[5] Narusawa, S.; Minematsu, N.; Hirose, K.; Fujisaki, H., 2002. Automatic extraction of model parameters from fundamental frequency contours of English utterances. *Proc. ICSLP'02*. Denver: U.S.A., 1725-1728.

[6] Gu, W.; Hirose, K.; Fujisaki, H., 2003. A method for automatic extraction of F_0 contour generation process model parameters for Mandarin. *Proc. ASRU'03*. St. Thomas: U. S. Virgin Islands.