A General Approach for Automatic Extraction of Tone Commands in the Command-Response Model for Tone Languages

Wentao Gu, Keikichi Hirose and Hiroya Fujisaki

University of Tokyo, Japan
{wtgu; hirose}@gavo.t.u-tokyo.ac.jp fujisaki@alum.mit.edu

Abstract

Although the command-response model for the process of F_0 contour generation has been successfully applied to many languages, the inverse problem, *viz.*, automatic derivation of the model parameters from an observed F_0 contour, is more challenging, especially for tone languages which have tone commands of both polarities. Since the polarities of tone commands cannot be inferred directly from the F_0 contour itself, the information on tone identity and timing need to be incorporated. The current study gives a general approach for the first-order estimation of tone command parameters for tone languages, taking Mandarin and Cantonese as two examples. After a rule-based recognition of the tone command patterns within each syllable, the timing and amplitude of tone commands will be deduced. The experiments show that the method gives good results of analysis for both the two dialects.

1. Introduction

A command-response model for the process of F_0 contour generation has been successfully applied to many languages including tone languages such as Mandarin [1] and Cantonese [2, 3]. Based on the formulation of the underlying physiological and physical mechanisms, the model can generate very close approximations to observed F_0 contours from a small number of linguistically meaningful parameters.

While it is straightforward to generate an F_0 contour from a set of commands, the inverse problem, *viz.*, derivation of command parameters from a given F_0 contour, cannot be solved analytically. In most previous studies, the commands are derived by the method of Analysis-by-Synthesis, *viz.*, first estimated manually by an expert and then optimized by a hill-climbing procedure of successive approximation. Since it is well known that the hill-climbing procedure can be trapped into local minima, a good initial estimation is critically important. In order to make the model practically useful in prosody modeling and speech synthesis, a fully automatic solution for the inverse problem is necessary.

Although several approaches have been proposed for the inverse problem for non-tone languages, very few works have been reported on tone languages, because it is more difficult to estimate tone commands with both positive and negative polarities in tone languages than those of accent commands with only positive polarity in most non-tone languages.

A successive approximation method with multi-stage first-order estimation has been proposed for Japanese with certain success [4]. The first-order estimation is done first on accent commands and then on phrase commands. Based on the same framework, we have proposed a modified method for Mandarin [5, 6] and recently for Cantonese [7], focusing on the first-order estimation of tone commands. In this study we will give a general discussion on the approach for any tone languages by taking Mandarin and Cantonese as two examples.

2. The F_0 contour model for tone languages

A command-response model describes F_0 contours in the logarithmic scale as the sum of phrase components, accent/tone components and a baseline level $\ln F_b$. 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.

The model can be formulated by the following equations:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{tj} \{G_t(t - T_{1j}) - G_t(t - T_{2j})\}, \quad (1)$$

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

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

The values of α, β and γ can be considered to be practically constant, and in the current study they are fixed at 3.0 (1/s), 20.0 (1/s) and 0.9 respectively based on our previous works. The following parameters in the model need to be estimated: A_{pi} and T_{0i} indicate the magnitude and time of the ith phrase command respectively, while A_{tj}, T_{1j} and T_{2j} indicate the amplitude, onset time and offset time of the jth tone command respectively, and the baseline frequency F_b is more or less speaker-specific but may vary from an utterance to another.

Mandarin and Cantonese, as the two most well-known dialects of Chinese, have four and nine lexical tones respectively, as listed in Table 1. It should be noted that the tone numbers in the two dialects are not corresponding, though the tones showing similar features may have similar historic origins. Mandarin does not have entering tones, whereas in Cantonese the syllables of entering tones (T7~T9) end with an unreleased stop coda /p/, /t/ or /k/, and each entering tone has its counterpart of non-entering tone, showing a similar F_0 pattern (hence the nine tones can be merged into six).

Especially, the two dialects show opposite characteristics of tone neutralization and tone *sandhi*. Mandarin has a neutral tone (T0), which does not have an intrinsic tone pattern but varies largely with the preceding tone, and any lexical tones in Mandarin can be neutralized in an unstressed syllable. Also, tone *sandhi* (*i.e.*, change of tone identity due to tonal context) in Mandarin has been well known, including T3 *sandhi*, medial T2 *sandhi* in a trisyllabic word, and tone *sandhi* for several particular morphemes. On the other hand, in Cantonese tone neutralization and tone *sandhi* are hardly observed.

Tone languages usually need both positive and negative

Table 1: *Tone systems of Mandarin and Cantonese.*

Tone feature		Tone number	Tone code	Tone commands
Mandarin	high	T1	55	positive
	rising	T2	35	negative - positive
	low	T3	21(4)	negative
	falling	T4	51	positive - negative
	neutral	T0	*context dependent*	
Cantonese	high level	T1	55	positive
	high rising	T2	35	negative - positive
	mid level	T3	33	zero
	low falling	T4	21	negative
	low rising	T5	13	negative - zero
	low level	T6	22	negative
	high level (entering)	T7	5	positive
	mid level (entering)	T8	3	zero
	low level (entering)	T9	2	negative

tone commands, and a specific tone language shows a specific set of tone command patterns. In previous study, we have revealed the tone command patterns for Mandarin [1] and Cantonese [2, 3], as also shown in Table 1.

T3 and T8 of Cantonese have no tone commands, T2 of both dialects and T4 of Mandarin have a pair of tone commands, and all the other tones have a single tone command (T5 of Cantonese has a negative command only in the early part of the syllable). Especially, T4 and T6 of Cantonese show the same command polarity but T4 gives larger absolute amplitude than T6. The command patterns of entering tones are similar to those of their respective counterparts of non-entering tones, except for a shorter duration. Neutral tone (T0) in Mandarin does not have a fixed pattern of tone command, but it depends on neighboring tones.

Statistical analyses of command parameters for both dialects [1-3] have shown that the onset time of tone commands is nearly constant ahead of the rhyme onset, while the offset time of tone commands varies approximately in a linear relation with the rhyme duration. This correspondence between tone and rhyme is believed to be a general nature of tone languages. The amplitude of tone commands, on the other hand, is more variable, but systematic differences in the amplitude can still be observed between certain tones [2, 3].

3. Overall framework for command extraction

The parameters of the model can only be derived by successive approximation from a good initial estimation. We use the same framework as proposed for Japanese [4] as given below, but modify the second stage (*i.e.*, first-order estimation of tone commands instead of accent commands) and focus our discussion on this particular module. It should be noted that this module is multilingual for tone languages. Namely, the algorithm is language-independent, while all the language-specific information can be stored in separate data tables.

(1) Pre-processing of an observed F_0 contour.

After a series of pre-processing, the observed F_0 contour is approximated by a set of piecewise 3rd-order polynomials which are continuous and differentiable everywhere.

(2) First-order estimation of tone command parameters.

Since the natural angular frequency α of the phrase control mechanism is much smaller than β of the tone control mechanism, phrase components are much more gradual than tone components. If we neglect the effect of phrase components, it can be inferred from Eq. (3) that the *positive maxima* (*p*-maxima) and *negative minima* (*n*-minima) of the first derivative (hence both are *inflection points*) of the F_0 contour should correspond approximately to the onsets and offsets of tone commands with a certain delay.

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

(4) Optimization of parameters by successive approximation.

4. First-order estimation of tone commands

4.1. Why a difficult task?

Detection of tone commands is much more difficult than that of accent commands, because the polarity of tone commands cannot be inferred directly from the F_0 contour itself.

A previous work on Mandarin [8] supposes that tone and phrase components correspond to high-frequency (and hence DC-free) and low-frequency components of F_0 contour respectively. This assumption, however, is not generally valid even for Mandarin, not to mention other tone languages which can be unbalanced in tones of positive and negative commands. And, it ignores the linguistic information of a given language.

In fact, even manual extraction of commands cannot work without linguistic information, because the modeling process is not only a curve fitting, but should also be linguistically meaningful. Without linguistic constraints, the task of tone command extraction is ill-posed because: (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 may not occur in pairs, and hence causes the problem of uncertainty in pattern connection.

To overcome this difficulty, it is necessary to incorporate the information of tone identity in each syllable, together with the corresponding timing (both syllable timing and rhyme timing, so as to give more accurate analysis). It is to be noted that the tone identity should coincide with phonetic realization instead of linguistic form. Namely, tone neutralization, tone *sandhi* and any other changes of tones should be identified and processed after appropriate linguistic and phonetic judgments.

4.2. Recognition of tone command patterns

The model assumes that an intrinsic tone command pattern is associated with each syllable of a particular lexical tone. In practice, however, the command pattern within a syllable can have many variations due to tone coarticulation, inaccurate syllable segmentation or imperfect preprocessing of F_0 contour. Therefore, a set of rules needs to be set up to help recognition.

The basic requirement is that a series of tone commands can be generated from the inflection points of the F_0 contour. Each inflection point corresponds to either onset or offset of a tone command, or in other words, corresponds to either ascent or descent edge of a pedestal. The starting and ending points for each edge of the pedestal can be defined at three levels: 1 (positive), 0 (baseline) and -1 (negative). Therefore we need to determine the starting and ending levels of the pedestal edge corresponding to each inflection point from the polarity of derivative. For example, for a *p*-maximum, we need to judge whether it is an onset of a positive command (0→1) or an offset of a negative command (-1→0) or both (-1→1, *i.e.*, a junction point between a negative and a positive commands).

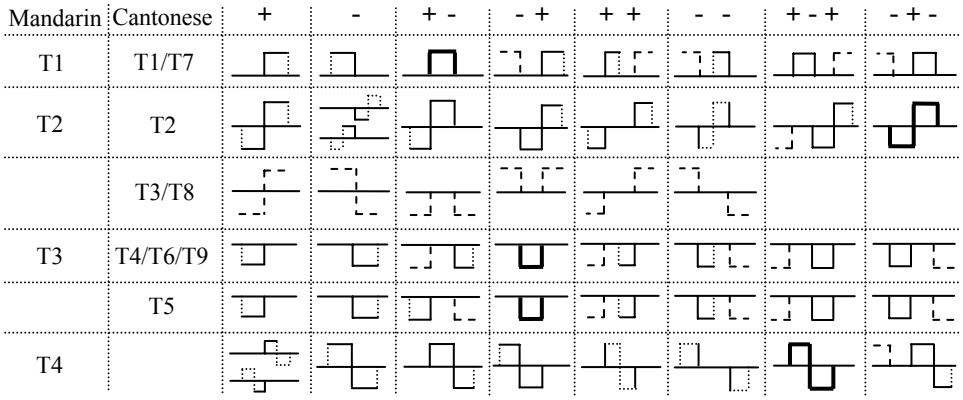


Figure 1: *Tone command patterns associated with different inflection points in each tonal syllable.*

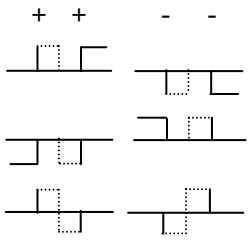


Figure 2: *Insertion of dummy points (onsets or offsets) between two inflection points of the same polarity.*

4.2.1. Intra-syllable command pattern recognition

At the first place, intra-syllable command pattern recognition is conducted, based on some heuristic rules to give the most probable match with the intrinsic command patterns. Figure 1 shows the most frequently observed patterns associated with different series of inflection points for the two dialects. Each column depicts the command patterns associated with the six categories of tonal syllables when a specific series of inflection points (+: *p*-maximum, -: *n*-minimum) are observed within the syllable. The thin solid lines indicate the recognized onsets/offsets of tone commands, while the dotted lines present a reference of the intrinsic tone command patterns. The patterns recognized as identical to the intrinsic ones are indicated by the thick solid lines. For the cases associated with two candidates, the patterns are then determined by the timing.

In Cantonese, T4 and T6 share the same rules because they differ only in amplitude, whereas T5 also applies similar rules since it shares the same polarity with T4/T6 (the difference in duration can be weakened in continuous speech). Especially, neutral tone needs more context-dependent process. Our study shows that most T0's in continuous speech of Mandarin show a negative command, and hence the rules similar as for T3 are employed, with an exceptional process for T0 preceded by T2.

Intra-syllable pattern recognition not only looks at the pattern within a syllable, but also looks into the tonal context to assure the inter-syllable consistency: (1) The dashed lines in Fig. 1 indicate the patterns which are only adopted when they are consistent with the neighboring tone commands, and when they contradict with the tonal contexts, these inflection points are deemed to result from noises and will be deleted; (2) If the first starting level or the final ending level of an entire syllable is 0, check whether it need to be extended to ± 1 so that the onset/offset is shared by the preceding or following tone.

For example, the intrinsic pattern of inflection points in a T1 syllable (for both dialects) should be '+ -', corresponding to the onset and offset of a positive command. In many cases, however, only one of these two inflection points is detected, indicating the onset or offset. Occasionally, an inflection point corresponding to the offset of the preceding command or the onset of the following command can fall in the current syllable (the former occurs more often, for the interpolated contour in the voiceless initial is less reliable), and hence the other five possible patterns of inflection points shown in Fig. 1 should also be taken into account. A dashed onset/offset is added if it coincides with the polarity of the neighboring tone command.

4.2.2. Command pattern refinement

Tone command patterns are not determined after intra-syllable recognition. First, the ending level of an inflection point may not match the starting level of the next inflection point. This is observed not only in some cases within a syllable, but more often across syllables. Second, the onset of utterance-initial tone command or the offset of utterance-final tone command may not be detected from the F_0 contour. Therefore, we apply the following rules to connect the individual patterns in each syllable into a set of well-formed tone commands (pedestals):

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

If the starting (or ending) level of the utterance-initial (or utterance-final) inflection point is not 0 (henceforth assume it in the *i*th syllable), check the syllables before (or after) it until the *k*th syllable is found (*k* can equal to *i*) such that all the syllables between *i* and *k* have the tone types consistent with the nonzero level. Insert a *dummy point* ahead of (or at the end of) the *k*th syllable to act as the utterance-initial command onset (or the utterance-final command offset).

(2) Connect the ending level of the current inflection point and the starting level of the next inflection point appropriately.

(2.1) If only one of these two levels is 0 and the two inflection points are opposite in polarity, which happens only across syllables (henceforth assume the nonzero level in the *i*th syllable while the zero level in the *j*th syllable): check all the syllables between *i* and *j*, and find the *k*th syllable such that all the syllables between *i* and *k* have tone types consistent with the nonzero level while all the syllables between *k* and *j* do not. If *k* equals *j*, change the level 0 to coincide with the nonzero one. Otherwise, insert a *dummy point* in the *k*th syllable and connect it with the nonzero level in the *i*th syllable.

(2.2) If the two levels differ and the two inflection points are of the same polarity, as illustrated in Fig. 2, a *dummy point* is inserted (as depicted by the dotted lines) between the two inflection points to recover the command offset (as shown in the 1st row) or onset (in the 2nd row) or both (in the 3rd row).

(2.2.a) If the two points are in one syllable or in two neighboring syllables, simply insert the *dummy point* in the middle of the two inflection points.

(2.2.b) Otherwise, make a similar search as in (2.1) until the *k*th syllable is found, and insert the *dummy point* there.

(3) Delete the inflection points for which both the starting and the ending levels remain at 0 after the above processes.

4.3. Estimation of timing/amplitude of tone commands

After command pattern recognition based on the inflection points, the parameters for a series of tone commands can be

derived from Eq. (3). As discussed in [7], the process depends on whether the distance between the neighboring two inflection points exceeds a threshold (0.2s) or not. It is to be noted that the amplitudes of step response functions at the command onset and offset (denoted as a_{t1j} and a_{t2j} respectively for the jth command) should be estimated first before deriving the command amplitude A_{tj} by the following equations [7]:

$$A_{tj} = \begin{cases} a_{t1j}, & \text{if } s(T_{1j}) = 0 \text{ and } (e(T_{2j}) \neq 0 \text{ or } d(T_{2j})), \\ -a_{t2j}, & \text{if } d(T_{1j}) \text{ and } e(T_{2j}) = 0, \\ -a_{t2j}/2, & \text{if } d(T_{1j}) \text{ and } e(T_{2j}) \neq 0, \\ (a_{t1j} - a_{t2j})/2, & \text{if } s(T_{1j}) = 0 \text{ and } e(T_{2j}) = 0, \\ A_{t,j-1} + a_{t1j}, & \text{if } s(T_{1j}) \neq 0 \text{ and } (e(T_{2j}) \neq 0 \text{ or } d(T_{2j})), \\ (A_{t,j-1} + a_{t1j} - a_{t2j})/2, & \text{if } s(T_{1j}) \neq 0 \text{ and } e(T_{2j}) = 0, \end{cases} \quad (4)$$

$$A_{tj} = \text{sgn}(a_{t1j}) \cdot A_{t\min}, \quad \text{if } \text{sgn}(A_{tj}) \neq \text{sgn}(a_{t1j}). \quad (5)$$

Here $s(t)$ and $e(t)$ indicate the starting and ending levels of the inflection point at t respectively, while $d(t)$ indicates that the point at t is an inserted *dummy point* instead. Equation (5) is employed to deal with the polarity reversion error. Finally, a back-tracing process as discussed in [6] will be introduced to correct the unnaturally biased command amplitudes.

5. Experimental results

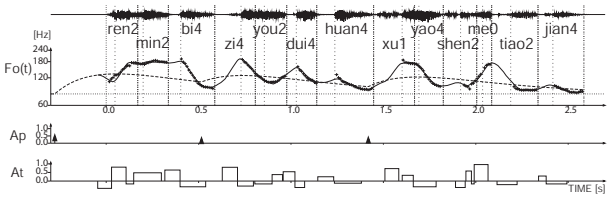
The speech data tested in the current study consist of 40 declarative Mandarin utterances (4~22 syllable long) and 54 declarative Cantonese utterances (5~16 syllable long), each read by a male native speaker of the respective dialect at the normal speech rate. The F_0 values are extracted by a modified autocorrelation analysis of the LPC residual signal. Syllable and rhyme boundaries are marked manually by visual inspection of the waveform and spectrogram. Also, by the aid of visual inspection, the initial value of F_b is set at 80Hz for the Mandarin speaker and 90Hz for the Cantonese speaker.

The performance of the first-order estimation of tone command parameters is evaluated by comparison with the results of manual analysis [1, 2]. When consecutive tone commands of the same polarity 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 tone coarticulation.

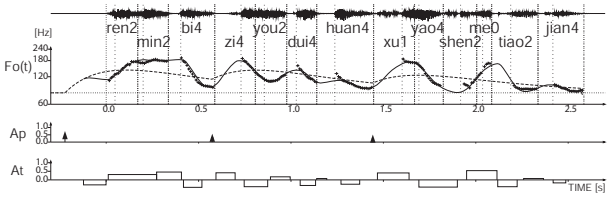
Among the total of 1071 syllables in the speech data, only 7 tone commands are mistakenly deleted, while no tone commands are incorrectly inserted. Hence both the miss and false alarm rates are very low. Furthermore, for the syllables that inherently have tone commands but fail to present any, tone commands inherent to the particular tone type can be forcedly assigned according to the constraints revealed in [1-3] so as to provide an initial value for successive approximation.

For the manual analysis, the average approximation error, defined as the average of RMS errors between the observed and the approximated $\ln F_0(t)$ within voiced intervals for all the utterances, is 0.039 for Mandarin and 0.024 for Cantonese, which are equivalent to a relative error of 3.9% and 2.4% in F_0 respectively. For the automatic analysis, the relative errors in F_0 by the first-order estimation are 16.2% for Mandarin and 10.7% for Cantonese, which reduce respectively to 5.8% and 4.1% after successive approximation. Therefore the automatic results give comparable accuracy as the manual results.

Results on example utterances of Mandarin and Cantonese are shown in Figs. 3 and 4 respectively, where panel (a) gives the manual result, whereas panel (b) gives the automatic result. It is seen that in both dialects the automatically approximated F_0 contours match the observed F_0 contours very well.

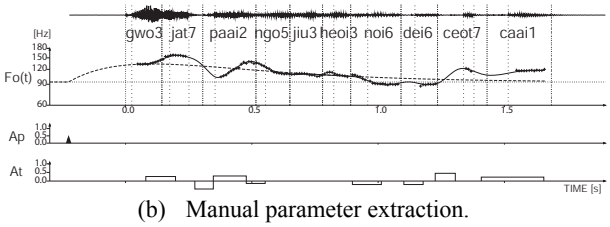


(a) Manual parameter extraction.

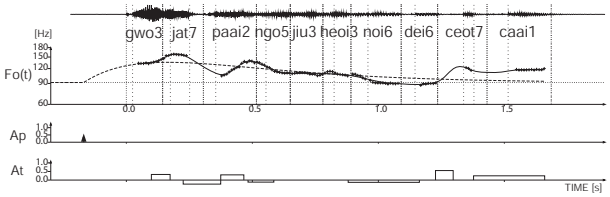


(b) Automatic parameter extraction.

Figure 3: *Analysis of a Mandarin F_0 contour by the model.*



(b) Manual parameter extraction.



(b) Automatic parameter extraction.

Figure 4: *Analysis of a Cantonese F_0 contour by the model.*

6. References

[1] Fujisaki, H.; Wang, C.-F.; Ohno, S.; Gu, W.-T., 2005. Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech Communication* 47 (1-2), 59-70.

[2] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2004. Analysis of F_0 contours of Cantonese utterances based on the command-response model. *Proc. ICSLP'04*, Jeju, Korea, 781-784.

[3] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2005. Identification and synthesis of Cantonese tones based on the command-response model for F_0 contour generation. *Proc. ICASSP'05*, Philadelphia, USA, 289-292.

[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. Japan* 43 (7), 2155-2168. (In Japanese)

[5] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2004. Automatic extraction of tone command parameters for the model of F_0 contour generation for Standard Chinese. *IEICE Trans. Inf. & Syst.* E87-D (5), 1079-1085.

[6] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2004. A method for automatic tone command parameter extraction for the model of F_0 contour generation for Mandarin. *Proc. Speech Prosody 2004*, Nara, Japan, 435-438.

[7] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2006. Tone command parameter extraction for the F_0 contour generation model for Cantonese. *Proc. Spring Meeting of Acoust. Soc. Japan*, 267-268.

[8] Mixdorff, H.; Fujisaki, H.; Chen, G.-P.; Hu, Y., 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. *Proc. Eurospeech'03*, Geneva, Switzerland, 873-876.