ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ANALYSIS AND SYNTHESIS OF CANTONESE F_0 CONTOURS BASED ON THE COMMAND-RESPONSE MODEL

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

[1]The University of Tokyo, [2]Shanghai Jiaotong University

ABSTRACT

Cantonese is a well-known Chinese dialect with a quite complex tone system. We have applied the command-response model to represent F_0 contours of Cantonese speech by defining a set of appropriate tone command patterns. In this paper, the analysis is extended to Cantonese utterances at three different speech rates. By incorporating the effects of tone coarticulation, word accentuation and phrase intonation, the model gives high accuracy of approximations to F_0 contours of Cantonese speech, and hence provides a much better means to quantitatively describe the F_0 contours than the traditional 5-level tone code system. The distributions of timing and amplitudes of commands are investigated, based on which a set of rules is used for synthesis of Cantonese F_0 contours. The validity of the current approach is confirmed by perceptual evaluation of synthetic speech of Cantonese.

1. INTRODUCTION

An accurate and quantitative representation of the essential characteristics of the F_0 contours of speech is necessary for both text-to-speech synthesis and automatic speech recognition, especially for tone languages. For this purpose, the command-response model for the process of F_0 contour generation [1], proposed by Fujisaki and his coworkers, is useful since it can generate very close approximations to observed F_0 contours from a relatively small number of linguistically meaningful parameters. The model has been successfully applied to tone languages including Mandarin [2] and Thai [3], and has recently been extended to Cantonese [4]. In this paper we apply the model to Cantonese at various speech rates and investigate the method for synthesizing F_0 contours by use of a set of rules.

2. CANTONESE TONE SYSTEM

Cantonese is one of the major Chinese dialects spoken by about 70 million people worldwide (in Guangdong and Guangxi provinces of China, Hong Kong, Macau, and in many overseas Chinese communities). Although Cantonese largely shares the same writing system and the same monosyllabic nature with Mandarin, it is much richer in the number of tone types. It is usually accepted that Cantonese has nine citation tones, which preserve the tonal categories of Middle Chinese (7th~10th century A.D.). Table 1 gives some traditional descriptions of all the nine citation tones.

The syllables of entering tones (T7, T8 and T9) end with an unreleased stop coda /p/, /t/ or /k/, and are comparatively shorter in duration than those of non-entering tones. Each entering tone has its counterpart of non-entering tone, showing a similar F_0 pattern. T7, T8 and T9 correspond to T1, T3 and T6 respectively. Therefore in some transcription schemes only six tones are given.

Table 1: Some traditional descriptions of Cantonese tones.

Tone name in Middle Chinese system		Tone * number	Pitch feature	5-level code
Non-entering tones	Upper-level	T1	High level	55
	Upper-elevating	T2	Mid rising	35
	Upper-departing	T3	Mid level	33
	Lower-level	T4	Low falling	21
	Lower-elevating	T5	Low rising	13
	Lower-departing	T6	Low level	22
Entering tones	Upper-entering	T7	High level	5
	Middle-entering	T8	Mid level	3
	Lower-entering	T9	Low level	2

* Note: T1~T4 here are different from those of Mandarin.

Traditionally a 5-level tone code system is adopted for Cantonese after Chao [5], though it varies somewhat from one reference to another. As the first approach for quantifying the tones, it provides a simplified canonical form for tones in isolated syllables.

However, there are some intrinsic limitations of this tone code system, especially when using it for synthesis purpose. First, the five levels are subjective and relative, and in continuous speech the actual F_0 values change with tonal context, word accentuation and phrase intonation. Second, this approach is only semi-quantitative, and cannot characterize the continuous nature of F_0 values. Third, the five levels in Chao's system are perceptually defined, but many researchers interpret them to be the beginning and the end F_0 values within a syllable, and approximate the F_0 contour with straight lines connecting these values. This is a misunderstanding of Chao's system, and the straight lines are too simple to describe the actual F_0 contours accurately.

3. THE COMMAND-RESPONSE MODEL FOR F_0 CONTOUR GENERATION

To overcome the intrinsic limitations of the traditional tone code system, we have introduced the command-response model for the generation process of F_0 contours of Cantonese [4]. The model has been shown to give quantitative descriptions to Cantonese F_0 contours with high accuracy.

Figure 1 shows the diagram of the model for tone languages. It describes F_0 contours in the logarithmic scale as the sum of phrase components, tone components and a baseline level. 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) of both positive and negative polarities 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.

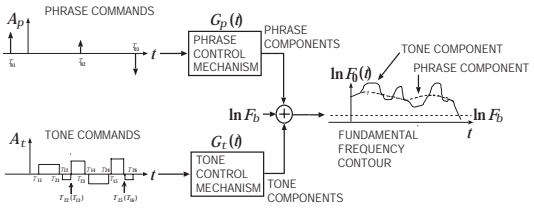

Figure 1: The command-response model for F_0 contour generation with both positive and negative tone commands.

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 details of the parameters are described in reference [1]. A set of tone command patterns needs to be specified for the model to fit a specific tone language.

This model incorporates the effects of tone coarticulation, word accentuation and phrase intonation simultaneously in an explicit way. Tone coarticulation is automatically taken care of by the transfer characteristics of the tone control mechanism. Word accentuation can be implemented either by magnifying the amplitudes or by lengthening the durations of tone commands. Phrase intonation is explicitly represented by the phrase components.

4. ANALYSIS OF CANTONESE F_0 CONTOURS

4.1. Speech data

Two sets of speech materials are used: Speech Material A is designed with a fixed carrier sentence, while Speech Material B includes various meaningful sentences. In our former work [4] only the utterances at normal speech rate of the speaker were analyzed. Here the utterances at three different speech rates are collected and analyzed: slow (4.2 syllables/s), normal (5.6 syllables/s) and fast (6.9 syllables/s).

Speech Material A consists of carrier sentences "hon3 gin3 __ faai3 gong2 ceot7 lai4" (Speak it out quickly when you see __), in each of which the target syllable *maa* or *ma*(*a*)*k*, carrying each of the nine tones, is embedded at the underlined position. Each sentence was uttered eight times at each of the three speech rates by a native male speaker of Cantonese (from Guangzhou). Although in the lexicon there are no characters uniquely corresponding to *maa*2 and *maa*3, the speaker was trained to utter pseudo-words for them. Carrier sentences with meaningful target words *wai*2 and *wai*3 were also recorded and analyzed.

Speech Material B consists of 20 declarative sentences each with 5~14 syllables. Each sentence was uttered three times at each of the three speech rates.

The speech signal was digitized at 10 kHz with 16bit precision. The fundamental frequency was extracted by the modified autocorrelation analysis of the LPC residual. Syllable boundaries and rhyme boundaries were labeled by visual inspection of the waveform and the spectrogram.

4.2. Tone command patterns

First, F_0 contours of Speech Material A were analyzed by the method of Analysis-by-Synthesis. Our work on the utterances at normal speech rate has shown the tone command patterns for Cantonese tones as below [4]:

T1: positive
T2: initially negative and later positive
T3: zero
T4: negative
T5: initially negative and later zero
T6: negative

Thus, T2 is considered to have a pair of tone commands. Also, it is to be noted that T4 gives a command pattern with the same polarity but larger absolute value of amplitude as compared with T6. As for the three entering tones, the command patterns are similar to those of their respective counterparts of non-entering tones (positive for T7, zero for T8, and negative for T9), but the duration is always shorter because voicing is interrupted by the unreleased stop coda.

With this set of tone command pattern definitions, very close F_0 approximations can be achieved. This was further confirmed by applying Analysis-by-Synthesis to approximate F_0 contours of Speech Material B. An example is shown in Fig. 2.

Analysis of slow and fast speech indicates the same set of tone command patterns. In other words, the tone command patterns are not affected by the speech rate. A comparison between utterances at different speech rates shows that the slower the speech is, the more phrase commands tend to be introduced within an utterance.

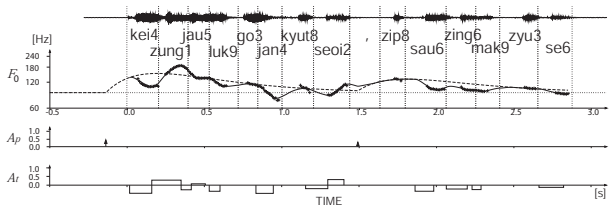

Figure 2: Analysis-by-Synthesis of the F_0 contour of a Cantonese utterance in Speech Material B.

4.3. Timing and amplitude of tone commands

Like Mandarin, a syllable in Cantonese can be divided into two parts: initial and final. The initial can be a consonant (unvoiced

or voiced), a semi-vowel or nil. The final is composed of the main vowel(s) and an optional nasal or stop coda. In Cantonese, the nasal /m/ or /ng/ can also form a syllable by itself. Such syllabic nasals are also regarded as finals. We define the rhyme, *i.e.* the portion carrying the tone, to be the final excluding the stop coda.

Figure 3 shows the timing of tone commands relative to the rhyme at normal speech rate for all the tones except for T3 and T8. The abscissa indicates the rhyme duration, while the ordinate indicates the timing relative to the rhyme onset. The lower and upper groups of points indicate the onsets and offsets of tone commands respectively. For T2, we assume that the onset of the 2nd tone command coincides with the offset of the 1st tone command just for simplicity. The top group of points shown for T2 indicates the offsets of the 2nd tone commands.

Some systematic tendencies are observed. The onsets of tone commands are concentrated within an interval, *viz*., 0~100 ms (for T4, T6 and T9) or 50~150 ms (for other tones) prior to the rhyme onset regardless of rhyme duration. The offsets of tone commands show a high correlation with the rhyme duration, and can be approximated by linear regression. Such tendencies suggest that timing may be constrained on those lines shown in Fig. 3. Similar patterns are also observed from the analysis of slow and fast speech. The distributions of parameters are hardly affected by the speech rate. This finding is consistent with that for Mandarin [2].

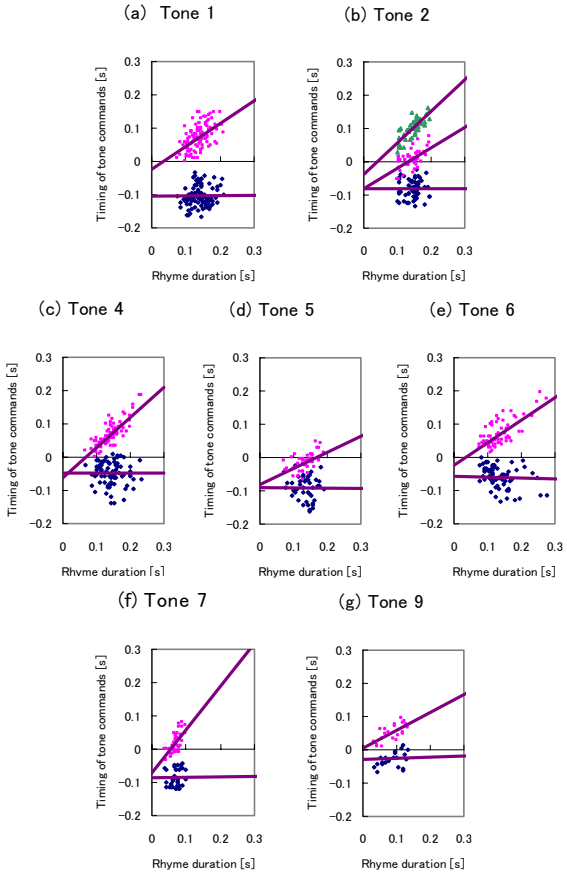

Figure 3: Tone command timing relative to the rhyme.

The amplitudes of tone commands are also investigated. They are shown to be quite scattered and the correlation with rhyme duration is very low. Such a wide range of amplitude distribution reflects a continuous strength of word accentuation. For practical applications, their amplitudes may be constrained to several discrete levels.

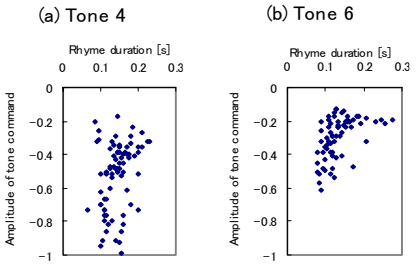

Figure 4: Tone command amplitudes of T4 and T6.

Figure 4 shows the relationship between tone command amplitude and rhyme duration at normal speech rate for T4 and T6, both of which are characterized by a negative tone command. It is observed that the amplitudes of T4 occupy a relatively more negative region, but there is a considerable overlap with that of T6. This suggests that the distinction between the tone command patterns for T4 and T6 may be influenced by the context.

5. SYNTHESIS OF CANTONESE F_0 CONTOURS BY RULES

Since the model can generate very close approximations to F_0 contours of Cantonese, it can be used for synthesizing Cantonese F_0 contours, as have been done for Mandarin [2] and Thai [3]. A set of rules are derived from the constraints described above to standardize the command parameters.

Table 2: Quantization of tone command amplitude.

Tone type	T1	T2		T4	T5	T6	T7	T9
Enhanced	0.35	-0.40	0.40	-0.85	-0.60	-0.45	0.50	-0.50
Normal	0.25	-0.25	0.30	-0.60	-0.40	-0.30	0.35	-0.375
Suppressed	0.15	-0.10	0.20	-0.35	-0.20	-0.20	0.20	-0.25

Table 3: Quantization of phrase command timing/magnitude.

Position		Timing [s]*	Magnitude
Utterance-initial	High	-0.25	0.55
	Medium		0.40
	Low		0.25
Utterance-medial	High	-0.30	0.25
	Medium	-0.20	0.15
	Low	-0.10	0.05

* The timing is relative to the rhyme onset of the first syllable in the corresponding phrase.

The timing of tone commands are determined by the linear regression lines for each tone type as shown in Fig. 3. The amplitudes of tone commands are quantized to three levels (suppressed, normal, and enhanced) for each tone type, as shown in Table 2 (note that T2 has two commands).

Based on a similar analysis of timing and magnitude of phrase commands, their values are also quantized depending on the position of the phrase commands in the utterance, as shown in Table 3. The timing of utterance-initial phrase command is fixed close to its mean value obtained by the analysis.

6. PERCEPTUAL EVALUATION

In order to test the validity of our analysis and rules for synthesis, two subjective evaluation experiments were conducted. The subjects are two native Cantonese speakers from Hong Kong. In each experiment, a set of PSOLA analysis-resynthesized speech was randomly presented to the subjects for evaluation.

6.1. Experiment 1 – tone identification

The purpose of Experiment 1 was to test the validity of the tone command configuration for each tone type. The utterances of a carrier sentence with target words *maa* or *ma*(*a*)*k* of nine different tones were resynthesized with model-based F_0 contours, where the tone commands for the target words were produced by the rules, with amplitudes set at each of the three levels. The subjects were asked to identify the target word from a list of characters with different tones. Two utterances were used for each tone type and each stimulus was presented three times. Due to the absence of characters uniquely corresponding to *maa*2 and *maa*3, carrier sentences with target words *wai*2 and *wai*3 were used for identification of T2 and T3 instead.

The identification result of the words with tone commands of normal amplitudes is 100% correct, which shows the validity of the rules for each tone type. For the words with tone commands of enhanced or suppressed amplitudes, most are correctly identified, except for the enhanced T6 and suppressed T4. Both subjects categorize all the six samples of enhanced T6 as T4, while Subject A also categorizes three samples (*i.e.* half of the samples) of suppressed T4 as T6.

This is due to the fact that the tone command patterns for T4 and T6 differ mostly (if not only) in amplitudes whose values are overlapping in their enhanced or suppressed cases. This implies that command amplitudes of T4 and T6 vary with the context, and perceptual distinction between T4 and T6 is also based on relative judgment and linguistic context.

6.2. Experiment 2 – naturalness evaluation

Experiment 2 evaluated the naturalness of the whole utterance with a 10-point scale (higher means better). Three utterances were used, each of which provided seven stimuli as listed below, and the judgment was made 10 times for each.

(1) the original F_0
(2) model-based F_0 with the minimum mean square error
(3) synthetic F_0 with only tone commands generated by rules
(4) synthetic F_0 with only phrase commands generated by rules
(5) synthetic F_0 with only command timing generated by rules
(6) synthetic F_0 with both timing and amplitude/magnitude of both tone and phrase commands generated by rules
(7) same as (6) except that amplitudes/magnitudes are only quantized to the normal or medium level

The mean scores of evaluation are listed in Table 4. The degradation of naturalness introduced by the model is negligible. After introducing all the rules, the degradation is still quite small even in the simplest case. This result verifies the validity of the model and the set of rules.

Table 4: Mean scores for the naturalness of F_0.

Stimulus / Subject	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Subject A	9.43	9.33	8.77	8.77	8.73	8.63	8.43
Subject B	9.63	9.70	9.67	9.67	9.70	9.47	9.43
Average	9.53	9.52	9.22	9.22	9.22	9.05	8.93

7. CONCLUSION

Our experiments have shown that with a set of well-defined tone command patterns, the command-response model can approximate F_0 contours of Cantonese utterances at various speech rates with high accuracy. Compared with the traditional 5-level tone code system, this model provides more accurate means to represent the F_0 contours of Cantonese speech quantitatively. All the findings on Cantonese tones in continuous speech given in references [6-8] can be explained by the proposed command-response model.

The relationship between tone command timing and rhyme duration shows that certain constraints can be applied, which, together with the quantization of command amplitude/magnitude, are used as a set of rules to synthesize Cantonese F_0 contours. The validity of the rules has been confirmed by perceptual evaluation of synthetic speech.

These rules for parameter standardization are still preliminary. In order to build a text-to-speech synthesis system, we need to find out how the command parameters can be derived automatically from text. We expect to do it in our future research.

8. REFERENCES

[1] Fujisaki, H., "Information, prosody, and modeling – with emphasis on tonal features of speech," *Proc. Speech Prosody 2004*, Nara, Japan, pp. 1-10, 2004.

[2] Wang, C., Fujisaki, H., Tomana, R. and Ohno, S., "Analysis of fundamental frequency contours of Standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation," *Proc. ICSLP 2000*, vol. 3, pp. 326-329, Beijing, China, 2000.

[3] Fujisaki, H., Ohno, S. and Luksaneeyanawin, S., "Analysis and synthesis of F_0 contours of Thai utterances based on the command-response model," *Proc. 15th ICPhS*, Barcelona, Spain, pp. 1129-1132, 2003.

[4] Gu, W., Hirose, K., and Fujisaki, H., "Analysis of F_0 contours of Cantonese utterances based on the command-response model," *Proc. ICSLP'04*, Jeju Island, Korea, 2004.

[5] Chao, Y.-R., *Cantonese Primer*, Harvard University Press, Cambridge, 1947.

[6] Li, Y., Lee, T. and Qian, Y., "Acoustical F_0 analysis of continuous Cantonese speech," *Proc. ISCSLP'02*, Taipei, pp. 127-130, 2002.

[7] Lee, T., Kochanski, G., Shih, C. and Li, Y., "Modeling tones in continuous Cantonese speech," *Proc. ICSLP'02*, Denver, USA, pp. 2401-2404, 2002.

[8] Li, Y., Lee, T. and Qian, Y., "F_0 analysis and modeling for Cantonese text-to-speech," *Proc. Speech Prosody 2004*, Nara, Japan, pp. 467-470, 2004.