Improved Prediction of Tone Components for F_0 Contour Generation of Mandarin Speech Based on the Tone Nucleus Model

Qinghua Sun, Keikichi Hirose**, and Nobuaki Minematsu****

*Graduate School of Engineering, **Graduate School of Information Science and Technology,
***Graduate School of Frontier Sciences, University of Tokyo, Japan
{qinghua, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Improved prediction of tone components was realized in our method for synthesizing sentence fundamental frequency (F_0) contours of Mandarin speech. The method is based on representing a sentence logarithmic F_0 contour as a superposition of tone components on phrase components as in the case of generation process model (F_0 model). The tone components are realized by concatenating their fragments at tone nuclei predicted by a corpus-based method, while the phrase components are generated by rules under the F_0 model framework. In the original method, tone components are assumed to have similar shapes as F_0 contours at tone nuclei. This is based on the assumption that the phrase components are almost flat throughout an utterance. However, this is not the case especially for phrase component initials. To cope with this problem, parameters representing tone components of tone nuclei are modified. Also, predicted parameters in earlier processes are used for the prediction of following processes. Result of the listening test conducted for synthetic speech with the generated F_0 contours by our methods and also by the HMM-based method confirmed the advantage of ours, especially the improved version.

1. Introduction

Introduction of selection-based scheme in speech synthesis largely improved quality of synthetic speech. However, there still remain problems if we view from the aspect of prosodic features. Although the control of prosodic features is an important issue in speech synthesis for any languages, it comes quite critical for speech quality in the case of Mandarin. As it is well known, Mandarin is a typical tonal language and each syllable with the same phoneme constitution has up to four tone types, each indicating different meaning. Fundamental frequency (F_0) contours of utterances should include these local tonal features in addition to the sentential intonation corresponding to syntactic/utterance structures. This situation makes F_0 movements of Mandarin sentences be more complicated than non-tonal languages like English, Japanese and so on. Therefore, control of F_0 contours (together with other prosodic features) becomes an important (and tough) issue in Mandarin speech synthesis.

Benefit of corpus-based methods over rule-based methods increases when handling complicated features. Naturally, most F_0 controls adopted in Mandarin speech synthesis are corpus-based using decision trees, neural networks linear regression analysis and so on [1-3]. Among all, the hidden Markov model (HMM) [4] is now commonly used for synthesizing speech of many languages, including Mandarin, because of its ability of handling segmental and prosodic features simultaneously and concatenating speech segments in statistical basis. Flexible control of speech styles is possible by adapting HMM to a new style using a small-sized speech corpus of that style. However, it still requires a certain size to keep the speech quality. Moreover, the method handles F_0 in frame-by-frame manner, which is not appropriate for prosodic features: prosodic features cover wider spans of utterances, such as words, phrases, and so on.

A better control of prosodic features (F_0 movements) in longer units is possible using the generation process model (F_0 model), which represents a logarithmic F_0 contour as the sum of phrase and tone components on a baseline level F_b [5]. These components are assumed as responses to phrase and tone commands of the corresponding control mechanisms. By controlling these commands instead of frame-by-frame F_0 values, good F_0 control can be realized even if the training data are limited. Moreover, since the command parameters (timings and magnitudes/amplitudes) are tightly and directly related to linguistic and para/non-linguistic information, a flexible control is possible by manipulating the commands. For Japanese, we have already developed a corpus-based method of generating F_0 contours from text in the framework of the F_0 model successfully [6]. However, this is not the case for Mandarin; it is rather difficult to prepare a necessary size of prosodic corpus for training. The corpus should have the command parameters, which are rather difficult to be extracted automatically from observed F_0 contours because of complicated F_0 movements.

These considerations led us to propose a method of F_0 contour generation for Mandarin speech synthesis (Figure1), where the tone components were generated by concatenating their fragments at tone nuclei predicted by a corpus-based scheme [7], and were superposed onto the phrase components, which were generated by a rule-based scheme on the basis of F_0 model [8]. Here, "tone nucleus" is defined as a portion of syllable, which possesses a stable F_0 pattern regardless of the context [9]. To prevent mismatch between tone and phrase components, the phrase components are generated first, and then the tone components are generated taking the features of phrase component into account (two-step scheme). The validity of the method was proved through a listening experiment of synthetic speech. The most significant benefit of the method over others (without decomposition) is the flexibility in F_0 contour generation. For instance, focal position can be controlled only by manually changing phrase components [10].

Tone nucleus is defined as the flat- F_0 part of a syllable for Tone 1 (T1) and Tone 3 (T3), the rising- F_0 part for Tone 2 (T2) and the falling- F_0 part for Tone 4 (T4). Since phrase components are mostly flat as compared to tone components, we assume the tone components have F_0 movements similar to those of F_0 contour, and approximate tone components of tone nuclei as flat F_0 patterns for T1 and T3, rising F_0 patterns for T2, and falling-F_0 patterns for T4. Although this assumption is good for almost

all the cases, and the synthetic speech using the generated F_0 contour sounded natural, there are occasional degradations mostly at phrase initials. To cope with this situation, we modified the process of tone nucleus F_0 pattern generation. Henceforth, the methods before and after the modification are called as original method and improved method, respectively.

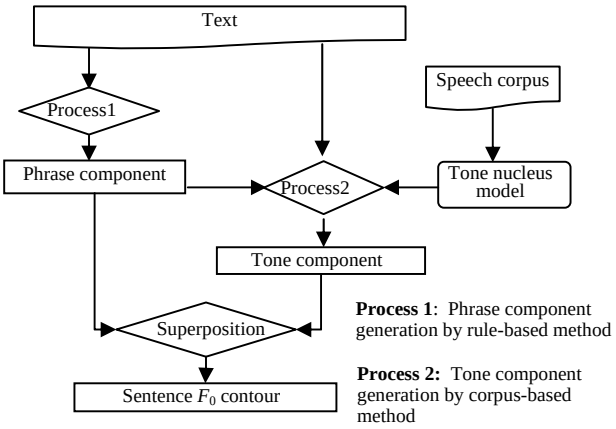

Figure 1: Two-step scheme of F_0 contour generation

The rest of the paper is organized as follows. Section 2 describes the method of phrase component generation. In Section 3, after a brief explanation on the tone nucleus, the original version for the method of tone component generation is given with its problems. After the detailed explanation of improved method also in Section 3, results of F_0 contour generation are given in Section 4. Generated F_0 contours are evaluated through a listening test of synthetic speech in the same section. Section 5 concludes the paper.

2. Generation of phrase components

A rule-based method was developed to generate the phrase components [9]. In the method, "prosodic word" is first defined as a chunk of syllables usually uttered in a tight connection: a prosodic word can be a word, a compound word, or a word chunk uttered together frequently. Then, at each prosodic word boundary, a phrase command with a certain magnitude is placed depending on the phrase component value at the boundary. Rules for placing are constructed based on the observations of 100 utterances by a female native speaker of Mandarin. Different from non-tonal languages, such as English and Japanese, tone components of Mandarin can have negative values. To prevent F_0 contour go below the baseline (F_b) even with negative tone components, phrase components should keep a certain value any time. The rules are constructed to satisfy this condition. For detail refer to [8].

3. Generation of tone components

3.1 Tone nucleus model

In Mandarin, each syllable can have up to four lexical tones, T1, T2, T3 and T4. They are characterized as high-level, mid-rising, low-dipping, and high-falling F_0 contours, respectively. Besides the lexical tones, there is also a so-called neutral tone (T0), which does not possess its inherent shape in the F_0 contour. Its F_0 contour varies largely with the preceding and following tones.

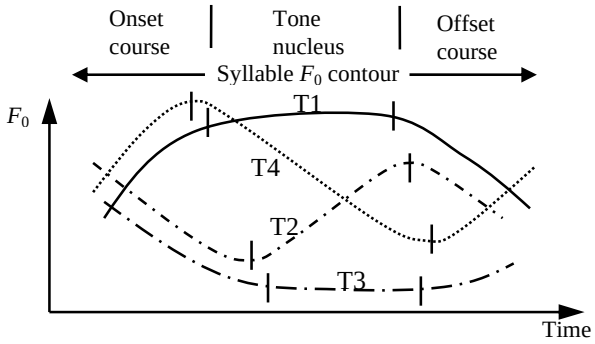

Figure 2: Tone nuclei for the four lexical tones.

For a syllable, not only its early portion but also voicing period at the ending portion is regarded as physiological transition period to/from the neighboring syllables. Based on this observation, a tone nucleus model, which divides a syllable F_0 contour into three segments according to their roles in the tone generation process, was proposed and applied to tone recognition successfully [9]. The three segments are called onset course, tone nucleus, and offset course, respectively. Only the tone nucleus is a portion where F_0 contour keeps the intrinsic pattern of the tone; the others are only the portions for physiological transitions. Figure 2 illustrates syllable F_0 contours for the four lexical tones with possible articulatory transitions. It shows how the three segments are defined on the F_0 contours. Among the three segments, only tone nucleus is obligatory, whereas the other two segments are optional; their appearance depends on voicing characteristics of initial consonant, syllable duration, context, and etc. (Since T0 does not possess its intrinsic F_0 contour, entire voiced segment is assumed as tone nucleus.)

3.2 Original method

Our method of tone component generation first predicts tone components only for tone nuclei of constituting syllables in a corpus-based way, and then concatenated them to generate an entire component for the utterance. It consists of the following processes:

1. For each syllable in the sentence to be synthesized, the onset and offset times of tone nucleus are predicted.
2. For each tone nucleus, several parameters representing the tone component are predicted. The parameters are different depending on the tone types as explained later.
3. Based on the predicted parameters, an F_0 pattern is generated for each tone nucleus.
4. The patterns are concatenated with each other to produce the entire tone components (of the speech to be synthesized). Although a smoother concatenation is possible by using such as 3rd order polynomials, they are concatenated using straight lines, because, in preliminary listening test, no clear difference was perceived in the quality of synthetic speech depending on the concatenation methods.

In the first and second steps above, the parameters are predicted using binary decision trees trained separately for each parameter. Inputs to a tree are the information extracted from input text, such as phonemic constitutions of syllables, number of

syllables in words, depths of syntactic boundaries, and so on [8]. Information on phoneme durations and pauses are also used, which may be predicted in a separate process in a total system of text-to-speech conversion.

Assuming the phrase components are mostly flat, a tone component of a tone nucleus shows a shape similar to the F_0 contour of the corresponding part. Based on this assumption, parameters for tone components of tone nuclei are defined as follows:

1. T1 and T3 are known as the "level tones," characterized by flat F_0 contours. Based on this observation, their tone nuclei are defined as portions with flat F_0 contours, each of which is represented by a single parameter, *i.e.*, average F_0 value.
2. For each of T0, T2 and T4, F_0 contours of tone nuclei are first normalized in time and frequency ranges, and then are clustered into several groups. The average contour for each group serves as a template to represent the shape of tone component of tone nucleus. The parameters include the absolute pitch range, average F_0 value, and template identity.

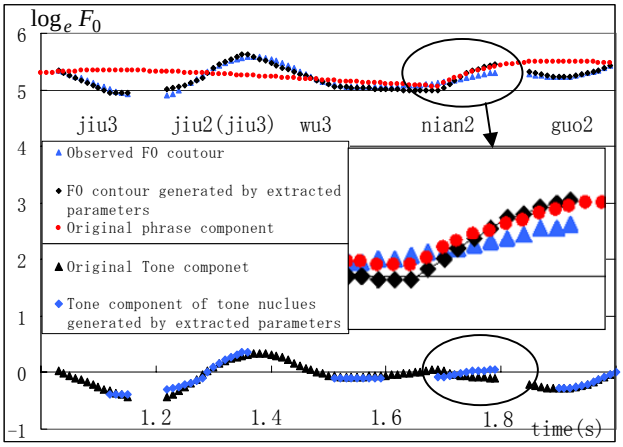

Figure 3: Example of F_0 contour for a Chinese utterance with its phrase and tone components. Only part ("jiu3 jiu2(jiu3) wu3 nian2 guo2") of the utterance is shown for visibility. The observed F_0 contour and its tone component are compared with those generated by the method. Here, "jiu2(jiu3)" indicates that the syllable "jiu" is uttered as T2 because of tone sandhi, though its original tone type is T3.

Using the same 100 news utterances used to construct rules for phrase component generation in section 2, we have trained binary decision trees for parameter prediction. For several sentences, speech synthesis was conducted using generated F_0 contours, and the listening test for synthetic speech indicated the validity of the proposed method. However, when F_0 contours are predicted directly without decomposing them into phrase and tone component, the resulting synthetic speech is evaluated to have quality similar to that by the two-step scheme (our method). One possible reason for this situation is that the two-step scheme causes occasional degradation around the phrase component initial. Figure 3 shows such an example: For the syllable "nian2" locating at the second phrase initial, the tone component of the tone nucleus shows slightly-falling feature in the observed F_0 contour, while rising feature when it is generated using parameters extracted from the observed F_0 contour. This is because no template for T2 has falling feature. This unmatched condition with observed F_0 contour may occur more frequently when dealing with speech other than the reading style (such as dialogue speech and emotional speech), where phrase components have larger values. In order to solve this issue, we modified the method of tone component prediction as shown in the next section.

3.3 Improved method

The improved method for tone component generation differs from the original version as follows in its way of representing tone components of tone nuclei:

1. For T1 and T3, the tone component is still represented as a straight line, but its slope coefficient has a value with the same absolute value and opposite sign with that of the slope of the linear regression line of the phrase component of the tone nucleus. This process is conducted so that the resulting F_0 contour of tone nucleus comes close to a level line.
2. For T2 and T4, 11 rising templates and 11 falling templates are prepared manually by observing tone components. These 22 templates are used for both T2 and T4.

Although parameters for tone component are predicted in parallel using the same input parameters in the original method, in the improved method, prediction is done sequentially in the order of onset time, offset time, and the rest parameters (average F_0 value, absolute pitch range and template identity, for T2 and T4). Information predicted in the former process is added to the succeeding prediction process: onset time is added to input parameters for offset time prediction, for instance. After predicting the onset and offset times for each tone nucleus, corresponding phrase component is identified and its slope coefficient and average F_0 value are calculated. These values are further added to the prediction of the average F_0 value (of the tone component) for all the tone types, and the absolute pitch range and template identity for T0, T2 and T4. The slope coefficients for T1 and T3 are not included in the prediction process: They are calculated from the slope coefficients of phrase components after the prediction of onset and offset times of tone nucleus.

4. Experiments on F_0 contour generation

In order to investigate the advantage of the improved method over the original one, a listening experiment was conducted for synthetic speech with F_0 contours generated by the two methods. The same 100 news utterances explained in section 2 were used to train both of the methods. Each utterance consists of about 50 syllables. Totally, the 100 utterances include 4839 syllables. First, all the F_0 contours were manually decomposed into tone and phrase components. Also, tone nucleus was searched for each syllable. For T2 and T4, the tone nucleus can be detected rather easily by searching peaks and valleys in F_0 contours. On the other hand, it is rather difficult to automatically find the flat F_0 portion for T1 and T3. Therefore, their tone nuclei were manually extracted. These syllables were used to train binary decision trees for predicting tone component parameters.

Further 30 sentences were selected from the speech corpus of the same speaker, and their F_0 contours were generated by the original and improved methods. Figure 4 shows the observed F_0 contour and the F_0 contours generated by the two methods for "ta1 yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1" (He joined the Chinese Workers' and Peasants' Red Army in April 1932). In most cases, the generated

F_0 contours for both methods are quite similar to the original F_0 contours.

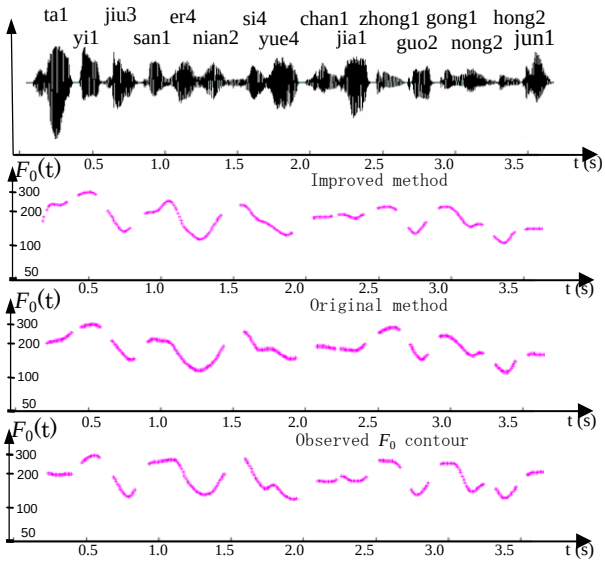

Figure 4: The samples of result. From top to bottom panels: original wave form, F_0 contour generated by the improved method, one by original method, one extracted from original speech.

A listening test was further conducted after synthesizing speech by substituting the original F_0 contours to the generated F_0 contours by TD-PSOLA. Also, speech synthesis was conducted using F_0 contours generated by the HMM-based speech synthesis. In order to combine 24-order mel-cepstrum with F_0 (and their delta and delta-delta values), MSD (Multi-Space probability Distribution) HMM was used [4]. Phone boundaries were fixed to those of the original utterance during HMM training and synthesis processes. The speech synthesis was done again by TD-PSOLA (by substituting the generated F_0 contours to the original contours).

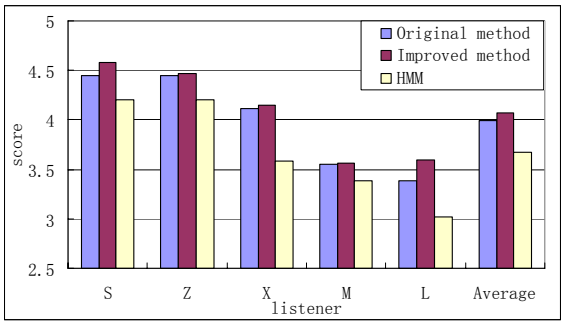

Figure 5: Result of the listening test.

Five native speakers of Mandarin were asked to evaluate the synthetic speech with a focus on prosody, using a five-point scoring: 5 (excellent), 4 (good), 3 (marginal), 2 (poor), and 1 (very poor). Totally, 90 utterances were synthesized and presented randomly. The result is shown for each listener as an average in Fig. 5.

Although the scoring fluctuates among listeners, the scores are the best for the improved method for all the listeners, though the difference between the original and improved methods is minor. It should be noted that both the original and improved methods provide clearly better results than the HMM. One possible reason for low scores for HMM is the amount of training data being not sufficient; our method works even when a small sized speech corpus is obtainable. In addition, as already pointed out in section 1, one of the major merits for our methods is their ability for "flexible" control of prosody [10].

5. Conclusion

An improvement was made for tone component generation in our method of F_0 contour synthesis for Mandarin speech. Validity of improvement was shown through comparative experiments. Listening test for synthetic speech showed the advantage of our method over an HMM-based method. Future research includes realization of various styles in synthetic speech.

The authors' sincere thanks are due to Prof. Renhua Wang, the University of Science and Technology of China for his providing us the Mandarin speech corpus.

6. References

[1] S. Chen, S. Hwang, and Y. Wang, "An RNN-base prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. on Speech and Audio Processing*, Vol.6, No.3, pp.226-239, 1998.

[2] J. Tao, and L. Cai, "Clustering and feature learning based F_0 prediction for Chinese speech synthesis," *Proc. ICSLP*, pp.2097-200, 2002.

[3] J. Ni, and K. Hirose, "Synthesis of fundamental frequency contours of standard Chinese sentences from tone sandhi and focus conditions," *Proc. ICSLP*, pp.195-198, 2000.

[4] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling". *Proc. IEEE ICASSP*, pp.229-232 1997.

[5] H. Fujiaski, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol. 5, No. 4, pp. 233-242, 1984.

[6] K. Hirose, K. Sato, Y. Asano and N. Minematsu, "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, No.3-4, pp.385-404, 2005.

[7] Q. Sun, K. Hirose, W. Gu, and N. Minematsu, "Generation of fundamental frequency contours for Mandarin speech synthesis based on tone nucleus model," *Proc. Interspeech-Eurospeech-*, pp.3265-3268, 2005.

[8] Q. Sun, K. Hirose, W. Gu, and N. Minematsu, "Rule-based generation of phrase components in two-step synthesis of fundamental frequency contours of Mandarin," *Proc. Speech Prosody*, pp.561-564, 2006.

[9] J. Zhang, and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition," *Speech Communication*, Vol. 42, Nos. 3-4, pp.447-466, 2004.

[10] Q. Sun, K. Hirose, and N. Minematsu, "Two-step generation of Mandarin F_0 contours based on tone nucleus and superpositional models", *Proc. ISCA Workshop on Speech Synthesis (SSW-6)*, pp.154-159, 2007.