MODELING TONES IN CONTINUOUS CANTONESE SPEECH

Tan Lee [1], *Greg Kochanski* [2], *Chilin Shih* [2] *and Yujia Li* [1]

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

[2] Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill
NJ 07974, U.S.A.
{gpk,cls}@bell-labs.com

ABSTRACT

Cantonese is a major Chinese dialect with a complicated tone system. This research focuses on quantitative modeling of Cantonese tones. It uses Stem-ML, a language-independent framework for quantitative intonation modeling and generation. A set of F_0 prediction models are built, and trained on acoustic data. The prediction error is about 11 Hz or 1 semitone. The resulting optimal model parameters are analyzed in accordance with linguistic knowledge. Key observations include: (1) There is no obvious advantage to model the entering tones separately. They can be considered as simply truncated versions of the non-entering tones; (2) Cantonese appears to have a declining phrase intonation; (3) Tones at initial positions of a phrase or a sentence tend to have a greater prosodic strength than those at the final positions; (4) Content words are stronger than function words; (5) Long words are stronger than short words.

1. INTRODUCTION

Automatic intonation generation is an important yet difficult problem in text-to-speech (TTS) research. Indeed, the exact acoustic realization of fundamental frequency (F_0) in natural human speech is determined by a number of physiological and linguistic as well as extra-linguistic factors. The task of intonation modeling is to characterize, preferably in a quantitative way, the effect of individual factors, and the interaction between them.

Recently, Kochanski and Shih have developed a language-independent framework, named Stem-ML, for prosody description and prediction [1]. Stem-ML has been successfully applied to the modeling of tones and intonation in Mandarin Chinese [2,3], and the synthesis of speaking and singing styles [4]. In the present paper, the use of Stem-ML is extended to the modeling of Cantonese tones.

Cantonese is a major Chinese dialect spoken by over 60 million people in Hong Kong, Southern China and overseas Chinese communities. Cantonese is well known for its complicated and interesting tone system. Previous work on Cantonese TTS was focused mostly on acoustic synthesis or waveform concatenation techniques [5,6,7]. There has been very little progress in intonation modeling and prediction for Cantonese. This paper describes our first attempt at quantitative F_0 prediction for continuous Cantonese speech.

2. TONES IN CANTONESE

2.1. Tone System

Like Mandarin, spoken Cantonese is seen as a string of monosyllabic sounds. Each Chinese character is pronounced as a monosyllable that carries a specific tone. A character may have multiple pronunciations, and a syllable typically corresponds to a number of different characters. A Cantonese syllable is divided into the *Initial* part and the *Final* part. There are 19 *Initials* and 53 *Finals* in Cantonese, in contrast to 23 *Initials* and 37 *Finals* in Mandarin.

Cantonese is often said to have nine citation tones that are characterized by different pitch patterns as shown in Figure 1. Cantonese preserves all of the tonal categories of Middle Chinese, namely, *ping* ("level"), *shang* ("rising"), *qu* ("going" or "departing") and *ru* ("entering"). The so-called "entering" tones occur exclusively with "checked" syllables, i.e. syllables ending in an occlusive coda [p], [t] or [k]. They are contrastingly shorter in duration than the "non-entering" tones. In terms of the pitch height, each entering tone coincides with a non-entering counterpart. In many transcription schemes, only six distinctive tones, labeled by numerals 1 to 6, are used [8]. In Sections 5.1 and 5.2, we will compare intonation models that employ six-tone and nine-tone classification respectively, to see if the three entering tones are indeed equivalent to their longer counterparts.

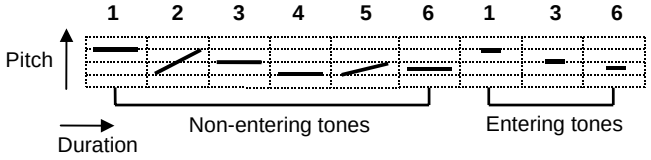


Figure 1. Tones in Cantonese: schematic description

2.2. Acoustic Realization

In spoken Chinese, tone is manifested in the F_0 movement across the voiced portion of a syllable. In Cantonese, the *Final* segment can be regarded as voiced while the *Initial* is either voiced or unvoiced.

Figure 2 depicts the average F_0 contours of the nine tones produced by a male subject who speaks native Cantonese. The averages were computed over 1,800 monosyllabic utterances that cover most of the tonal syllables used in today's Cantonese.

In Cantonese, four of the non-entering tones (Tones 1, 3, 4 and 6) have flat or slightly declining F_0 patterns while the other two (Tones 2 and 5) show different rates of rise of F_0.

Discrimination among these tones relies much more on the relative height than the shape of F_0 profiles. This is unlike Mandarin, where the four main tones have quite distinctive shapes of F_0 profiles, *i.e.* (high) level, rising, falling-rising and falling.

Figure 3 shows the F_0 contour of an example utterance. It is a Cantonese spoken digit string "43969705214758". The speaker is female and her pitch range is approximately 200 – 300 Hz. Each Cantonese digit is a tonal syllable. The F_0 profiles in the continuous utterance deviate considerably from the "canonical" patterns given as in Figures 1 and 2. For instance, the utterance has four occurrences of Tone 1, corresponding to the digits "3", "7", "1" and "7". Their F_0 values are 287 Hz, 280 Hz, 264 Hz and 256 Hz, respectively, showing a noticeable sentential down-drift. The influence of tonal context is also substantial. The F_0 profiles of Tone 1 carried by the second and the twelfth digits are alike since they have similar neighboring tones, but the other two occurrences of Tone 1 exhibit fairly different F_0 movements.

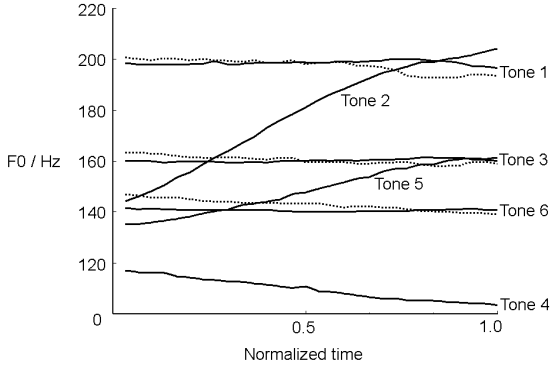


Figure 2. F0 profiles of different tones uttered by a male speaker. The dashed lines are derived from the respective entering tones

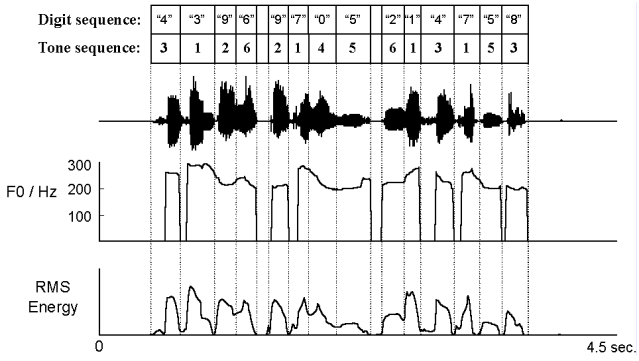


Figure 3. A Cantonese digit string uttered by a female speaker

3. STEM-ML

Stem-ML abbreviates Soft Template Mark-up Language [1]. Superficially, it is a tagging system for intonation mark-up and specification. More importantly, Stem-ML implements an underlying mathematical model that translates the mark-up tags into quantitative intonation, *i.e.* F_0 contour. Each Stem-ML tag defines a set of parameters for the mathematical model. Together, the parameters and the model generate F_0 deterministically.

In Stem-ML, the pitch contour is generated from the concatenation of local tones. Each tone type is represented by a *soft template*. By "soft", we mean that the accent template is subject to substantial modification caused by its nearby accents and/or the phrasal trend. The degree of modification is controlled by a parameter, named *strength*. Such a concept of "prosodic strength" is related how carefully the accent should be articulated. The F_0 contour is calculated as a balance between minimizing speech effort and accuracy of the communication of the tone [2]. The balance is controlled by a weight that corresponds to the prosodic strength of a syllable. If the strength is large, the F_0 contour will follow the tone's specification accurately, while if the strength is small, minimizing speech effort will be most important, and the tone will compromise with its context.

Stem-ML tags and their associated parameters are cleanly divided into local and global ones. The local parameters are attached to tones or other types of local accents while the global parameters are used to represent speaker-specific information [1]. A Stem-ML model is built on a particular speech corpus. Tone types and locations need to be marked on each utterance, using Stem-ML tags. The best values of model parameters are determined in an iterative way to minimize the difference between the predicted F_0 contour and the acoustic data. This is done by a Levenberg-Marquardt algorithm with numerical integration [2].

4. CANTONESE TONE MODELING

4.1. Design of the Models

The design of our Cantonese models largely follows the Mandarin models as described in [2]. Some of the important points are summarized below:

1) Each lexical tone class is described by one template. That is, syllables with the same tone are all described by the same template. Each template consists of 5 pitch values.

2) Each word occurrence is assigned a *strength* parameter. For each syllable (or equivalently each tone occurrence) in a word, the strength is derived by assuming a metrical pattern that is defined only by the word length. Let S_w be the word strength and $M_{L(w),i}$ be the metrical weight of the i^{th} syllable in the word w, which has the length $L(w)$. The actual strength of the i^{th} syllable, $s_{w,i}$, is given by,

$$s_{w,i} = S_w \cdot M_{L(w),i}$$

Such a design facilitates the analysis of individual word strengths without using an excessive number of parameters.

3) The phrase intonation is assumed to be a straight line. It is represented by one point located at the beginning and one at the end of the phrase.

4) Important global parameters include
ctrshift – offset of template's center from syllable's center
wscale – length of template relative to syllable length
add – nonlinearity in the mapping between perceived pitch and measured F_0
base – baseline pitch (in Hz)
smooth – smoothing time of pitch curve (in seconds)
Detailed explanation and usage of Stem-ML tags can be found in [1].

Two different models were investigated. Being denoted by T6 and T9, they adopt the 6-tone and the 9-tone classification respectively. In all cases, model parameters were initialized randomly, though restricted into "reasonable" ranges.

4.2. Speech Data

To facilitate the research on Cantonese intonation, a large-scale speech corpus was developed at the Digital Signal Processing Laboratory of the Chinese University of Hong Kong. The corpus contains 1,300 continuous speech utterances recorded from a female speaker with professional training in narration. All utterances were manually transcribed into syllable-level pronunciations. Sub-syllable time alignment was done by HMM forced alignment.

Only a small portion of the corpus is used in our modeling experiments. It consists of 30 utterances, which are divided into 3 sets. The contents of these data sets are summarized as in Table 1. Most of our experiments used only Set A. The other two data sets were for cross validation of the results. Word boundaries were marked manually on the Chinese text. Pitch extraction was done automatically using the "get_f0" program of the ESPS software [9], followed by manual inspection.

	Set A	Set B	Set C
No. of phrases*	38	41	40
No. of words	229	225	209
No. of syllables	431	410	389
Total duration	94.1 sec.	89.5 sec.	87.0 sec.
No. of F_0 data points**	5,179	4,859	4,614
Mean F_0 value (range)	191.7 Hz (118 - 346 Hz)	192.1 Hz (119 - 337 Hz)	195.6 Hz (123 - 353 Hz)

* Phrases are defined as those separated by a perceivable pause
** Only voiced frames with valid F_0 values are used.

Table 1. A summary of the three sets of speech data

5. RESULTS & DISCUSSION

5.1. Accuracy of F_0 Fitting

For the T6 models, the RMS errors of F_0 prediction are 11.7 Hz, 10.5 Hz and 11.9 Hz for data set A, B and C respectively. Figure 4 gives the fitting results of two phrases. Similar to the modeling of Mandarin, the accuracy is best for smaller or slower pitch excursions. Large and fast pitch excursions are less accurately captured. These errors are partially caused by a contribution from segmental effects. If segmental effects were properly modeled or removed, such errors would be smaller.

For the T9 models, the attained prediction errors are 11.0 Hz and 9.3 Hz for Set A and B respectively. It seems that using a finer tonal classification is slightly advantageous. But this result is by no means conclusive. We examined the fitting results of T6 and T9 models for each utterance in Set A and didn't see any significant difference between the two models.

5.2. Tone Shapes

Figure 5(a) and 5(b) give the tone shapes that are derived from a T6 model and a T9 model. They are considered to be "canonical" tone shapes in the sense that the contextual effect is averaged out. The modeled tones preserve the relative pitch levels of the isolated tones (see Figure 2). This is in line with our previous experience in automatic recognition of Cantonese tones [8]. We strongly believe that perception of Cantonese tones relies more on the relative pitch levels than on the pitch movement.

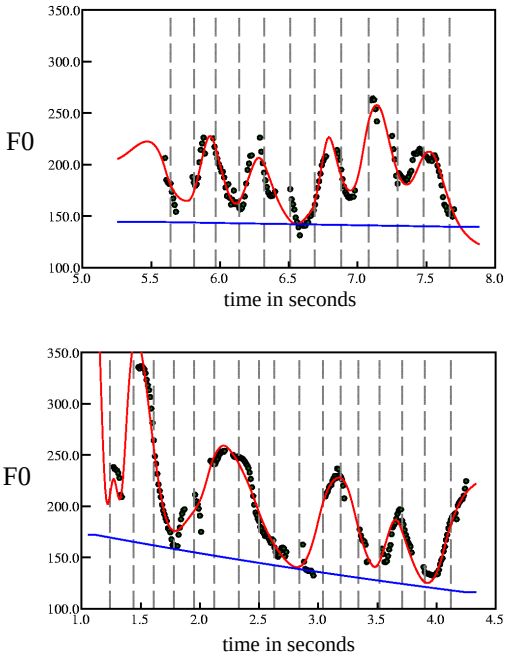
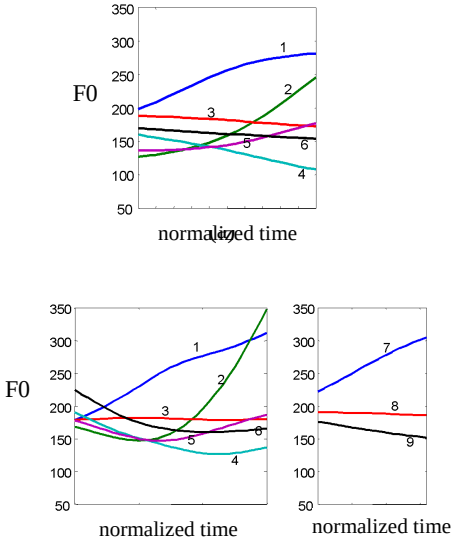


Figure 4. Examples of the F0 fitting results. Solid curves are the predicted F0 curves while the filled circles are the acoustic targets. Solid straight lines indicate the phrase intonation. Vertical dashed lines indicate the centers of the tones.



(b)

Figure 5. Tone shapes resulted from (a) a T6 model, and (b) a T9 model

Except for the Tone 1, the modeled tone shapes look very similar to the canonical shapes. The Tone 1 contour shows a substantial pitch rise. We suspect that this is a speaker specific

characteristic. Additionally, this may be partially caused by an uneven distribution of tonal context in the corpus. Among all Tone 1 occurrences in Set A, about 65% have low-pitch tones, *i.e.* Tone 3, 4, 5 and 6, as their left context, which we expect would pull down the left edge of the tone template.

As shown in Figure 5(b), the T9 model produces nine tone shapes, among which 1 to 6 are non-entering tones and 7 – 9 are entering tones. It can be seen that the entering tones are at the same pitch level as their non-entering counterparts. Thus we see little reason to represent Cantonese with the larger set of nine tones: the difference between entering tones and non-entering tones is primarily duration.

5.3. Phrase Curves

Figure 6 shows the typical phrase curves in an utterance from Set A. Among all of the phrases in Set A, 85% are declining. The overall average slope is -10.9 Hz per second.

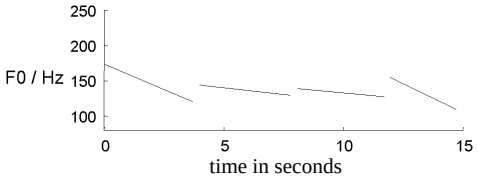


Figure 6. Phrase curves produced by a T6 model

5.4. Analysis of Prosodic Strength

In this section, we analyze the fitted strength values to see if they can be connected with linguistic features of the text. The analysis of the strength parameters was focused on:
1) The relation between strength and word position;
2) The relation between strength and part-of-speech;
3) The relation between strength and word length.

The results are given as in Figure 7(a) – (d). A word at the beginning of a sentence tends to be stronger than that at the middle or the end. Within a phrase, phrase-initial and phrase-middle words are comparably strong. A phrase-final word is very weak.

Content words such as nouns, verbs and adverbs found to be the strongest, and function words such as conjuncts are weak. This result is expected. Unexpectedly, the strengths of particles are in the middle of the range. A more detailed analysis is needed to find out the reasons behind this.

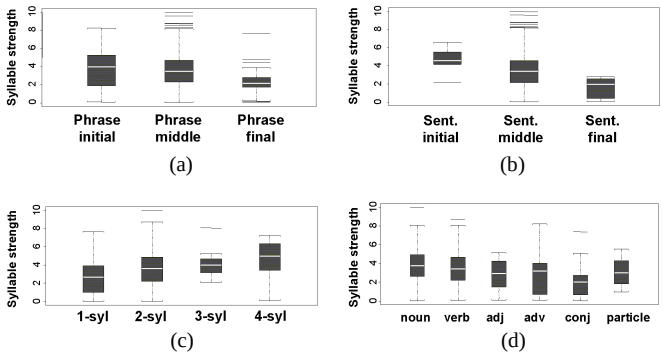


Figure 7. Analysis results of prosodic strength

Longer words are generally stronger than the shorter ones. The same observation was made for Mandarin.

As for the metrical pattern, it is observed that, in a two-syllable word, the first syllable is stronger than the second one. In a three-syllable word, the first two syllables are at the same strength level while the ending syllable is relatively weak.

6. CONCLUSIONS

In this paper, we fitted the F_0 contour of natural sentences in Cantonese. A model with approximately one freely adjusted parameter per word, representing its prosodic strength, attains a fitting accuracy of approximately 11 Hz, or 1 semitone. We compared the 6-tone classification against the 9-tone one, and saw no obvious advantage to the more complex tonal system: both worked essentially identically. Consequently, we believe that the entering tones should be considered as truncated versions of the non-entering tones.

Many of the properties of Cantonese seem similar to Mandarin: syllables tend to be stronger at the beginning of sentences and phrases, and weaker at the ends; longer words tend to be prosodically stronger; and content words tend to be stronger. Unlike Mandarin, Cantonese appears to have a declining phrase curve on most phrases.

7. ACKNOWLEDGEMENT

This research is substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK4291/00E). The first author would like to thank Dr. Chin-Hui Lee and Dr. Joseph Olive of Bell Laboratories, for their great support that made this collaborative work possible.

8. REFERENCES

[1] Kochanski, G. P. and Shih, C., "Prosody modelling with soft templates," accepted for publication by *Speech Communication*.

[2] Kochanski, G. P. and Shih, C., "Automatic modelling of Chinese intonation in continuous speech," in *Proceedings of EUROSPEECH 2001*, pp.911-914.

[3] Yuan, J., Shih, C. and Kochanski, G. P. "Comparison of declarative and interrogative intonation in Chinese," to appear in *SPEECH PROSODY 2002*.

[4] Shih, C. and Kochanski, G. P., "Prosody control for speaking and singing styles," in *Proceedings of EUROSPEECH 2001*, Vol.1, pp.669-672.

[5] Lee, Tan et al., "Micro-prosodic control in Cantonese text-to-speech synthesis," in *Proceedings of EUROSPEECH 1999*, Vol.4, pp.1855-1858.

[6] Fung, T. Y. and Meng, H., "Concatenating syllables for response generation in spoken language applications," in *Proceedings of ICASSP 2000*, Vol.2, pp.933-936.

[7] Law, K.M. and Lee, Tan, "Cantonese text-to-speech synthesis using sub-syllable units," in *Proceedings of EUROSPEECH 2001*, Vol.1, pp.991-994.

[8] Lee, Tan et al., "Using tone information in Cantonese continuous speech recognition," to appear in *ACM Transactions on Asian Language Information Processing*.

[9] Talkin, D. and Lin, Derek, "ESPS/waves online documentation", Entropic Research Laboratory.