IDENTIFICATION AND SYNTHESIS OF CANTONESE TONES BASED ON THE COMMAND-RESPONSE MODEL FOR F_0 CONTOUR GENERATION

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

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

ABSTRACT

Cantonese is a well-known Chinese dialect with a quite complex tone system. We have successfully applied the command-response model to represent F_0 contours of Cantonese speech by defining a set of appropriate tone command patterns. It provides an efficient means to describe Cantonese F_0 contours with high accuracy. In this paper, both qualitative and quantitative descriptions of the command patterns of nine tones are presented. The various cues for identifying each tone are investigated, based on which a set of rules is derived to synthesize F_0 contours of Cantonese. Perceptual experiments are also conducted to test the validity of the rules for each tone and to evaluate the naturalness of synthetic speech based on those rules.

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. It is even more challenging for Cantonese which is well-known for its system of nine citation tones. For this purpose, the command-response model for the process of F_0 contour generation [1], proposed by Fujisaki and his coworkers, has been applied to Cantonese [2]. It can generate very close approximations to observed F_0 contours from a relatively small number of linguistically meaningful parameters. In this paper we further define the tone command patterns for Cantonese by a set of quantitative rules, which is then used for synthesizing F_0 contours of Cantonese.

2. CANTONESE TONE SYSTEM

Cantonese is one of the major dialects of Chinese spoken by over 70 million people worldwide (in Guangdong and Guangxi provinces of China, Hong Kong, Macau, and in many overseas Chinese communities). 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 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 some transcription schemes only give six tones.

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	High 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

* T1~T4 here are different from those of Mandarin.
** Guangzhou speakers have two varieties of T1, *i.e.* 55 (high level) and 53 (high falling), but young speakers in Guangzhou tend to merge 53 into 55, as Hong Kong speakers do.

Traditionally a 5-level tone code system is adopted for Cantonese after Chao [3], though it varies somewhat from one reference to another. It provides a simplified canonical form for tones in isolated syllables. However, it has some intrinsic limitations due to the relative, subjective and symbolized nature of the codes, and it is unable to give quantitative representations of continuous F_0 contours for purpose of speech synthesis.

Some recent works [4, 5] try to use a set of F_0 templates derived from a certain corpus to characterize each tone quantitatively. However, those templates vary with the corpus.

3. THE COMMAND-RESPONSE MODEL FOR F_0 CONTOUR GENERATION

A novel approach based on the command-response model for the generation process of F_0 contours, has been successfully introduced to Mandarin [6] and Cantonese [2], to overcome the intrinsic limitations of the traditional tone code system.

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 produce phrase components through the phrase control mechanism, giving the global shape of F_0 contour, while the tone commands 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. The details of the formulation are described in the work [1]. A set of tone command patterns needs to be specified for the model to fit a specific tone language.

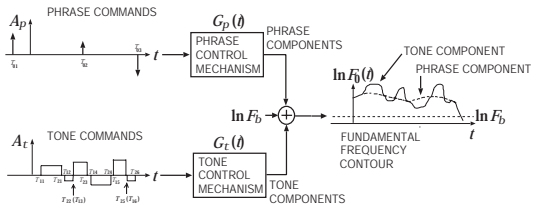

Figure 1: The command-response model for F_0 contour generation with both positive and negative tone commands.

4. ANALYSIS OF CANTONESE TONES

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. The materials were recorded by two native speakers, Speaker A (male) from Guangzhou and Speaker B (female) from Hong Kong, whose utterances represent the two most popular regional dialects of Cantonese. The utterances at different speech rates were collected for the same text: slow (3.5 syllables/s), normal (4.1 syllables/s) and fast (5.4 syllables/s) for Speaker A. Since the analysis results for the two speakers are quite similar, only the data analysis for Speaker A will be presented in this paper.

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 (only syllables of T7 and T9 consist of /a/ instead of /aa/), is embedded at the underlined position. Although in the lexicon there are no characters uniquely corresponding to *maa*2 and *maa*3, the speakers were trained to utter pseudo-words for them. Carrier sentences with meaningful target words *wai*2 and *wai*3 were also recorded and analyzed. Besides, Speech Material A also consists of the carrier sentences in which bi-syllabic words of tone *sandhi* are embedded. Each sentence was uttered eight times at each of the three speech rates.

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

4.2. Qualitative patterns of tone commands

First, F_0 contours of Speech Material A are analyzed by the method of Analysis-by-Synthesis. Our former work [2] indicates that the command patterns for Cantonese tones are as follows:

T1: positive
T2: initially negative and later positive
T3: zero
T4: negative (larger absolute amplitude than T6)
T5: initially negative and later zero
T6: negative (smaller absolute amplitude than T4)
T7: positive (with stop coda)
T8: zero (with stop coda)
T9: negative (with stop coda)

Thus, only T2 has a pair of commands. It is also to be noted that T4 and T6 are distinguished quantitatively. The command patterns of entering tones are similar to those of their respective counterparts of non-entering tones, but the duration is always shorter since voicing is interrupted by the unreleased stop coda.

A *loose* correspondence can be observed between the set of command patterns and the traditional 5-level tone code system, if we take the mid level 3 as a reference and map the higher and lower levels to positive and negative commands respectively. However, a command produces a dynamic F_0 curve instead of a static F_0 value through the tone control mechanism, and can capture shape variations by adjusting the timing and amplitude.

With this set of tone command pattern definitions, very close F_0 approximation can be made to speech at various speech rates. This is confirmed by applying Analysis-by-Synthesis to F_0 contours of Speech Material B. An example is shown in Fig. 2, where the relative error of approximation in F_0 is only 2.3%.

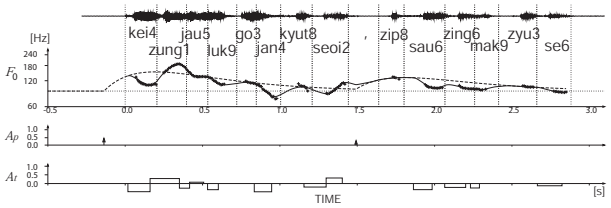

Figure 2: Analysis-by-Synthesis of the F_0 contour of a Cantonese utterance in Speech Material B.

4.3. Tone *sandhi* in Cantonese

Although there is no phonological tone *sandhi* in Cantonese, it is generally accepted that two kinds of *habitual* tone *sandhi* exist: high level tone *sandhi* and high rising tone *sandhi* [7]. In some specific words, tones of some syllables will by habit be changed into high level tone or high rising tone. The high rising tone *sandhi* occurs more frequently but less stable across speakers.

Analysis of carrier sentences with bi-syllabic target words of tone *sandhi* shows that the command patterns of the two kinds of tone *sandhi* are consistent with that of T1 and T2 respectively. Therefore, based on the command-response model, tone *sandhi* can be defined as the modification of intrinsic command pattern.

4.4. 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 form a syllable by itself. Such syllabic nasals are also regarded as finals. We define the rhyme, *i.e.* the portion carrying tones, to be the final excluding the stop coda.

In our work [2] we have shown some systematic tendencies of the timing and amplitude of tone commands from the analysis of Speech Material B. Here we present the analysis on the target syllables at various speech rates in Speech Material A, for intrinsic patterns can be observed the best in a well controlled context. There are 24 samples of syllables for each tone type.

Figure 3 shows the command timing relative to the rhyme at three speech rates for all the tones except 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. The top group of points for T2 indicates the offsets of the 2nd commands (the onset of the 2nd command is assumed to coincide with the offset of the 1st command).

It is observed that the onsets of tone commands are located around a constant distance prior to the rhyme onset regardless of rhyme duration, while 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 a set of straight lines. It is to be noted that the offset of T1 here shows very good linear regression tendency, while in the figure given in our work [2] it shows the lowest correlation among the nine tones. This is due to the fact that in continuous speech Speaker A from Guangzhou may use two varieties of T1 and hence the offsets shown in our work [2] are actually distributed between two linear lines of different slopes.

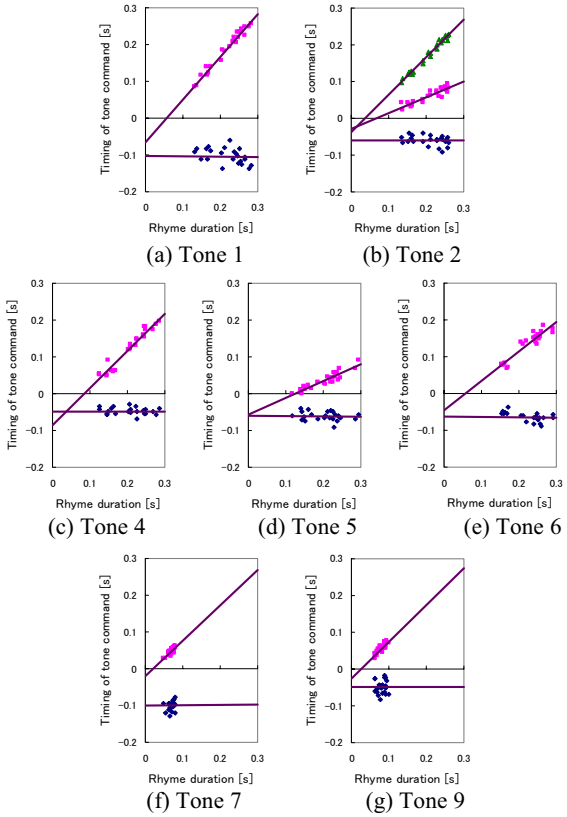

Figure 3: Tone command timing relative to the rhyme.

On the other hand, the amplitudes of tone commands are quite scattered and the correlation with rhyme duration is very low. The wide range of amplitude distribution reflects a continuous strength of word emphasis. For practical applications, their amplitudes may be constrained to several discrete levels. Different from shown in the figure given in our work [2], the distributions of amplitudes of T4 and T6 here are distinctly separated. It indicates that the intrinsic command patterns of T4 and T6 are clearly distinguished by amplitude but in continuous speech the ranges of amplitudes of T4 and T6 may overlap due to various contextual effects.

4.5. Quantitative identification of tones

For the purpose of speech synthesis and recognition, quantitative descriptions of tones are required. For example, T4, T5 and T6 are all characterized by a negative command, and hence they need to be quantitatively defined in terms of timing or amplitude.

Table 2 gives the statistics of command timing (onset and offset), command amplitude and rhyme duration based on the target syllables in carrier sentences. For offset time, the correlation coefficients and the slopes of linear regression lines as shown in Fig. 3 are given. Note that T2 has two commands.

Table 2: Quantitative analysis of Cantonese tones.

			T1	T2		T4	T5	T6	T7	T9
Tim-ing	onset [s]	Mean	-0.10	-0.06	–	-0.05	-0.06	-0.06	-0.10	-0.05
		S. D.	0.02	0.01	–	0.01	0.01	0.01	0.01	0.02
	offset	Corr.	0.99	0.91	0.99	0.97	0.92	0.93	0.80	0.84
		Slope	1.16	0.43	1.02	1.01	0.45	0.80	0.96	1.00
Amp.	Mean		0.22	-0.26	0.42	-0.66	-0.45	-0.19	0.34	-0.31
	S. D.		0.03	0.06	0.09	0.13	0.08	0.04	0.06	0.07
Mean rhyme dur.			0.21	0.21		0.21	0.20	0.22	0.07	0.08

The major difference between T4 and T6 is in the absolute value of command amplitude, being significantly larger for T4 than for T6, without overlapping in the measured data. On the other hand, the primary difference between T5 and T4/T6 is not in the amplitude but in the duration of command as indicated by the slopes of linear regression lines – the slope for T5 is about half of those for T4 and T6. This is in line with our qualitative description that T5 has no command late in the syllable.

Other findings include: (1) the onsets of positive commands (T1 and T7) occur significantly earlier than those of negative commands; (2) the entering tones show significantly larger absolute amplitudes than their counterparts of non-entering tones; (3) the rhyme duration does not show any significant difference among non-entering tones.

The tones without tone commands, T3 and T8, are not listed in Table 2. The mean durations of their rhymes are also calculated: 0.22s for T3 and 0.14s for T8. It is to be noted that the durations of T7 and T9 listed in Table 2 cannot be compared directly with those of non-entering tones because syllables of T7 and T9 are composed of shorter vowels than others. The result confirms that rhyme duration of entering tones is significantly shorter than that of their counterparts of non-entering tones.

5. SYNTHESIS OF CANTONESE F_0 CONTOURS

Since the model can generate very close approximations to F_0 contours of Cantonese, it can be used for synthesizing the F_0 contours. A set of rules is derived from the quantitative analysis on both Speech Material A and B as described above and in our work [2] respectively, to standardize the command parameters.

The timing of tone commands are determined by the constant lines and linear regression lines for each tone type as shown in Fig. 3. The amplitudes of tone commands are quantized to three levels for each tone type (set close to μ and $\mu \pm 1.5\sigma$ of each distribution respectively), as given in Table 3.

Table 3: 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 4: Quantization of phrase command timing/magnitude.

Position	Level	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.

Based on a similar analysis of timing and magnitude of phrase commands, their values are also quantized depending on the position of phrase commands in the utterance, as given in Table 4. The timing of utterance-initial phrase command is fairly stable and hence is fixed close to its mean value.

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 four native speakers of Cantonese from Hong Kong. Subjects I and II are male, while Subjects III and IV are female who both have professional experiences in speaking Cantonese. 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. First, the utterances of a carrier sentence with target words *maa* or *ma(a)k* of 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 of *maa* or *ma(a)k* with different tones. Two utterances were used for each tone type and each stimulus was presented three times. Due to the absence of characters corresponding to *maa*2 and *maa*3, carrier sentences with target words *wai*2 and *wai*3 were used instead, to be identified from a list of characters *wai* of different tones.

The result of identification 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. All the subjects categorize all the six samples of enhanced T6 as T4, and Subject III also categorizes three samples (*i.e.* half of the six 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, of which the values are overlapping in their enhanced or suppressed cases. Since in continuous speech command amplitudes of T4 and T6 vary with the context, it implies that perceptual distinction between T4 and T6 is based on relative judgment and linguistic context. Another implication is that emphasis on T4 and T6 may be indicated by cues other than command amplitude.

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 fixed to the normal or medium level

The mean scores of evaluation are listed in Table 5. The degradation of naturalness introduced by the model is negligible. After introducing all the rules, the degradation is still quite small even in the most simplified case, especially for Subjects I and II. This result confirms the validity of the model and the set of rules.

Table 5: Mean scores for the naturalness of F_0.

Stimulus / Subject	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I	9.63	9.70	9.67	9.67	9.70	9.47	9.43
II	8.90	8.90	8.90	8.90	8.90	8.90	8.90
III	9.43	9.33	8.77	8.77	8.73	8.63	8.43
IV	9.83	9.73	9.67	9.67	9.63	9.67	9.33
Average	9.45	9.42	9.26	9.26	9.24	9.17	9.02

7. CONCLUSION

With a set of appropriate tone command patterns, the command-response model has been shown to approximate F_0 contours of Cantonese utterances at various speech rates with high accuracy. The patterns of nine tones can be identified by polarity, amplitude or duration of commands. Such a set of quantitative definitions, together with the quantization of command amplitude/magnitude, can be used as a set of rules to synthesize Cantonese F_0 contours. The validity of the current approach has been confirmed by perceptual evaluation of synthetic speech.

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] 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.

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

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

[5] 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.

[6] Wang, C., Fujisaki, H., et al., "Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command-response model," *Proc. Eurospeech'99*, Budapest.

[7] Rao, B., Ouyang, J., and Zhou, W., *Guangzhou Dialect Dictionary*, The Commercial Press, Hong Kong, 1981.