# HAKKA PITCH-CONTOUR PARAMETER GENERATION USING A MANDARIN-TRAINED PITCH-CONTOUR MODEL

*Hung-Yan GU and Shiue-Jen LI*

Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei
guhy@mail.ntust.edu.tw http://www.csie.ntust.edu.tw

## ABSTRACT

In this paper, using an existing pitch-contour model of a Chinese dialect (Mandarin here) to generate pitch-contour parameters for another dialect's sentences (Hakka here) is studied. This can be generally viewed as a pitch-contour model adaptation problem. We study this problem in hope to save tedious labors and research time needed to build a pitch-contour model for a specific Chinese dialect. This approach is also useful for synthesizing speech of a weak dialect to help reserve it from disappearance. In this study, we have built a prototype Hakka speech synthesis system. Except the pitch-contour parameters, the other prosodic parameters are generated by rules. For signal-waveform synthesis, TIPW method previously proposed is adopted. Two sets of Hakka and Min-Nan sentences are synthesized, respectively, for evaluation experiments. Syllable signals for Min-Nan are borrowed from the recorded Hakka syllables, and pitch-contour parameters are generated using the same approach. Because of timing difference and different difficulties in sentence contents, the comprehension and naturalness-level scores for Hakka synthetic speech only reach 91.87% and 79.5%, respectively. But for the set of Min-Nan synthetic speech, the scores obtained are 97.1% and 85.5%, respectively.

## 1. INTRODUCTION

In Chinese speech synthesis, prosodic parameters are known to be the key factors that determine the naturalness level of a synthetic sentence [1, 2]. Among the different kinds of prosodic parameters (duration, amplitude, pitch-contour, and syllable-preceding pause), pitch-contour for a syllable is the most important one. Therefore, many researchers have studied to develop a general and good pitch-contour model [3, 4, 5, 6]. Also, it is intended that this model can be equally applied to most of the Chinese dialects. Note that most Chinese dialects (e.g., Mandarin, Hakka, Min-Nan, Guang-Dong, etc.) are syllable prominent, and are tone languages.

Here, we consider furthermore the problem whether a pitch-contour model (or duration model, etc.) trained with spoken sentences of a Chinese dialect can be adapted directly to generate pitch-contour parameters for another dialect's sentences. That is, spoken sentences of the target dialect need not be acquired for model adaptation. This problem is meaningful and important because tedious labors and enormous research time/budget can be saved if the adapted model's performance is acceptable. Otherwise, it will be necessary to do the laborious works, defining sentence contents, recording sentences of a dialect, verifying segment boundaries and pitch peaks that are automatically or manually labeled, etc. In addition, there are many Chinese dialects that are mother tongues of small groups of people, and are facing disappearance. Since these dialects are used by small groups and have no commercial profit, almost no researchers are willing to study computer speech synthesis for them. Therefore, studies on adapting a popular dialect's pitch-contour model (or other parameter model) to generate a weak dialect's pitch-contour parameters can have contribution to help reserve weak dialects from disappearance.

In this paper, a direct adaptation method is studied and proposed, which makes use of a Mandarin-trained pitch-contour model to generate pitch-contour parameters for a Hakka sentence. The Mandarin pitch-contour model used here is originally constructed in previous study to generate sentence-pitch-contour parameters for Mandarin sentences [6]. It may be suspected how can a Mandarin pitch contour model be used to generate Hakka sentences' pitch contours since Hakka has more lexical-tones than Mandarin. The method developed will be explained in details in next section. Here, note that Hakka is a family and includes several sub-dialects, e.g., Hoi-Liuk (海陸), Si-Rhan (四縣), Ta-Pu (大埔), etc. These sub-dialects have most of their base-syllables (disregarding tones) in common, but different in number of lexical-tones and tone-shape assignments to Chinese characters. For example, the character "家" (home), of phonetic representation /ga-1/, is pronounced with falling tone in Hoi-Liuk but pronounced with rising tone in Si-Rhan. Actually, the tone-shape assignments in Hoi-Liuk and Si-Rhan are always complementary as explained above for the character "家". Also, the number of lexical-tones in Hoi-Liuk is seven but is only six in Si-Rhan. In this study, Hoi-Liuk is selected to represent Hakka and be the target for adapting the Mandarin trained pitch-contour model because it has more lexical-tones and more base-syllables. If an adaptation method can be developed for Hoi-Liuk, we think this method can also be applied to the other sub-dialects.

## 2. PITCH-CONTOUR MODEL ADAPTION

We had previously constructed a Mandarin sentence-pitch-contour model that performs well and is ready for use [6]. However, it is hard to train an intrinsic Hakka pitch-contour model because lots of time and labors must be spent again, and a good cooperative and native Hakka speaker is very difficult to look for. Therefore, we are motivated to study if the Mandarin model can be used directly (without acquiring any training sentences) to generate pitch-contour parameters for Hakka sentences. If this approach performs well, tedious labors and research time/budget can then be saved. In general, using an existing model to generate pitch-contour parameters for another dialect's sentences can be viewed as a model-adaptation problem. However, in this study, we do not change any model parameter value of the Mandarin pitch-contour

model. The adaptation is made in the reverse direction. That is, the comprising syllables' lexical-tones of a Hoi-Liuk Hakka sentence are first mapped to their corresponding Mandarin lexical-tones, and the mapped Mandarin tones are then used as input to the Mandarin pitch-contour model. The reason why this adaptation method is workable will be explained in Section 2.1 while the details of the tone-mapping rules are described in Section 2.2.

### 2.1 Context Dependent Pitch-Contour Shapes

Mandarin has five lexical-tones. They are distinguished by order numbers assigned to them (i.e., Tone 1, 2, 3, 4, 5), or by tone-shape derived names, i.e., high-level, mid-rising, dip-falling-and-rising, and high-falling tones. Such shape-derived names can only portray the pitch-contour shape of a syllable when it is uttered isolatedly. These names are inappropriate to portray the pitch-contour shape of a syllable uttered within a sentence. For example, the pitch-contours of Tone 4 uttered under different contexts have very different shapes as shown in Fig. 1. The three shape classes shown in Fig. 1 are obtained by

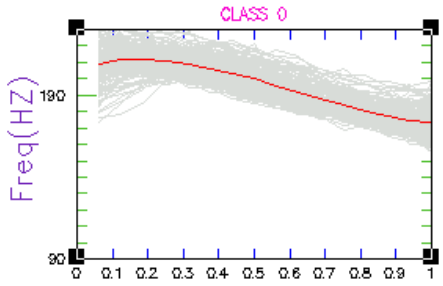


**Fig. 1(a) First class.**

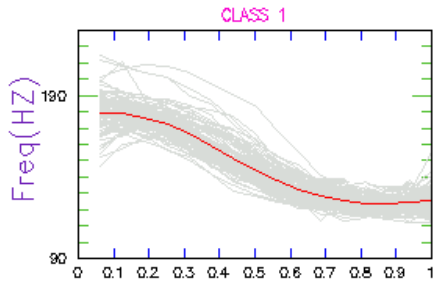


**Fig. 1(b) Second class**

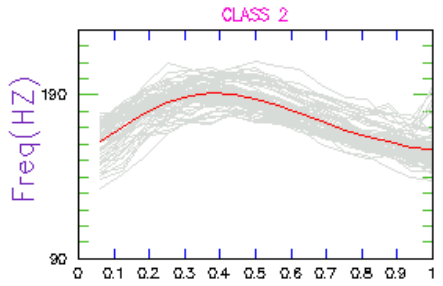


**Fig. 1(c) Third class.**

**Fig. 1 Three classes of pitch-contour shapes for Tone 4.**

vector quantizing the collected pitch-contours of Tone 4 uttered within different sentences [6]. Note that all the pitch-contours in Fig. 1(c) rise first and then fall. Therefore, the name "high-falling" is not appropriate for this shape class. In addition, although the two shape classes shown in Fig. 1(a) and 1(b) are of similar trend (falling), however, their absolute frequency heights are very different, one higher and one lower (high-falling vs. mid-falling). In condition of these kinds of differences, all of the pitch-contours are still perceived as of Tone 4. This indicates that people can by-pass the context-affected boundary parts and extract out the central part of a syllable's pitch-contour. Also, people can tolerate the difference in frequency height and pick out the shape-trend to recognize the carried lexical-tone correctly.

According to the observations above, we think that slight differences in frequency-height or at boundary-parts need not be worried for correct recognition of the carried lexical-tone. Furthermore, we think it is feasible to approximate a pitch-contour shape used in a target dialect (e.g., Hakka) with a shape class, of similar shape trend in central part, trained in a working dialect (e.g., Mandarin). Since there are usually several shape classes (e.g., 8) trained for each lexical-tone of the working dialect, we can hence select the class that is most similar, in frequency-height and boundary-parts, to the target pitch-contour shape to be approximated. Then, the possible decrease in naturalness level caused by slight difference in frequency-height and boundary-parts can be minimized. In this study, we propose a practical way to do such an approximation, or called pitch-contour model adaptation. The details are as the following. First, lexical-tone mapping rules between the target dialect, Hoi-Liuk, and the working dialect, Mandarin, are designed according to lexical-tone shapes in isolated uttering. Using these lexical-tone mapping rules, the comprising syllables' lexical-tones of a Hoi-Liuk sentence can then be mapped to their corresponding Mandarin lexical-tones. After lexical-tone mapping, the sequence of Mandarin lexical-tones and other information (e.g., word boundary marks) are put into the Mandarin pitch-contour model to generate the pitch-contour parameters for the target Hoi-Liuk sentence. In practice, it is encountered there is a Hoi-Liuk lexical-tone that cannot find a Mandarin lexical-tone with ideally similar tone shape. Therefore, we have to study to remedy (or post-process) the generated pitch-contour parameters in order to obtain better prosodic presentation. The details of the remedy processing are described in Section 2.3.

### 2.2 Lexical-Tone Mapping Rules

There are totally seven lexical-tones in Hoi-Liuk Hakka. Two of the seven are abrupt tones that always accompany syllables that have stop-ended consonant. Also, abrupt tone syllables are pronounced in shorter duration than ordinary lexical-tones. Tone numbering, tone-shape names, and relative tone-height values for the seven tones are as listed in Table 1. Note that in convention the tone number 6 is not used whereas the number 8 is used. Also, the tone-height values of the two abrupt tones

**Table 1 The seven lexical-tones of Hoi-Liuk Hakka**

| **Tone number** | 1 | 2 | 3 | 4 | 5 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| **Shape name** | **Fall-ing** | **Ris-ing** | **Low-dip** | **High-abrupt** | **High-level** | **Mid-level** | **Low-abrupt** |
| **Tone-height value** | 53 | 24 | 21 | 55 | 55 | 33 | 21 |
| **Example syll. & char.** | Fu-1 夫 | fu-2 虎 | fu-3 富 | fuk-4 福 | fu-5 湖 | fu-7 護 | fuk-8 復 |

(numbered 4 and 8) are not of distinct pitch-contour shapes but pronounced in shorter duration due to stop-consonant ending. Therefore, the high-abrupt tone (numbered 4) can be viewed as a variant of the high-level tone (numbered 5) with shorter duration (about three forth of an ordinary tone's duration). Similarly, the low-abrupt tone (numbered 8) can be viewed as a variant of the low-dip tone (numbered 3) with shorter duration.

At first glance, it may be questioned how can the larger set of seven lexical-tones of Hoi-Liuk Hakka be mapped to the smaller set of five lexical-tones of Mandarin. But according to the observation above, we know that only five lexical-tones excluding the abrupt tones need to be further considered. First, the falling tone (numbered 1) has distinct shape, "falling". It can be ideally mapped to the forth tone of Mandarin. Next, the rising tone (numbered 2) has distinct shape, "rising", and can be ideally mapped to the second tone of Mandarin. As to the low-dip tone (numbered 3), it has similar shape to the leading half of the third tone of Mandarin (tone-height value 214 if uttered isolatedly). However, the third tone of Mandarin will just have tone-height value, 21, if uttered with another leading or following syllable. Therefore, it is still ideally to map the low-dip tone of Hoi-Liuk to the third tone of Mandarin. Next, consider the high-level tone (numbered 5). It is apparent this tone can be ideally mapped to the first tone of Mandarin because both of them are of the same shape, "high-level". Last, consider the mid-level tone (numbered 7). Unluckily, it has no apparent correspondent tone in Mandarin. Thus, we consider to map it to the neutral tone or third tone of Mandarin according to its tone-height value, 33. To determine which is better to choose, we resort to perceptual evaluation. According to evaluation of synthesized Hoi-Liuk sentences, we find that the mapping to the third tone of Mandarin is better. This can be attribute to the more varying tone-height of the neutral tone. Nevertheless, the difference in tone-shape trend between Hoi-Liuk mid-level tone and Mandarin third tone can still be felt sometimes, level vs. dip. Therefore, we study this problem and develop a remedy method as described in Section 2.3. In summary, the lexical-tone mapping rules from Hoi-Liuk Hakka to Mandarin are as given in Table 2.

**Table 2 Tone mapping rules from Hoi-Liuk to Mandarin**

| **Hoi-Liuk tone number** | 1 | 2 | 3 | 4 | 5 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| **Mandarin tone Mapped** | 4 | 2 | 3 | 1 | 1 | 3 | 3 |

### 2.3 Pitch-Contour Shape Remedying

In preliminary evaluation of the synthesized Hoi-Liuk Hakka sentences, we find almost of the generated pitch-contours sound good with certain level of naturalness. However, the generated pitch-contours for the mid-level tone (numbered 7) are sometimes perceived as slightly dipping rather than leveling, and thus leads to strange feeling although it does not cause erroneous recognition of the carried lexical-tone. Therefore, we begin to study this problem, and find out an effective but simple remedy method. In this method, the generated pitch-contour parameters for the mid-level tone are post-processed in the third step. The detailed processing steps are as following. (1) First, the mid-level tone is still mapped to Tone 3 of Mandarin as the other Hoi-Liuk tones are mapped to their correspondent Mandarin tones. (2) The mapped Mandarin tone sequence and word-boundary mark information are put into the Mandarin pitch-contour model to generate pitch-contour parameters. (3) The generated pitch-contour parameters for each mid-level tone are remedied. That is, rotate the pitch-contour in counter-clock direction to decrease the shape-angle from dipping-like to more leveling-like. This shape-angle decreasing operation is as illustrated in Fig. 2.

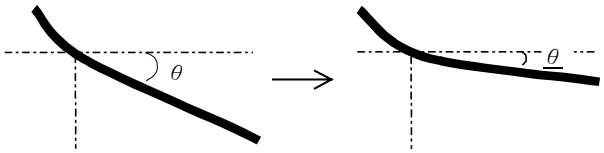


**Fig. 2 Shape-angle decreasing**

In our speech synthesis system, pitch-contour parameters for a syllable are a frequency vector of 16 dimensions. Each dimension represents a normalized time point [6, 9]. Therefore, it is straightforward to perform such kind of shape-angle decreasing operation. Suppose ($f_0$, $f_1$, …, $f_{15}$) is the pitch-contour frequency vector for a mid-level tone syllable. We first find the peak value among $f_3$, …, $f_{15}$. The dimensions, $f_0$, $f_1$, and $f_2$, are not considered because they are very probable at boundary transition part. Then, the peak frequency value, $f_p$, is used to decrease the shape angle according to Equation (1).

$$f_p = \underset{3 \le k \le 15}{MAX} \; f_k$$

$$\overline{f_k} = \left( \frac{(f_k / f_p) - 1}{3} + 1 \right) \times f_p \;, \quad k = 3, \ldots, 15 \qquad (1)$$

With the pitch-contour shape remedying processing, the synthesized Hakka syllables of mid-level tone are now perceived normal and better in naturalness-level. That is, the tone-shape trend, level, is felt all times rather than the trend, dip, sometimes.

## 3. SYNTHETIC SPEECH EVALUATION

Since the pitch-contour model used here is entirely trained by Mandarin spoken sentences, others may naturally question if such a model can be used to generate pitch-contour parameters for Hakka speech synthesis. To show that our approach is indeed workable, we hence decide to build a prototype Hakka speech synthesis system. Besides the pitch-contour parameters, the other considered prosodic parameters are syllable duration, syllable amplitude, and syllable-preceding pause. In this study, those prosodic parameter values are generated by rules that are adopted and modified from previous studies [7, 8]. As to signal-waveform synthesis, the method, Time-Proportioned Interpolation of Pitch waveform (TIPW) [9, 10], is adopted and improved here. Because TIPW is originally developed for Mandarin synthesis and Mandarin has no abrupt tone, therefore, we have to solve the synthesis problems, syllable-duration adjusting and final-stop waveform concatenation, for abrupt-tone Hakka syllables. According to our study, the stop-consonant waveform at the final part of an abrupt-tone syllable can be synthesized in the same way as the stop-consonant at the initial part.

Currently, only small amount (about 2,200) of multi-character words of Hakka is collected in the dictionary. Therefore, input-text to our prototype system is mainly phonetic syllable

sequence although Chinese characters are also accepted (but often translated to wrong syllables). Using this prototype system, two sets of sentences are synthesized for evaluation experiments. One set of five Hakka sentences is synthesized earlier than the other set of five Min-Nan sentences when the prototype system is still under development and suffered with bugs. Min-Nan sentences are synthesized by similarly mapping their lexical-tones to Mandarin lexical-tones, and borrowing the recorded Hakka syllables for signal-waveform synthesis processing. In source-signal collection, each Hakka syllable is pronounced only in high-level tone or high abrupt tone, and is recorded only one time. That is, the pitch-contour for each synthesized Hakka syllable is determined by the pitch-contour model and not by the recorded syllable signal.

In the evaluation of Hakka speech synthesis, 21 Hakka familiar persons are invited to listen to the synthetic sentences, write down the Chinese characters spoke, and give a score for perceived naturalness-level. Similarly, 20 other Min-Nan familiar persons are invited to evaluate the synthetic Min-Nan sentences. Due to timing difference in synthesis-program execution, different difficulties in sentence contents, and using of a notebook computer's poor speaker for playing, the comprehension rate and naturalness-level score for Hakka synthetic sentences only reach 91.87% and 79.5%, respectively. But for Min-Nan synthetic sentences, the comprehension rate and naturalness-level score are, however, 97.1% and 85.5%, respectively. Besides off-line evaluation, we have set up a web page for on-line evaluation through Inter-network. That is, those who are interested in our prototype system can browse the web site **http://guhy.ee.ntust.edu.tw/hakka/** and input a phonetic syllable sequence. Then, the sequence will be processed to synthesize out Hakka speech signal immediately.

## 4. CONCLUSION

In this study, a simple but effective approach is proposed for pitch-contour model adaptation. Actually, a pitch-contour model originally trained by and for Mandarin is adapted to generate pitch-contour parameters for Hakka speech synthesis. The way of adaptation is simple because no Hakka spoken sentences need to be recorded beforehand to do the adaptation processing, and no model parameter value modification is made. This approach is also effective. When the adapted pitch-contour model is used to generate the pitch-contour parameters, the synthesized Hakka and Min-Nan sentences are both evaluated to have acceptable comprehension rates and naturalness-level scores. More detailed, the adaptation is done in the reverse direction. The lexical-tones of Hakka syllables in a sentence are mapped to their corresponding Mandarin lexical-tones. Because one of the Hakka lexical-tones cannot be ideally mapped, we therefore study and propose a remedy method to post-process this lexical-tone's pitch-contour parameters. Afterward, the generated pitch-contour shapes for all Hakka lexical-tones are perceived well.

## 5. REFERENCE


[1] Shih, Chilin and Richard Sproat, "Issues in Text-to-Speech Conversion for Mandarin", Computational Linguistics & Chinese Language Processing, Vol. 1, No. 1, pp. 37-86, 1996.

[2] Wang, Ren-Hua. "Overview of Chinese Text-to-Speech System", International Symposium on Chinese Spoken Language Processing (ISCSLP'98), Singapore, 1998.

[3] Lee, L. S., C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System", IEEE trans. Speech and Audio Processing, Vol. 1, pp. 287-294, 1993.

[4] Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE trans. Speech and Audio Processing, Vol. 6, No.3, pp. 226-239, 1998.

[5] Wu, C. H. and J. H. Chen, "Prosody Generation in a Chinese TTS System based on a Hierarchical Word Prosody Template Tree", ROCLING X, Taipei, pp. 262-266, 1997.

[6] Gu, Hung-Yan and Chung-Chieh Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", International Symposium on Chinese Spoken Language Processing (ISCSLP2000), Beijing, pp. 125-128, 2000.

[7] Lee, L. S., C. Y. Tseng, and M. Ouh-Young, "The Synthesis Rules in a Chinese Text-to-Speech System", IEEE trans. ASSP, Vol. 37, No. 9, pp. 1309-1320, 1989.

[8] Chiou, H. B., H. C. Wang, and Y. C. Chang, "Synthesis of Mandarin Speech based on Hybrid Concatenation", Computer Processing of Chinese and Oriental Languages, Vol. 5, pp.217-231, 1991.

[9] Gu, Hung-Yan and Wen-Lung Shiu, "A Mandarin-syllable Signal Synthesis Method with Increases Flexibility in Duration, Ton and Timbre Control", Proc. Natl. Sci. Counc. ROC(A), vol. 22, No.3, pp. 385-395, 1998.

[10] Gu, Hung-Yan, "Notes for the Syllable-Signal Synthesis Method: TIPW", International Symposium on Chinese Spoken Language Processing (ISCSLP1998), Singapore, SS-B3, 1998.


## ACKNOWLEDGEMENT


This work was sponsored by national science council under the contract number NSC89-2218-E-011-011.