Rule-based Generation of Phrase Components in Two-step Synthesis of Fundamental Frequency Contours of Mandarin

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

In this paper, a rule-based method was developed for realizing phrase components in our two-step generation of fundamental frequency (F0) contours of Mandarin. The scheme assumes (logarithmic) F0 contours as superposition of tone components on phrase components, which are further assumed to be responses of phrase commands. In general, possibility of a new phrase command comes higher at deeper syntactic boundaries, but is also affected by the distance from the preceding phrase command. A long interval from preceding phrase command causes a flat F0 contour close to baseline, which is not the case in human speech. In the case of tonal languages such as Mandarin, tone components can be negative. Hence, to give a margin for downward F0 movement, phrase components need to be kept above a certain level, which requires more frequent phrase commands as compared to non-tonal languages. Based on these facts, simple rules were constructed for phrase component generation. Speech synthesis was conducted using F0 contours generated by the method. The result of listening test showed a good control of F0 contours being realized by the method.

1. Introduction

Due to recent ability of computers for processing a large amount of speech data, selection-based waveform-concatenation schemes are adopted in most commercial speech synthesis systems realizing a rather high quality of speech. However, improvement of speech quality reveals problems in intonation, rhythm, and etc. To solve these problems, control of prosodic features becomes an important issue in speech synthesis.

Utterances of tone languages show more complicated movements in their fundamental frequency (F_0) as compared to non-tonal languages. For instance, in Mandarin, each syllable can have up to four tonal types, each of which shows a distinctive F_0 pattern. In addition to these local tonal features, F_0 shows movements in longer spans, corresponding to syntactic/utterance structures. This situation makes F_0 contours of Mandarin utterances more complicated than non-tonal languages like English, Japanese and so on. Therefore control of F_0 contours becomes a crucial issue in Mandarin speech synthesis.

Most F_0 controls adopted in Mandarin speech synthesis are corpus-based using decision trees, neural networks, hidden Markov models, and linear regression analysis [1-3]. However, most of them predict syllable F_0 contours without explicit consideration on the F_0 movement in longer units such as words, phrases, and so on. This situation causes certain unnaturalness in synthesized speech. A better control of F_0 movements in longer units in synthetic speech is possible using the F_0 contour generation process model (henceforth, F_0 model), which represents a logarithmic F_0 contour as the sum of phrase and tone components [4]. Phrase components are assumed to be responses to impulse-like phrase commands, while tone components are assumed to be responses to step-wise tone commands. Tone components for Tones 1 and 3 are respectively generated from positive and negative commands, while those for Tones 2 and 4 are respectively generated from pairs of positive-negative commands and negative-positive commands [5]. This model was originally developed for Japanese, where accent components were assumed instead of tone components, and was successfully used in our corpus-based method for F_0 contour generation [6]. In the case of Mandarin, however, it is rather difficult to develop a fully corpus-based method, because of complicated F_0 features; arrangement of enough training corpora with reliable information on F_0 model command is not an easy task.

Fundamental frequency contours are considered to include both language specific and universal features. Features for tone components may be mostly language specific, while those for phrase components may be mostly language universal. This is because phrase components are tightly related to higher-level linguistic information, such as syntactic structure, discourse structure, and so on. Through analysis on Japanese F_0 contours, we already have an ample knowledge on how phrase components are related to linguistic information, and may be able to construct rules for generating phrase components for Chinese speech synthesis through analogy of Japanese. Moreover, phrase commands extracted from observed F_0 contours of Mandarin may include lager errors than tone components do; corpus-based method generation is not appropriate for phrase components.

These considerations led us to a new method of F_0 contour generation for Mandarin speech synthesis, where the tone components were generated by corpus-based method and superposed onto the phrase components, which are generated by a rule-based method. In general, F_0 pattern of a syllable in a Mandarin utterance shows a relatively stable shape at its center regardless of the context, called "tone nucleus." To make the most of this stability, fractions of tone components are first predicted only for tone nuclei, and then are concatenated to produce entire tone components [7]. Since, phrase and tone components are tightly related, there are occasionally "strange" F_0 movements if both components are generated independently. To cope with this situation, we developed a two-step scheme, where information on generated phrase components is included in the inputs for prediction of tone components.

Our preliminary experiment showed that a good quality of speech is possible even with phrase components generated by a simple set of rules. However, close observation of Mandarin F_0 contours reveals that phrase components have rather different distribution from those in Japanese. In the case of

Mandarin, phrase commands are frequently positioned, so that phrase components can usually be above a certain level to give a margin for negative tone components.

Based on these observations, we have newly developed rules to generate phrase components from text input. After generating phrase components by the rules, sentence F_0 contours are generated by predicting tone components through the two-step scheme.

The rest of the paper is organized as follows. Section 2 describes the characteristics of phrase components of F_0 contour in Mandarin as compared to those in Japanese. Then the rules for generating phrase components are presented in Section 3. After a brief explanation on the two-step scheme in Section 4, experimental results on F_0 contour generation are shown in Section 5. Section 6 concludes the paper.

2. Phrase components in Mandarin and Japanese

It is observed that phrase components are related to syntactic structures; phrase commands tend to occur at deeper syntactic boundaries. However, they are also affected by the human habits of utterance; there is a certain limit in the distance between two succeeding phrase commands. We showed that a good control of phrase components was possible for Japanese by a set of simple rules, which are basically placing larger phrase commands at deeper syntactic boundaries and adding phrase commands at shallower syntactic boundaries to keep the distance between two succeeding phrase commands blow a threshold [8]. These rules, however, cannot be applied to Mandarin as they are. Figures 1 and 2 show the F_0 contour of example utterances for Japanese and Mandarin, respectively. It is clear that phrase commands occur more frequently in Mandarin than in Japanese: in normal speech rate, the distance between two succeeding phrase components could be around 15 morae (2.1 sec) for Japanese, while it is mostly less than 7 syllables (1.4 sec) for Mandarin. Frequent phrase commands in Mandarin may be due to the fact that tone components can have negative values causing sharp declination in F_0 contours below phrase components. Phase component should always be above a certain level so that, in principle, F_0 does not go below the baseline F_b even with negative tone components. This may be a possible reason for shorter distances between phrase commands in Mandarin. Therefore, we developed rules for phrase command generation of Mandarin by placing priority in the control of phrase component values at phrase boundaries as explained in the next section.

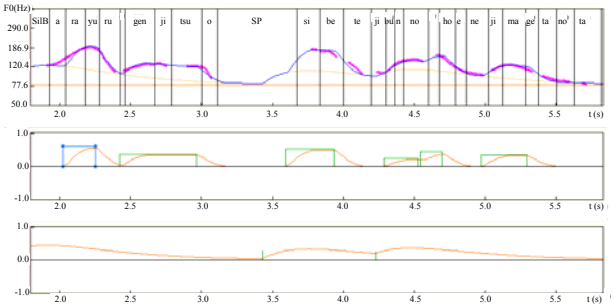

Figure 1: An example of F_0 contour of Japanese utterance "arayuru geNjitsuo subete jibuNnohooe nejimagetanoda ((He) twisted all the reality to his side.)." From top to bottom: observed F_0 contour with its F_0 model approximation, accent components/commands, phrase components/commands.

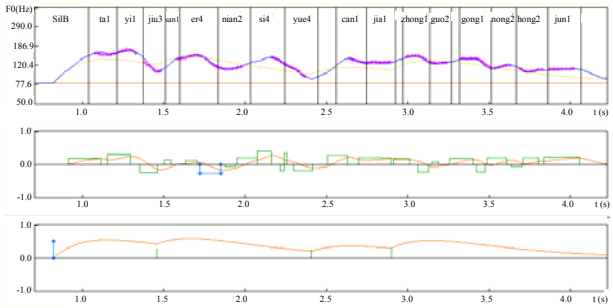

Figure 2: An example of F_0 contour of Chinese utterance "ta1 yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1 (He joined the Chinese Workers' and Peasants' Red Army in April 1932.)." From top to bottom: observed F_0 contour with its F_0 model approximation, tone components/commands, phrase components/commands.

3. Generation of phrase components

3.1. Observation of speech samples

Speech corpus used for the experiments consists of 100 news utterances by a native female speaker of Mandarin. Each utterance includes around 50 syllables on average. From F_0 contours of these utterances, 1264 phrase commands were extracted manually. Out of these commands, 888 commands located neither at utterance-initial (henceforth, SilB) nor immediately after a short pause (longer than 200 ms, abbreviated as SP) are selected, and their distribution depending on the F_0's at their position (therefore, at the initial points of corresponding phrase components) is shown in Figure 3. The F_0's are grouped in 10 Hz ranges as shown in the horizontal axis of the figure. The baseline F_b is around 120 Hz for all the utterances. It is clear from the figure that a majority (513) of the phrase components start from the range of 150 Hz - 190 Hz. The cases starting from values below 140 Hz are rather limited, and no case below 130 Hz.

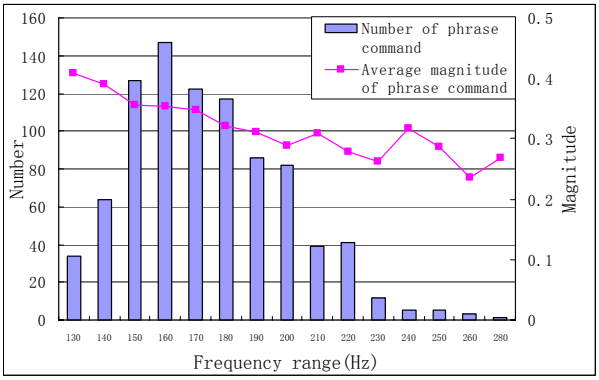

Figure 3: Distribution of phrase commands sorted for the F_0 values at the command positions in 10 Hz ranges. "130" in the abscissa means the range from 130 Hz to 140 Hz, and so on. Average phrase command magnitude for each frequency range is also shown.

An extended analysis of the distribution is possible by introducing the notion of "prosodic word", which we will define as a chunk of syllables usually uttered in a tight connection. So, a prosodic word can be a word, a compound word, or a word chunk uttered together frequently. For

example, the sentence shown in Figure 2 can be segmented as follows:

(ta1) | (yi1 jiu3) | (san1 er4 nian2) | (si4 yue4) | (can1 jia1) | (zhong1 guo2) | (gong1 nong2) (hong2 jun1).

Here, a pair of parentheses embraces an element (syntactic) word, while "|" indicates prosodic word boundary. Phrase components can be grouped into the following cases. Henceforth, for simplicity of explanation "phrasal F_0" is used to indicate the F_0 value contributed by phrase components.

Case 1: A phrase command is always positioned at SilB or immediately after SP. The average magnitude is 0.61 for those at SilB, 0.59 immediately after SP longer than 300 ms, and 0.45 immediately after SP between 200ms and 300 ms.

Case 2: When phrasal F_0 falls into the range between 150 Hz and 190 Hz at a prosodic word boundary, a phrase command is positioned there. It takes different magnitudes according to the number of preceding phrase commands after SilB or SP till the current position as shown in Table 1. In Tables 1-5, "number of phrase commands" means "number of phrase commands between preceding SilB/SP and current phrase command (including current one)."

Case 3: A phrase command is observed with phrasal F_0 higher than 190 Hz, when, if the command being deleted, phrasal F_0 falls below 150 Hz at the next prosodic word boundary. It takes different magnitudes according to the number of preceding phrase commands after SilB or SP till the current position as shown in Table 2 (when phrasal F_0 falls into the range between 190 Hz and 230 Hz) or Tables 3 (when phrasal F_0 falls into the range between 230 Hz and 290 Hz).

In Tables 1, 2 and 3, considering the statistical credibility, the results calculated from little number of samples are abandoned, and are not used for the following rules of phrase component generation.

Table 1: Relationship between magnitude and position of phrase commands when phrasal F_0 falls into 150Hz~190Hz range.

Number of phrase commands	Number of samples	Average magnitude
1	27	0.462
2	171	0.360
3	154	0.353
4	108	0.351
5	58	0.293
6	20	0.286
7	1	0.280
8	1	0.237

Table 2: Relationship between magnitude and position of phrase commands when phrasal F_0 falls into 190Hz~230Hz range.

Number of phrase commands	Number of samples	Average magnitude
1	1	0.423
2	118	0.317
3	77	0.284
4	33	0.279
5	14	0.260
6	2	0.299
7	4	0.307

Table 3: Relationship between magnitude and position of phrase commands when phrasal F_0 falls into 230Hz~290Hz range.

Number of phrase commands	Number of samples	Average magnitude
2	24	0.267
3	3	0.269
4	1	0.330

3.2. Rules for phrase component generation

Based on the observations on phrase components in section 3.1, the following set of rules is developed for phrase command assignment:

Rule 1: Place a phrase command with magnitude 0.6 at SilB or after an SP longer than 300 ms. Place a phrase command with magnitude 0.47 after a SP shorter than 300 ms but larger than 200 ms.

Rule 2: Check all the prosodic word boundaries without an SP in a left-to-right manner from the utterance initial. If phrasal F_0 at the current boundary falls into a lower range (set at 150Hz ~ 190Hz according to Figure 3), place a phrase command with magnitude as shown in Table 4, depending on the number of preceding phrase commands. Table 4 is decided by referring to Table 1.

Rule 3: During the process of rule 2, when phrasal F_0 at the current prosodic word boundary falls below the lower range, go back to the preceding boundary and place a phrase command there with magnitude shown in Table 5 depending on the feature of preceding phrase commands. Table 5 is decided by referring to Tables 2 and 3. If a phrase command has already been placed at the preceding boundary, or if "number of phrase commands" or "phrasal F_0" is out of the range of Table 5, skip to rule 4.

Rule 4: Split the prosodic word before the current word boundary into two smaller prosodic words. Then go back to apply rules 2 and 3 on the newly inserted prosodic word boundary.

An additional rule is applied to the timing of phrase commands. The distance of the phrase command ahead of the corresponding prosodic boundary is set as follows: 150 ms for the phrase commands larger than 0.5, 50 ms for the commands smaller than 0.3, and 80 ms for those in between.

Table 4: Magnitude of phrase command placed at the current prosodic word boundary when phrasal F_0 falls into the lower range.

Number of phrase commands	2	3	4	5	≥6
Magnitude of phrase command	0.36	0.35	0.35	0.29	0.29

Table 5: Magnitude of phrase command placed at the preceding prosodic word boundary when phrasal F_0 falls below the lower range at the current prosodic word boundary.

F_0 at immediately preceding prosodic word boundary	190Hz~230Hz				230Hz~280Hz
Number of phrase commands	2	3	4	5	2
Magnitude of phrase command	0.32	0.28	0.28	0.26	0.29

4. Generation of F_0 contours

Tone components are then generated by the corpus-based method and superposed onto the phrase components to produce sentence F_0 contours [7]. Taking account that the tone components are tightly related to the phrase components, the two-step scheme is adopted where generated phrase command information is added to the inputs for the tone command prediction [9]. It was shown in our previous experiment that independent generation of phrase and tone components (one-step scheme) causes a serious degradation in the final F_0 contour.

5. Experiment on the F_0 contour generation

Nine utterances were selected from 100 utterances of the corpus mentioned in section 3.1, and their F_0 contours were generated by the developed method. Tone components are generated by the binary decision trees, which were trained using the same speech corpus [7, 9]. Figure 4 shows the generated F_0 contour together with its phrase components for the sentence appeared in Figure 2.

Although the F_0 contours generated by our method are rather close to the observed ones, some differences are observed between their phrase components, as shown in Figures 2 and 4. Since a rather wide variety is allowed in the F_0 contour of a sentence, these differences may not directly means the degradation; a listening evaluation of synthetic speech is required.

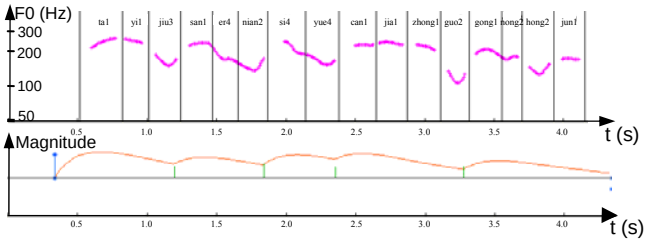

Figure 4: F_0 contour generated by the two-step scheme (top), and its phrase components/commands (bottom).

Speech synthesis (TD-PSOLA) was conducted by substituting the original F_0 contours to the generated F_0 contours. 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). The F_0 contours were generated by three different approaches: *o-gt*, *r-go*, and *r-gt*, where *o* and *r* indicate the original and the generated phrase components respectively, while *gt* and *go* indicate the tone components generated by the two-step scheme and by the one-step scheme, respectively. The synthetic speech samples were presented in a random order to five native speakers of Chinese. The average score for each speaker is shown in Figure 5. Scores above 4 are obtained by the method of *r-gt* (generated phrase plus two-step scheme) for all the listeners except one. It should be noted that the scores are almost the same as the case *o-gt* using original phrase components but much better than in the case *r-go* using phrase component generated by one-step scheme. These results indicate that the proposed rule-based method for phrase component generation works quite well under the two-step scheme.

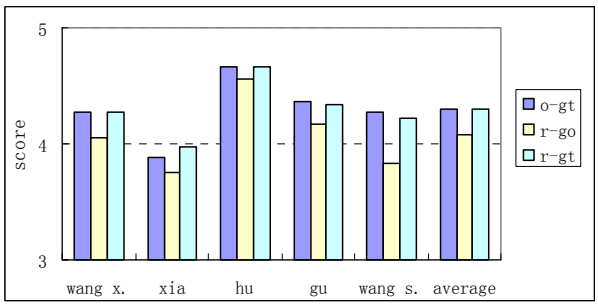

Figure 5: Results of listening test.

6. Conclusions

A rule-based method for phrase component generation was proposed in our F_0 contour synthesis method of Mandarin speech. It is clearly shown that the F_0 contour with high naturalness can be generated by the proposed method. Duration control is now in the scope of our research work, which is necessary for constructing a full system of Mandarin speech synthesis.

7. 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*, Denver, pp. 2097-2100, 2002.

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

[4] 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.

[5] K. Hirose, H. Lei, and H. Fujisaki, "Analysis and formulation of prosodic features of speech in standard Chinese based on a model of generating fundamental frequency contours," *J. Acoust. Soc. Japan*, Vol. 50, No. 3, pp. 177-187, 1994.

[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. Eurospeech*, Lisbon, pp. 3265-3268, 2005.

[8] K. Hirose, and H. Fujisaki, "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.

[9] Q. Sun, K. Hirose, W. Gu, and N. Minematsu, "Use of Phrase Information in Generation of Mandarin F_0 Contours based on the Tone Nucleus Model," *Record of Autumn Meeting, Acoustical Society of Japan*, pp. 329-320, 2005. (in Japanese)