Generation of Fundamental Frequency Contours for Mandarin Speech Synthesis based on Tone Nucleus Model

Qinghua Sun, Keikichi Hirose**, Wentao Gu**, 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, wtgu, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A new method for generating sentence F_0 contours of Mandarin speech is proposed. The method assumes the F_0 contour generation process model, but generates the tone and phrase components in different ways and sums them to produce a sentence F_0 contour. The tone component is generated concatenating F_0 patterns of tone nuclei, which are predicted by a corpus-based scheme (binary decision trees). Experiments of F_0 contour generation were conducted by using 100 news utterances by a female speaker. The results showed that the method could generate F0 contours close to those of target speech. A perceptual evaluation was also conducted on the synthetic speech using the F0 contours generated by the method. An average score of 4.5 in a 5-point scale indicates the high naturalness, verifying the validity of the method.

1. Introduction

Recently, quality of synthetic speech has been largely improved by introducing selection-based waveform concatenation schemes. However, there still remain problems if we view from the prosodic features. Or we should say that the high quality in each sound emerges problems in 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 is well known, Mandarin is a typical tone language, in which each syllable with the same phoneme constitution can have about four tones indicating different meanings. Fundamental frequency (henceforth, F_0) contours of utterances should include these local tonal features in addition to the sentential intonation corresponding to higher-level structures. This situation makes F_0 movements of Mandarin sentences more complicated than non-tone 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.

The benefit of corpus-based methods over rule-based methods increases when handling complicated features. Naturally, most of the F_0 control schemes adopted in recent Mandarin speech synthesizer are corpus-based using decision trees, neural networks, hidden Markov models, and linear regression analysis. However, most of them are predicting syllable F_0 contours without explicit considerations on the F_0 movement in longer units such as word, phrase and so on [1-3].

The F_0 contour generation process model (henceforth F_0 model) originally developed for Japanese [4] has been successfully extended to Mandarin by introducing negative commands [5]. The Mandarin version assumes tone commands instead of accent commands, and represents a logarithmic F_0 contour as the sum of phrase and tone components. A close approximation to an observed F_0 contour has already been shown to be possible [6], and therefore a better control of prosodic features in synthetic speech might be possible by using the model. Corpus-based generation of F_0 contours in the framework of F_0 model is feasible when we have enough training data with tone and phrase command information. Unfortunately, this is currently not the case. Through the Analysis-by-Synthesis process with manually (and carefully) assigned initial parameters, the tone and phrase commands can be extracted from observed F_0 contours in high accuracy. Although we have developed a scheme to automate the command extraction process, the result is still not satisfactory [7]. This situation makes the construction of training corpus with F_0 model command information difficult.

These considerations led us to a new method of F_0 contour generation for Mandarin speech synthesis, where the tone components were generated by concatenating F_0 patterns of tone nuclei, predicted by a corpus-based method, and then were superposed to the phrase components. Here, "tone nucleus" is defined as a portion of syllable, which shows a stable F_0 pattern regardless of the context [8]. By first generating F_0 patterns only for tone nuclei of constituting syllables and then concatenating them, a smooth sentence F_0 contour can be generated.

Although the current experiment is limited to the generation of tone components, we are planning to generate phrase components by a rule-based scheme on the basis of the F_0 model. The F_0 contours are considered to consist of both language specific and universal characteristics. Features for tone components may be mostly language specific, while those for phrase components may be mostly language universal, because they are tightly related to higher-level linguistic information, such as syntactic structure, discourse structure, and so on. We have already realized a rule-based control of phrase components for Japanese speech synthesis and revealed that a good quality is possible even with simple rules [9]. Similar rules are applicable to control the phrase components in Mandarin speech synthesis.

The rest of the paper is organized as follows. Section 2 describes the tone nucleus model. The detail of the proposed method is shown in Section 3. The results on F_0 contour generation and speech synthesis are given in Section 4. Section 5 concludes the paper.

2. Tone nucleus model

We divide each syllable of Mandarin into two parts: initial consonant and a final vocalic part. The initial consonant can be voiced or voiceless, and the final vocalic part consists of vowel(s) and an optional nasal coda.

In Mandarin, there are four lexical tones attached to each syllable. They are referred to as T1, T2, T3 and T4, which are characterized by high-level, high-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 tone. The neutral tone occurs not only on certain particles; any lexical tones can be neutralized in an unstressed syllable, for example, in the second syllable of some bi-syllabic words.

For a syllable F_0 contour, only its later portion, approximately corresponding to the final vocalic part, is regarded to bear tonal information, whereas the early portion is regarded as physiological transition period from the previous tone. It was also found that there are often cases where voicing period in the ending portion of a syllable also forms a transition period of vocal vibration and contributes nothing to the tonality. From this consideration, 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 [8]. The three segments are called onset course, tone nucleus, and offset course, respectively, which are defined as follows:

1. Onset course is an F_0 transition from the preceding syllable to the onset target of the tone nucleus. This segment covers the initial consonant and the transition period of the final vocalic part.
2. Tone nucleus is a portion where F_0 contour keeps the basic pattern of the tone unless it is affected by high-level prosodic factors such as neutralization, contextual effect, focus, phrasing, and etc. This segment covers the nucleus of the final vocalic part.
3. Offset course is an F_0 transition from the offset target of the tone nucleus to the following syllable. This segment holds the ending course of the final vocalic part.

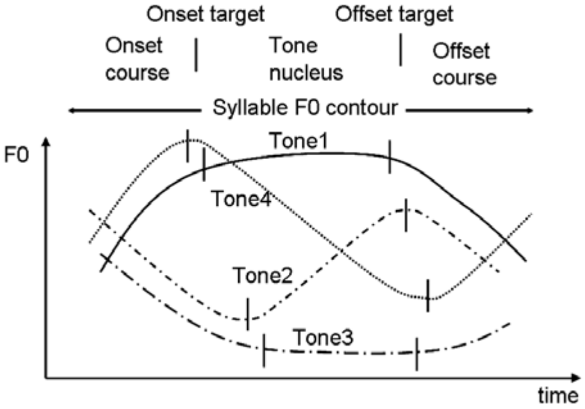

Figure 1: Tone nuclei for the four lexical tones.

Figure 1 illustrates syllable F_0 contours with possible articulatory transitions for the four lexical tones. 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. These observations led us to an idea of generating F_0 pattern only for tone nuclei, and to concatenate them to produce a sentence F_0 contour. (Note: In the original tone nucleus model, tone nucleus for T3 was not defined.)

From the speech data shown in section 4, 1094 T2 samples for training were selected and clustered into five groups after normalizing in time and frequency. For each group, an F_0 template is calculated as the average of the samples in the group. Figures 2 and 3 respectively show the templates for entire syllable and tone nucleus. It is clear that five F_0 templates are quite different when they are viewed for the entire syllable, but are similar when viewed for the tone nucleus only.

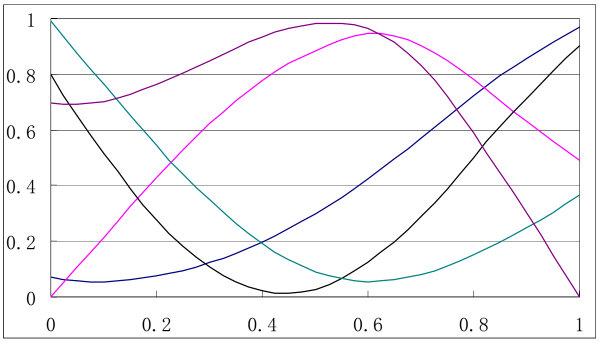

Figure 2: Syllable F_0 templates for T2. The vertical and horizontal axes are respectively normalized frequency and time.

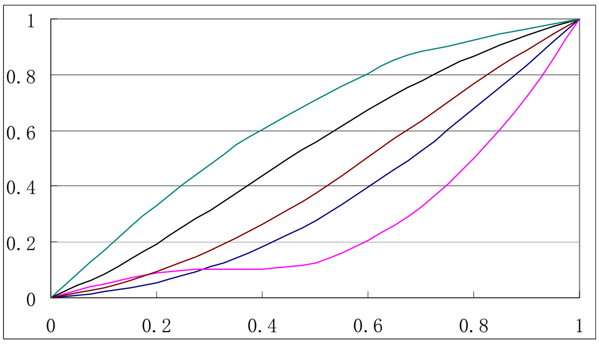

Figure 3: Tone nucleus F_0 templates for T2, obtained. The vertical and horizontal axes are respectively normalized frequency and time.

3. Method of F_0 contour generation

As already mentioned in Section 1, the proposed method for F_0 contour generation is to generate tone components by a corpus-based scheme and to generate phrase components by a rule-based scheme on the basis of the F_0 model. The current paper is focused on the generation of tone components. The tone components are generated through the following process:

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 shape of 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 3rd order polynomials, they are concatenated using straight lines, because, in preliminary listening test, no clear difference is perceived on the quality of synthetic speech using different concatenation methods.

In the first and second steps above, the parameters are predicted using binary decision trees trained separately for each parameter. Although it is possible to differently select the inputs for each tree, they are commonly selected in the current experiment as shown in Table 1. Taking the limited size of available training data into account, initial consonants are grouped into 5 categories: (1) /b/, /d/, /g/, /p/, /t/, /k/; (2) /z/, /zh/, /j/, /c/, /ch/, /q/; (3) /f/, /s/, /x/, /h/, /sh/; (4) /r/, /l/, /m/, /n/; and (5) null. The final vocalic part has two categories: with and without nasal coda. The boundary depth, from shallow to deep, gives 6 categories: intra-word syllable boundary, word foot boundary, prosodic word boundary, prosodic phrase boundary, punctuated break boundary, and sentence boundary.

Table 1: Inputs for the predictor.

Input	Category
Initial consonant of current syllable	5
Final vocalic part of current syllable	2
Final vocalic part of preceding syllable	2
Initial consonant of following syllable	5
Tone of current syllable	5
Tone of preceding syllable	5
Tone of following syllable	5
Duration of initial consonant	Continuous
Duration of final vocalic part	Continuous
Duration of voiced part	Continuous
Boundary depth between preceding and current syllables	6
Boundary depth between current and following syllables	6
Position of syllable in current breath group	Natural num.
Number of syllables in current word	Natural num.
Position of current word in sentence	Natural num.
Duration of short pause preceding to current syllable	Continuous or 0
Duration of short pause following to current syllable	Continuous or 0

Based on the discussion in Section 2 and careful observation of F_0 contours of Mandarin speech, the tone nuclei are defined for each tone type as shown in Table 2. Since T0 shows no inherent F_0 contour, a stable definition of tone nucleus is difficult, and hence we assume the entire voiced segment of the syllable as tone nucleus for T0. The parameters for representing the tone components of nuclei are as follows:

1. For T1 and T3, tone nuclei are defined as the flat portion, which is represented by a single parameter, *i.e.*, average F_0 value.
2. For T0, T2 and T4, tone components for nuclei are normalized both in time and pitch range, and the normalized contours are then clustered into several groups. The average contour for each group serves as a template to represent the tone component of nucleus. The parameters include the absolute pitch range, average F_0 value, and template identity.

When generating a contour for tone nucleus, for T1 and T3, it is approximated as a level line at the predicted F_0 level. For T0, T2 and T4, the tone nucleus contour is generated as follows:

1. Select one of the templates.
2. Adjust (expand or shrink linearly) the selected template to fit the time span and pitch range of the tone nucleus.
3. Place the adjusted template at the frequency level indicated by the predicted average F_0 value.

Table 2 summarizes the parameters need to be predicted.

Table 2: Definition of tone nucleus and output parameters of the predictor for each tone type.

Tone type	F_0 contour feature	Parameters *
T1	Flat F_0 with high level	Average F_0
T2	Rising F_0	Average F_0, Pitch range
T3	Flat F_0 with low level	Average F_0
T4	Falling F_0	Average F_0, Pitch range
T0	No specific feature (Entire voiced segment)	Average F_0, Pitch range

*Also onset/offset times and template identity for Tones 2, 4, 0.

4. Experiments

Experiments were conducted on the generation of F_0 contours. The speech data are 100 news utterances by a native female speaker of Mandarin. Each utterance consists of about 50 syllables. A smoothing process based on the piecewise 3rd order polynomials [10] was applied to F_0 contours of these utterances. The resulting smoothed F_0 contours were then used for the experiments.

First, all the F_0 contours were manually decomposed into tone and phrase components. In the current experiments, only the tone components were generated by the proposed method; the phrase components were kept as they were. Then, tone nucleus was searched for each syllable. For T2 and T4, a nucleus can be detected rather easily by searching for peaks and valleys of 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. Among the tone nucleus samples thus obtained, for each tone type, the first 50 samples were selected as testing data, while the remainders were used for training as shown in Table 3.

Table 3: Number of tone nucleus samples used in the experiments.

Tone type	T1	T2	T3	T4	T0
Training	992	1094	685	1520	298
Testing	50	50	50	50	50

For each of T0, T2 and T4, the training samples were clustered into 11 groups to generate 11 templates. The number of templates was decided from the observation of F_0 contours of training samples.

As mentioned already in Section 3, when generating tone components, the onset and offset of tone nucleus were positioned in the syllable first. Then, for T0, T2 and T4, the template identity, pitch range, and average F_0 value are predicted, while for T1 and T3, only the average F_0 value is predicted. A binary decision tree was constructed for each of above parameters of each tone type using the training data shown in Table 3.

Table 4 shows the root mean square (RMS) errors in the predicted onsets and offsets of tone nuclei for the testing data. According to different choices of time reference (voiced segment or entire syllable) and normalization, there are four approaches in representing the onset/offset times. Since no systematic and clear difference was found between the four choices, we just selected "normalize with the whole syllable" as the version for the speech synthesis experiment below, for its accuracy is better and its binary decision trees is easy to be understood by humans. The prediction errors are mostly less than 10% of the syllable length.

Table 4: RMS errors in predicting onsets and offsets of tone nuclei (in seconds). V and S denote voiced segment and entire syllable, respectively.

		Without norm.		With norm.	
Tone	Ref.	Onset	Offset	Onset	Offset
T1	V	0.0070	0.0156	0.0207	0.0156
	S	0.0070	0.0100	0.0069	0.0127
T2	V	0.0242	0.0134	0.0292	0.0094
	S	0.0220	0.0140	0.0170	0.0098
T3	V	0.0200	0.0180	0.0196	0.0191
	S	0.0260	0.0270	0.0261	0.0204
T4	V	0.0174	0.0116	0.0267	0.0083
	S	0.0170	0.0080	0.0181	0.0103

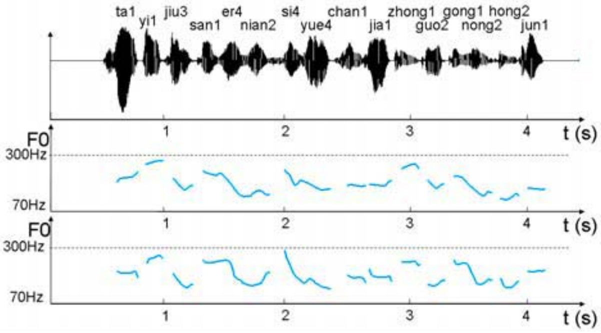

Figure 4: From top to bottom: waveform of synthesized speech, F_0 contour generated by the proposed method, and observed F0 contour of target speech. The utterance is "ta1 yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1" (He joined the Chinese Workers' and Peasants' Red Army in April 1932).

We selected 9 utterances, which consist of syllables for testing only. Their F_0 contours were generated by superposing the tone components produced by the proposed method on the original phrase components of the utterances. Then the speech synthesis (TD-PSOLA) was conducted by substituting the original F_0 contours to the generated contours. As clearly shown in Fig. 4, the generated F0 contour is quite close to the observed F0 contour. In Fig.4, the RMS error of F0 in log domain between generated speech and original (natural) speech is about 0.105. The quality of synthetic speech was evaluated with a focus on prosody, using a five-point score: 5 (excellent), 4 (good), 3 (acceptable), 2 (poor), and 1 (very poor). We used 18 utterances including 9 of synthetic speech and 9 of original speech for listening test. These utterances were presented in a random order to three native speakers. The average score was 4.5, indicating high naturalness of the synthetic speech.

5. Conclusion

A new method of generating F_0 contours for Mandarin speech synthesis was proposed. The method generates tone and phrase components of F_0 contours differently: corpus-based method for tone components and rule-based method for phrase components. The tone components were generated by concatenating F_0 patterns predicted for tone nuclei. We are now developing a full system for F_0 contour generation from text.

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

6. References

[1] Chen, S., Hwang, S., and Wang, Y., "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] Tao, J. and Cai L., "Clustering and feature learning based F_0 prediction for Chinese speech synthesis," *Proc. ICSLP*, Denver, pp.2097-200, 2002.

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

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

[5] Fujiaski, H., Hirose, K., Halle, P., and Lei, H., "Analysis and modeling of tonal features in polysyllabic words and sentences of the standard Chinese," *Proc. ICSLP*, Kobe, pp.841-844, 1990-10.

[6] Wang, C., Fujisaki, H., Tomana, R., and Ohno, S., "Analysis of fundamental frequency contours of Standard Chinese in terms of a command-response model and its application to synthesis by rule of intonation," *Proc. ICSLP*, Beijing, pp.326-329, 2000.

[7] Gu, W., Hirose, K., and Fujisaki, H., "Automatic extraction of tone command parameters for the model of F_0 contour generation for Standard Chinese," *IEICE Trans. Information & Systems*, Vol. E87-D, No. 5, pp. 1079-1085, 2004.

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

[9] Hirose, K. and Fujisaki, H., "A system for the synthesis of high-quality speech from texts on general weather conditions," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, Vol.E76-A, No.11, pp.1971-1980, 1993.

[10] Narusawa, S., Minematsu, N., Hirose, K. and Fujiaski, H., "A method for automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. IEEE ICASSP*, Orlando, pp.509-512, 2002.