F0 Analysis and Modeling for Cantonese Text-to-Speech

Yujia Li, Tan Lee and Yao Qian

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong
{yjli; tanlee; yqian}@ee.cuhk.edu.hk

Abstract

This paper presents a study on the control of fundamental frequency (F0) in Cantonese text-to-speech (TTS) systems. The surface F0 contour of an utterance is considered as the combination of tone-related local components and phrase-level long-term variation. A novel method of F0 normalization has been developed to effectively separate them. Statistical analysis is performed for the phrase curves and the tone contours extracted from a large speech corpus, and the results are summarized into regular patterns. These patterns are used as the basic templates in a non-parametric F0 model, from which utterance-level F0 contours can be generated. Perceptual test shows the naturalness of speech naturalness is significantly improved by the new F0 model. The MOS increases by 0.65 over a five-point scale.

1. Introduction

For the generation of highly natural synthetic speech, proper control of F0 (or fundamental frequency) is of primary importance. F0 contour is one of the major acoustical manifestations of supra-segmental features such as tone, pitch accent and intonation, which are all critical to perceptual naturalness of human speech. However, F0 is a highly variable acoustical parameter. It is determined by both the physical and the linguistic aspects of speech production. While the value of F0 measured at a particular instant indicates the vibration frequency of the speaker's vocal cords, the time-varying F0 contour carries abundant information about the linguistic structure of the sentence. In addition, the linguistic significance of F0 is language-dependent. For most Western languages, F0 is a feature of lexical accent and intonation. For many Asian languages like Chinese and Thai, the meaning of a syllable is determined by its F0 pattern.

The most effective way of understanding F0 variation is via the analysis of natural human speech. Existing approaches can be categorized into *acoustical analysis* and *analysis-by-synthesis*. Acoustical analysis deals with the acoustical measurements directly and attempts to find out how the surface F0 contour depends on a particular factor of interest. The approach of analysis-by-synthesis typically involves a parametric production model that attempts to approximate the observed contour. The optimized parameters in the best approximation reveal the underlying contributions of the respective factors [1,2].

This paper presents an investigation on the variation of F0 in natural Cantonese speech, with the goal of establishing an effective mechanism of prosody control in Cantonese text-to-speech applications. Cantonese is one of the major Chinese dialects spoken by over 70 million people in South China, Hong Kong and overseas [3]. Our investigation starts with statistical analysis of F0 contours extracted from a large amount of continuous speech data. A novel method of F0 normalization has been developed to separate tone-related local components from phrase-level long-term variation. In particular, phrase-level intonation, word-level tone patterns and their contextual variation have been studied. Accordingly, a non-parametric prosody model is established for automatic F0 generation in a Cantonese TTS system that we developed earlier [4]. The effectiveness of the prosody model is then evaluated by subjective listening test.

2. Tones in Cantonese

2.1. Tone system

Spoken Cantonese is made up of a sequence of monosyllabic sounds. Each Chinese character is pronounced as a single syllable that carries a specific tone. Each syllable consists of an *Initial* and a *Final* part. Cantonese is said to have nine citation tones that are characterized by different stylized pitch patterns as illustrated in Figure 1. The so-called "entering" tones occur exclusively with "checked" syllables, i.e. syllables ending in an occlusive coda /p/, /t/ or /k/. They are contrastively shorter in duration but coincide with a non-entering counterpart in terms of pitch contour. In many transcription schemes, only six distinctive tone categories, labeled as Tone 1 to Tone 6, are defined [5].

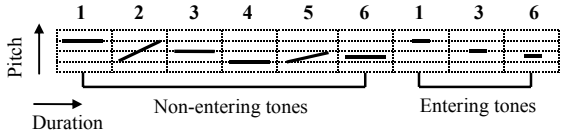

Figure 1: *Tones in Cantonese: schematic description*

2.2. Acoustical realization

Acoustically, tone is realized by the F0 movement across the voiced portion of a syllable. In a Cantonese syllable, the Final can be regarded as voiced while the Initial is either voiced or unvoiced. Figure 2 depicts the measured F0 contours of the nine tones, which were computed by averaging over 1,800 monosyllabic utterances spoken by a male native speaker. The utterances cover most of the tonal syllables used in today's Cantonese. It is seen that the acoustical realization of tones in isolation reflects the schematic patterns very well.

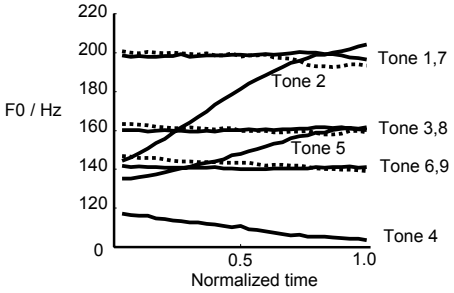

Figure 2: *F0 profiles of Cantonese tones uttered in isolation*

In continuous speech, the actual realizations of tones may vary greatly. An example is shown as in Figure 3. The utterance contains a Cantonese sentence. The curve in the upper part is the extracted F0 contour and the lower part is the concatenation of the respective schematic tone patterns. The characters and their respective tone identities are also shown in alignment with the F0 contour. Clearly, in continuous speech, tones are not realized as their canonical patterns. Even the same tone can be realized quite differently, like the five occurrences of Tone 3 in the example. In addition, at the syllable boundaries, especially when the tones are opposite in terms of levels, there is an obvious tendency that the tone contours compromise with each other to make a smooth transition. By examining the whole F0 contour, it is found that the later the position is, the lower height the tone is realized with, as seen from the occurrences of Tone 3 and Tone 1 (marked with circles).

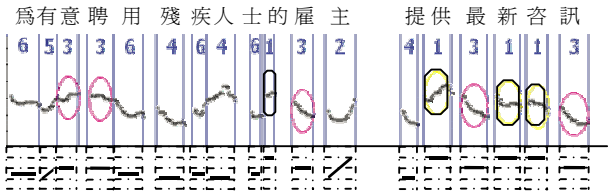

Figure 3: *F0 contour of a continuous speech utterance*

3. Speech Corpus: CUProsody

This research is done based on a large corpus of continuous speech, namely *CUProsody*, which was developed at the Digital Signal Processing Laboratory of the Chinese University of Hong Kong [7,8]. The corpus was recorded from a trained female speaker. It consists of 1,200 newspaper sentences and 1,000 spontaneous sentences. Only the newspaper sentences are used in this research. The speech data in *CUProsody* were manually annotated at orthographic and phonemic levels. Each utterance is accompanied with a Chinese sentence and the respective syllable pronunciations labeled by the Jyutping (LSHK) scheme [5]. The text content of each utterance was manually segmented into words.

All utterances were automatically segmented at Initial and Final level using the forced alignment technique with a set of pre-trained hidden Markov models (HMM). Subsequently the duration of Initial and Final segments were obtained. F0 contours were automatically extracted using the function "get_f0" of the ESPS software [6].

Table 1: *A summary of the CUProsody corpus [7]*

Overall range of F0: 140-300Hz					
No. of sentences = *1200*; Average utterance length: *66 syllables*					
No. of words = *38743*; Average word length: *2.05 syllables*					
1-syllable words	2-syllable words	3-syllable words	4-syllable words		
22.9%	*58.6%*	*11.4%*	*5.5%*		
No. of syllables = *79528*; Average syllable duration: *0.192s*					
Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
25.3%	*12.6%*	*16.1%*	*17.1%*	*6.4%*	*22.6%*

4. F0 Normalization

In order to capture tone contours precisely from an utterance, we have proposed a method of F0 normalization based on a properly estimated phrase curve [7]. The phrase curve reflects the long-term trend of F0 movement over an intonation phrase. Suppose that all syllables in the phrase carry the same tone. The phrase curve is defined as a straight line that best approximates the F0 variation across the whole phrase.

However, in most cases, syllables generally carry different tones. To facilitate the estimation of the phrase curve, we proposed to use a set of relative tone ratios to convert one tone to another tone with an equivalent F0 level. The creation of such tone ratios is based on the assumption that, for the same speaker, the relative positions of different tones remain largely invariant locally, i.e. between neighboring syllables. Details about the computation of tone ratios can be found in [7,8]. Table 2 shows a six-by-six matrix of the tone ratios, computed based on 1,200 utterances in the *CUProsody*, where R_{ij} denotes the relative height ratio of Tone *i* over Tone *j*.

Table 2: *Relative tone ratios derived from CUProsody*

		j					
	R_{ij}	1	2	3	4	5	6
i	1	0.97	1.39	1.28	1.60	1.39	1.35
	2	0.71	0.99	0.92	1.11	0.95	0.97
	3	**0.80**	**1.07**	**1.02**	**1.32**	**1.13**	**1.13**
	4	0.65	0.91	0.83	1.08	1.00	0.94
	5	0.71	0.99	0.93	1.16	1.02	1.01
	6	0.73	1.01	0.95	1.22	1.07	1.05

In our research, Tone 3 is selected as a reference tone, to which all tones would be converted. Given an occurrence of Tone k, we can convert its height to an equivalent height as if what Tone 3 should be in this position, by multiplying with the conversion ratio R_{3k}. For example, if the height of an occurrence of Tone 4 is $150Hz$, the equivalent height of Tone 3 would be equal to $150Hz \times 1.32 = 198Hz$. In this way, all tones' heights in the phrase are converted to Tone 3's. The phrase curve is then obtained by performing linear regression over these converted tone heights.

Subsequently, F0 normalization is done by dividing each original tone contours by the corresponding F0 value on the phrase curve. It was observed that the variance of tone contours would be much reduced (up to 30%) by the proposed method of F0 normalization [8].

5. F0 Analysis

We assume that the F0 contour of a Cantonese utterance is the combination of phrase-level intonation movement and local tone contours. Similar assumption has been widely adopted in other research [2] [9] [10].

5.1. Phrase curves

At sub-utterance level, intonation phrase boundaries were detected by a pause longer than 0.35 seconds. As a result, there are a total of 4,973 phrases available for our analysis of phrase intonation. Most phrases show declining F0. The average slope of phrase curves is -2.13 Hz/syllable. This perfectly agrees with the results that we attained with the approach of parametric modeling with Stem-ML [11].

An utterance may consist of several intonation phrases whose contents are inter-related. Figure 4 shows the averaged phrase curve pattern of all utterances that consist of four phrases. It is observed that the phrase curve depends on its position. Especially, both the initial value and downshift slope of the first phrase are significantly greater than those of

succeeding phrases. F0 reset can be clearly observed at the phrase boundary.

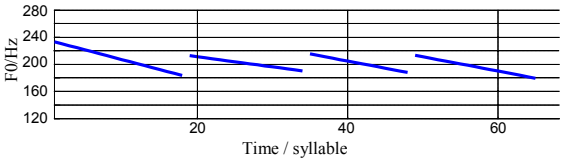

Figure 4: *Averaged phrase curve of four-phrase utterances*

5.2. Tone contours

Context-independent tone contours

For each of the six tones, an averaged contour is computed from the normalized five-point contours of all occurrences in the database. The results are shown in Figure 5. Obviously, the tone contours in continuous speech deviate greatly from their canonical patterns in isolated case. For almost all tones, the beginning section of the contour shows a much greater variation than the ending section. This suggests that tone co-articulation from the left context is more significant than that from the right one.

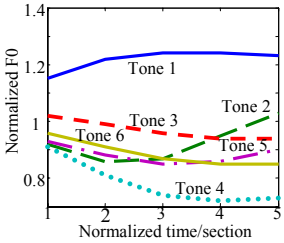

Figure 5: *Context-independent tone contours*

Co-articulated contours of disyllabic words

In order to find the templates that can carry tone co-articulation, we expand our analysis to lexical word. In particular, we focus on disyllabic words, which form the majority of the lexical words in Cantonese. All the disyllabic words with the same tone combination are grouped together and an averaged F0 contour is computed. Figure 7 shows a few examples of co-articulated tone contours of disyllabic words. In Figure 6(a), the three word contours differ from each other because of the different identities of the second tone. Figure 6(b) shows three tone combinations, which have different identities of the first tone. It can be seen that the contour corresponding to the first tone tends to resemble the context-independent single-tone case, regardless of the identity of the succeeding tone. The second tone shows a much severely co-articulated contour. It starts by following the height of the preceding tone and then gradually resumes its own position.

Cross-word contours

In this section, cross-word contours, i.e. neighboring tones locating at the boundary of two connected lexical words, are investigated. In Figure 7, the solid line draws the contour of two disyllabic words with tone combination of 1-4. It is just the direct connection of two averaged disyllabic word contours without considering any cross-word effect. The dashed line is the measured cross-word contour of tone combination 4-1. It is quite obvious that the cross-word effect plays an important role in continuous speech.

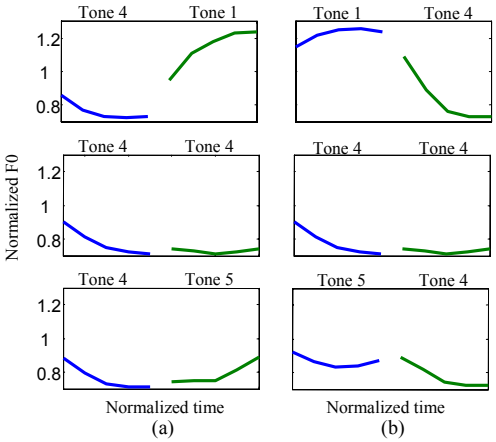

Figure 6: *Examples of co-articulated tone contours*

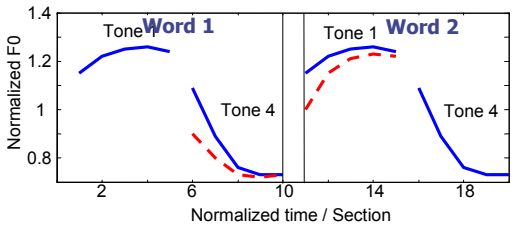

Figure 7: *Comparison of disyllabic word contour and cross-word contour for the tone combination 4-1*

Phrase-initial contours

The initial tone of a sentence or a phrase is considered to have no left neighbor. This special case is analyzed separately in our analysis. The results are shown as in Figure 8. It is very clear that phrase-initial tones have very different behavior from the context-independent tone patterns. Without a left neighbor, the beginning sections of phrase-initial tones are more voluntary to approximate the canonical target pattern.

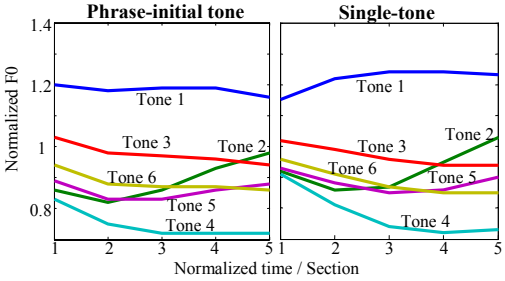

Figure 8: *Comparison of averaged phrase-initial tone and context-independent tone patterns*

6. F0 Modeling

Our baseline Cantonese TTS system is a sub-syllable based synthesizer developed at the Digital Signal Processing Lab of the Chinese University of Hong Kong [4]. This system only provided a single F0 template for each Cantonese tone.

Based on the results and observations of F0 analysis described above, a more sophisticated F0 model is developed. In this enhanced model, F0 is described at both phrase and word levels. At phrase level, F0 targets are described as linear

phrase curves. At word level, the templates include word, phrase-initial and cross-word tone contours.

To implement enhanced F0 model, the text analysis module segment a sequence of Chinese input into monosyllabic and disyllabic words. According to the tone combination of each word, the word templates are selected and connected together. Then the phrase-initial template and cross-word templates are applied to refine the connected F0 contour. The refined parts are the initial three points of initial tone, the initial three points and the last point of word tone contours. We change the F0 at word boundary especially the word initial part to capture smooth transition between words and meanwhile keep the smoothness within the word. To generate the ultimate F0 contour, the local tone contours in normalized F0 value are scaled with the respective phrase curve. Each point on the tone contour is multiplied by a syllable-dependent scaling factor, which is computed by

$$S_{p,i} = Initial_p + Slope_p \cdot i \quad (2)$$

where p represents the p^{th} phrase in a sentence and i represents the i^{th} syllable in a phrase. Finally the F0 contour is recovered to normal values and phrase curve is also implemented. At each phrase boundary, a break of fixed duration of 0.35 second is inserted.

7. Performance Evaluation

To test the performance of the enhanced F0 model, a perceptual test has been carried out. The material is 100 sentences randomly selected from newspaper. The length of sentences is from 9 to 66 characters. Over half of them are multi-phrase sentences. Each phrase contains about 10 syllables. To avoid excessive workload for the subjects and learning effect [12], the sentences are divided into 10 groups. Each tester is only allowed to access one group and each group is tested by two subjects. For each sentence, 3 different versions were generated:

Version 1 – baseline model: context-independent single-tone patterns
Version 2 – enhanced 1 model: word contour; phrase curve;
Version 3 – enhanced 2 model: word contour; phrase curve; phrase-initial tone contour; cross-word contour.

The subject is first asked to read the sentence on the screen, and then listen to the three versions of the synthetic speech, which appear in a random sequence. Each sentence is allowed to be accessed not more than three times. Finally, the subject is requested to give a Mean Opinion Score (MOS) in a five-point scale for each version.

Table 3 shows the result of the listening test. It can be seen that both versions with enhanced F0 models are statistically graded higher than the baseline. The naturalness of TTS output has been significantly improved. With the incorporation of word contours, phrase-initial contours, cross-word contours as well as phrase curves, the enhanced model Version 2, attains an MOS of 3.43, which is 0.65 higher than the baseline version.

Table 3: *Result (MOS) of listening test*

Version	Baseline	Enhanced 1	Enhanced 2
Mark	2.78	3.28	3.43

8. Conclusions

This paper has presented an investigation on the analysis of F0 contours in continuous Cantonese speech. It is found that Cantonese has a clear left-to-right control pattern. At phrase level, declining intonation is consistently observed. Co-articulated tone contours at word level and across words have been analyzed. The resulted phrase curves and context-dependent tone patterns have been utilized to establish a non-parametric model for F0 prediction. To generate smooth F0 contour for an input sentence, the statistically derived templates are integrated in a compromised way by concatenating, overlapping and adding. Subjective listening test confirms the effectiveness of the newly developed F0 model.

9. Acknowledgement

This research is partially supported by a Research Committee Funding (Direct Grant) from the Chinese University of Hong Kong and an Earmarked Research Grant (Ref: CUHK 4219/00E) from the Hong Kong Research Grants Council.

10. References

[1] Wang, C.F. et al, 2000. Analysis of fundamental frequency contours of standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation. In *Proceedings of ICSLP 2000*, 326-329.

[2] Kochanski, G.P.; Shih, C., 2003. Prosody modeling with soft templates. *Speech Communication*, Vol. 39, No.3-4, 311-352.

[3] Grimes, B.F. (Eds.), 2003. *ETHNOLOGUE: Languages of the World (14th Edition)*, http://www.sil.org/ethnologue (Internet Version), SIL International.

[4] Law, K.M.; Lee, Tan, 2001. Cantonese text-to-speech synthesis using sub-syllable units. In *Proceedings of EUROSPEECH 2001*, 991-994.

[5] Linguistic Society of Hong Kong (LSHK), 1997. *Hong Kong Jyut Ping Characters Table* (粵語拼音字表). Linguistic Society of Hong Kong Press (香港語言學會出版).

[6] Talkin, D.; Lin, Derek. ESPS/waves online documentation. Entropic Research Laboratory.

[7] Li, Y.J.; Lee, Tan; Qian, Y., 2002. Acoustical F0 analysis of continuous Cantonese speech. In *Proceedings of ISCSLP 2002*, 127 - 130.

[8] Li, Y.J., 2003. *Prosody Analysis and Modeling for Cantonese Text-to-Speech*. MPhil. Thesis, Department of Electronic Engineering, the Chinese University of Hong Kong.

[9] Holm, B.; Bailly, G., 2000. Generating prosody by superposing multi-parametric overlapping contours. In *Proceedings of ICSLP 2000*, 203-206.

[10] Dong, M.; Lua, K.T., 2002. Pitch contour model for Chinese text-to-speech using CART and statistical method. In *ICSLP 2002*.

[11] Lee, Tan et al, 2002. Modeling tones in continuous Cantonese speech. In *Proceedings of ICSLP 2002*, 2401-2404.

[12] Zhang, J.L.; Dong, S.W.; Yu, G., 1998. Total quality evaluation of speech synthesis systems. In *Proceedings of ICSLP 1998*, 60-63.