Generation of F_0 Contours for Mandarin Speech in Combination with Rule-based and Corpus-based Methods

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

*Graduate School of Information Science and Technology, **Graduate School of Engineering, University of Tokyo, Japan

{hirose, qinghua, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A method was developed for synthesizing sentence fundamental frequency (F_0) contours of Mandarin speech. It is based on representing an F_0 contour in logarithmic frequency scale 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. To keep the time-alignment between the phrase and tone components, the phrase components are first generated and their information is added to the inputs for the prediction of tone nucleus F_0 patterns. Result of listening tests on synthetic speech with the synthesized F_0 contours verified the validity of the developed method. Furthermore, it was shown through an experiment of word emphasis that a flexible F_0 control was possible by the proposed method.

1. Introduction

Introduction of selection-based waveform concatenation in speech synthesis largely improved quality of synthetic speech. However, there still remain problems if we view from the prosodic aspect. Although the control of prosodic features is an important issue in speech synthesis for any languages, it becomes crucial for Chinese. As it is well known, Mandarin is a typical tonal language and each syllable with the same phoneme sequence 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 speech 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 issue in Mandarin speech synthesis.

Several rule-based methods were developed for controlling F_0 contours in Mandarin speech synthesis [1]. Although the rule-based methods are ideal in realizing various speech styles, it is not an easy task to extract rules from observed F_0 contours. The 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 [2,3]. Among all, the hidden Markov model (HMM) is now commonly used for synthesizing speech of many languages, including Mandarin, because it can handle segmental and prosodic features simultaneously and concatenates speech segments in a statistical basis [4]. Flexible control of speech styles is possible by adapting HMM to a new style when a small-sized speech corpus of that style is available. However, it still requires a certain size to keep the speech quality. Moreover, the method handles F_0 in a 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 for the F_0 movement in longer units in synthetic speech is possible using the generation process model of F_0 contour model (F_0 model), which represents a logarithmic F_0 contour as a superposition of tone components on phrase components placed on a baseline level [5]. This model was used successfully in the corpus-based method of generating F_0 contours of Japanese [6]. The method required speech corpus with F_0 model commands for the training process, which was arranged efficiently using the method of automatic extraction of F_0 model commands from speech waveforms. However, in the case of Mandarin speech, automatic extraction comes difficult because of its complicated F_0 movements [7]. Although several efforts are going on, corpus-based F_0 contour generation fully based on the F_0 model is less feasible in the case of Mandarin.

While a syllable F_0 contour shows a stable pattern when it is uttered in isolation, it changes a lot when uttered in a sentence. This situation requires a number of templates for syllable F_0 contours, when a sentence F_0 contour is generated as a concatenation of syllable F_0 contours. Close observation of syllable F_0 contours indicates that a syllable F_0 contour consists of beginning and ending parts, which are transients from and to adjacent syllables, and mid part, which possesses rather stable F_0 pattern regardless of the tonal context [8]. The mid part with a stable F_0 pattern is often called as "tone nucleus."

These considerations led us to propose a 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 were superposed onto the phrase components, which were generated by a rule-based scheme on the basis of F_0 model [9]. By first generating F_0 patterns for tone nuclei of constituting syllables and then concatenating them, a smooth sentence F_0 contour can be generated only from a limited speech corpus.

Although speech synthesized using generated F_0 contours sounded natural, there was occasional degradation when phrase components and tone components were handled independently. In the case of Japanese, the independent handling of components did cause no clear degradation [10]. This is because, in Japanese, inter-syllabic movements of F_0 contours, viz., relative F_0 values of syllables, are important for realization of lexical accent, which is not corrupted so much due to changes in phrase components. However, in Mandarin, intra-syllable movements represent tone types, and they are often largely affected by the phrase components; for instance, rising F_0 contour characterizing T2 may appear as a falling contour when the phrase component position shifts. To cope with mismatches between two components, we developed a two-step scheme, where the phrase components were generated first, and then the tone components were generated taking the features of generated phrase components into account (Figure 1).

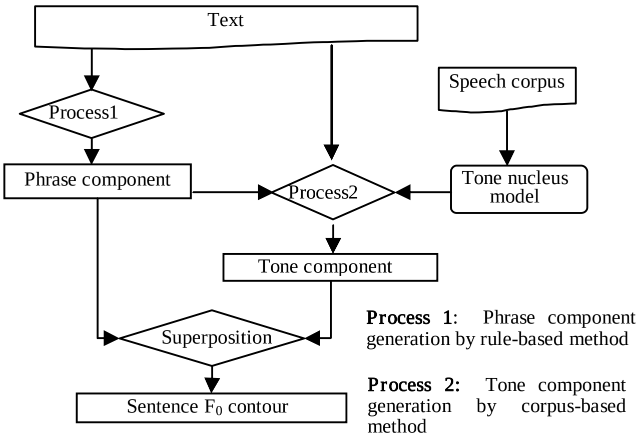

Figure 1: Two-step scheme of F_0 contour generation.

The most significant benefit of the proposed method over others without decomposition is the flexibility in F_0 contour generation: by manually controlling phrase components, we can easily generate F_0 contours with different utterance structures. In Mandarin, it is claimed that a word with emphasis is usually accompanied by a new phrase component with a large magnitude. Following to this claim, an experiment was conducted whether the control of emphasis position in a sentence is possible or not, by manually changing phrase component and generating F_0 contours using the proposed method.

The rest of the paper is organized as follows. Section 2 gives rules for phrase component generation, after showing differences found in Japanese and Chinese phrase components. In Section 3, tone nucleus is first explained and then the original and improved methods of tone component generation are given. Generated F_0 contours are evaluated through a listening test of synthetic speech in the same section. Section 4 describes the full speech synthesis system constructed using the developed methods. It also includes comparison of synthetic speech quality with that by HMM-based speech synthesizer. An experiment on word emphasis was conducted in Section 5. Section 6 concludes the paper.

2. Generation of phrase component

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. Therefore, rules developed for other languages are somewhat applicable for the control of phrase components in Mandarin. We tried to apply the rules developed for the control of phrase components of Japanese to Mandarin, and found out some differences in phrase components between two languages: in the case of Mandarin, phrase components occur more frequently than Japanese [11].

2.1. Phrase components of Mandarin speech

It is generally observed that phrase components are related to syntactic structures, and, therefore, their commands tend to occur at deeper syntactic boundaries. However, phrase components are also affected by the human habits of utterance: there is a certain limit in the distance between two succeeding commands. We showed that a proper control of phrase components was possible for Japanese by a set of simple rules, which were based on placing larger phrase commands at deeper syntactic boundaries, and adding supplementary phrase commands at shallower syntactic boundaries to keep the distance between two succeeding phrase commands blow a threshold [6]. These rules, however, cannot be applied to Mandarin speech as they are.

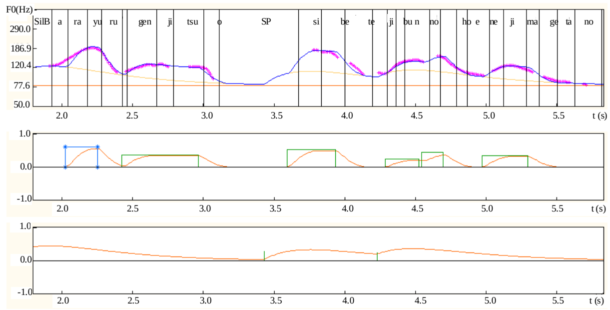

Figure 2: Example of F_0 contour of Japanese utterance "arayuru geNjitsuo subete jibuNnohooe nejimagetanoda ((He) twisted all the reality to his side.)." From top to bottom: observed F_0 contour with its F_0 model approximation, accent components/commands, and phrase components/commands.

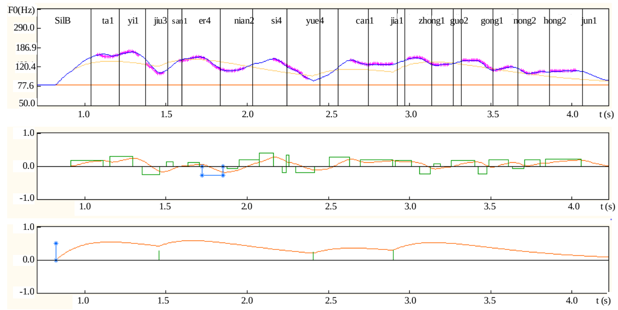

Figure 3: Example of F_0 contour of Chinese utterance "ta1 yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1 (He joined the Chinese Workers' and Peasants' Red Army in April 1932.)." From top to bottom: observed F_0 contour with its F_0 model approximation, tone components/commands, and phrase components/commands.

Figures 2 and 3 respectively show F_0 contours of Japanese and Mandarin utterances with the best approximations by the F_0 model. It is clear from the figures that phrase commands occur more frequently in Mandarin than in Japanese. It was observed that, in normal speech rate, the distance between two adjacent phrase components were 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 are considered to be due to the fact that tone components can have negative values causing sharp declination in F_0 contours below phrase components. Phrase component should always be kept above a certain level so that, in principle, F_0 does not go below the baseline even with negative tone components. Based on these observations, rules on phrase command generation for Mandarin speech synthesis are developed in the next section by placing priority in keeping certain values of phrase components at prosodic word boundaries.

2.2. Rules for phrase component generation

From a Mandarin speech corpus for speech synthesis consisting of 300 utterances by a native female speaker, arranged at University of Science and Technology of China, 100 utterances were selected for the analysis of phrase components. After extracting F_0 contours from the utterances, their phrase components were manually decomposed. Based on the statistics of 1264 samples found, the following rules were constructed, which assign phrase commands at "prosodic word" boundaries. Here, prosodic word is 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. For example, the sentence shown in Figure 4 can be segmented as follows:

(yu4ji4) | (quan2nian2) | (liang2shi6)(zong2chan3liang4) | (ke3da2) | (er4shi0dian3) | (qi1wu3yi4)(gong1jin1)

Here, a pair of parentheses embraces an element (syntactic) word, while "|" indicates prosodic word boundary. Since prosodic words are subject to change by the speaking styles, such as speech rates, it cannot be decided uniquely only from the texts. Although assignment of prosodic word boundaries is an importance issue, boundaries labeled in the corpus were used in the current paper.

Rule 1: Place a phrase command with magnitude 0.6 at the silence locating at the beginning of the sentence (SilB) or after a pause longer than 300 ms. Also, place a phrase command with magnitude 0.47 after a pause shorter than 300 ms but longer than 200 ms. (The pause lengths are predicted beforehand by a separate process. See section 4.)

Rule 2: Check all the prosodic word boundaries without pauses longer than 200 ms in a left-to-right manner from the utterance initial. If phrasal F_0 (F_0 value of phrase component plus baseline value) at the current boundary falls into a range (set to 150Hz ~ 190Hz for the speaker), place a phrase command with magnitude as shown in Table 1, depending on the number of preceding phrase commands between preceding SilB/pause and current phrase command (counting the current one). If the phrasal F_0 is larger than 190 Hz, skip to the next prosodic word boundary without placing any phrase command.

Rule 3: During the process of rule 2, when phrasal F_0 at the current prosodic word boundary falls below the range, go back to the preceding boundary and place a phrase command there with magnitude shown in Table 2 depending on the feature of preceding phrase commands. If a phrase command has already been placed at the preceding boundary, or if "number of phrase commands" or "phrasal F_0" does not fall into the cases listed in Table 2, skip to rule 4.

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

An additional rule is applied to the timings of phrase commands. The phrase command is placed ahead of the corresponding prosodic boundary as follows: 150 ms for the phrase commands with magnitude 0.6, 50 ms for the commands smaller than 0.3, and 80 ms for others.

Table 1: Magnitude of phrase command placed at the current prosodic word boundary when phrasal F_0 falls into the 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 2: Magnitude of phrase command placed at the preceding prosodic word boundary when phrasal F_0 falls below the range at the current prosodic word boundary.

Phrasal 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

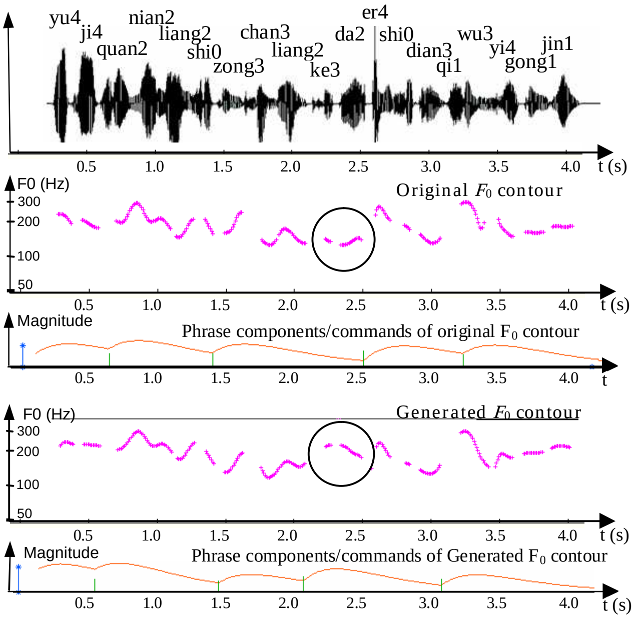

Figure 4: From top to bottom: waveform of synthesized speech, observed F_0 contour of target speech, its phrase components/commands, generated F_0 contour, and its phrase components/commands. The circles indicate the portion causing degraded sounds when the generated F_0 contour is used for speech synthesis.

The generated phrase components are evaluated through a speech synthesis experiment. Sentence F_0 contours were synthesized by superposing tone components of original utterances onto the generated phrase components, to prevent errors in tone component prediction affecting the evaluation.

The pause and phone durations were copied from the original utterances. Speech synthesis was conducted by replacing original F_0 contour with synthesized F_0 contour using TD-PSOLA scheme [12].

Figure 4 shows the waveform of synthesized speech for "yu4 ji4 quan2 nian2 liang2 shi0 zong3 chan3 liang4 ke3 da2 er4 shi0 dian3 qi1 wu3 yi4 gong1 jin1 (It is estimated that the output of grain can be improved to 2.075 billion kilograms in the whole year)," together with original F_0 contour (middle) and synthesized F_0 contour (bottom). Although the difference between two contours seems minor, results of a listening test indicated a considerable degradation in the synthetic speech quality. Since the phrase components generated by the rule do show no unnatural movements (even though they are different form the original ones), the reason of the degradation is considered to be the mismatch between phrase components and tone components, which can be suppressed by predicting tone components using information of phrase components (see section 3).

3. Generation of tone component

3.1. Tone nucleus model

In Mandarin, there are four lexical tones attachable to a syllable. They are referred to as T1, T2, T3 and T4, and are characterized by 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.

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 [8]. 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.

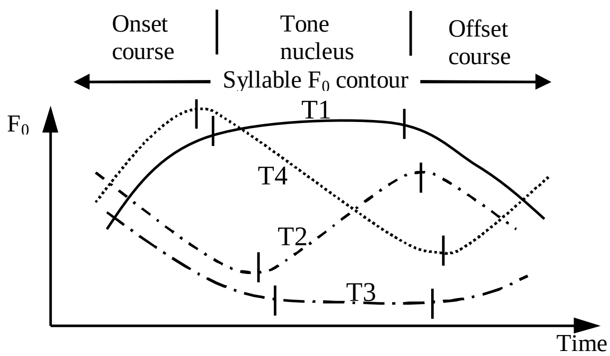

Figure 5: Tone nuclei for the four lexical tones.

Figure 5 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.

3.2. Original method of tone component generation

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 [13]. 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.

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, which can be extracted from input text, such as phonemic constitutions of syllables, number of syllables in words, depths of syntactic boundaries, and so on (Table. 3). In the current paper, labels attached to the corpus were used as these inputs. The inputs also include the information of generated phrase components, such as number of syllables in current phrase, magnitude of phrase command, and so on (two-step scheme). 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. (See section 4.)

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 11 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.

The same 100 news utterances used to construct rules for phrase component generation in section 2 were again used to train the method. 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.

The F_0 contour in Figure 6 shows one when the generated tone components are superposed on the rule-generated phrase components in Figure 4. It is clear that the F_0 contour of the circled part comes close to that of the target utterance.

Table 3: Inputs to the predictors.

Inputs to the predictor	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
Position of syllable in current phrase	Natural num.
Number of syllables in current phrase	Natural num.
Number of phases in current breath group	Natural num.
Position of phrase in current breath group	Natural num.
Position of breath group in sentence	Natural num.
Current phrase command magnitude	Continuous
Timing of current phrase	Continuous

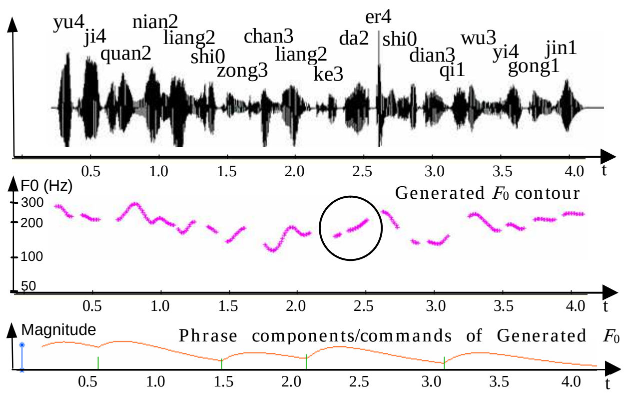

Figure 6: Waveform of synthesized speech (top) together with F_0 contour generated by two-step scheme (middle) and its phrase components/commands (bottom) for the same utterance shown in Figure 4.

For several sentences, speech synthesis was conducted using generated F_0 contours. The listening test for synthetic speech indicated the validity of the proposed method. However, when F_0 contours were predicted directly without decomposing them into phrase and tone component, the resulting synthetic speech was evaluated to have quality slightly higher than that by the two-step scheme [13]. One possible reason for this situation is that the two-step scheme causes occasional degradation around the phrase component initial. 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 may 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 of tone component generation

The improved method for tone component generation differs from the original version as follows in its way of representing tone components of tone nuclei [14]:

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. (No change for T0.)

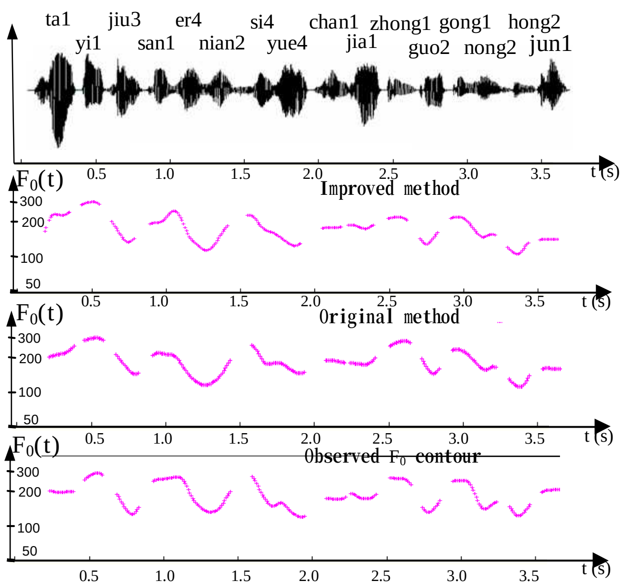

Figure 7: From top to bottom: original waveform, F_0 contour generated by the improved method, one by original method, and one extracted from original speech.

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 T0, 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.

Figure 7 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.

3.4. Listening experiment

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. For this purpose, 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. A listening test was 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).

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 Figure 8.

Although the scores fluctuate among listeners, they 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.

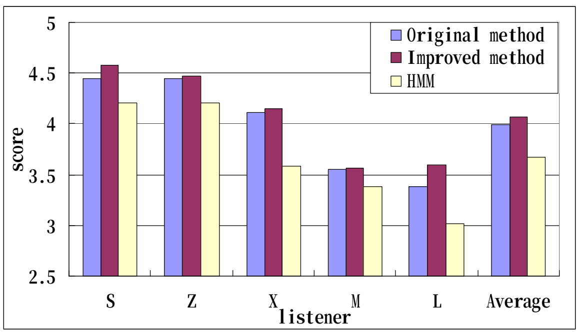

Figure 8: Result of the listening test on the quality of prosody.

4. Experiments on speech synthesis

To investigate the validity of the proposed method of F_0 contour generation when applied in a TTS system, a full speech synthesis system was constructed using the HMM-based speech synthesis (instead of TD-PSOLA) [15]. The phone HMMs were Mel-cepstrum based. It consists of the following processes:

1. Analyze the input text to extract information necessary for speech synthesis. The information is the same as the one used in the F_0 contour generation.
2. Predict phone durations and short pauses using a decision tree.
3. Generate F_0 contours by the (improved) method.
4. Generate 24-order mel-cepstrum coefficients and make the voice/unvoiced decision for each frame by the HMM-based speech synthesis.
5. Generate speech waveform using MLSA filter.

Hundred and seventy utterances from the same female speaker were added to the 100 utterances used in the previous sections, resulting in 270 utterances, which included 15392 phones and short pauses, were used to train the decision tree for duration prediction and the (context dependent) phone HMMs. Firstly, an experiment of duration prediction was carried out. For comparison, durations were also predicted by the HMM-based method, where durations were calculated from probability of state transitions. Although phone HMMs are usually trained after concatenating them without apparently using phoneme boundary information of the training corpus (concatenated training), they are also trained using the manually extracted phoneme boundaries as constraints. This is because the phoneme boundary information is used to train the decision trees. Eight hundreds and fifty five phones and short pauses for 30 utterances, which were not used for the training, were predicted by the decision tree-based method and the two-versions of HMM-based method, respectively. The root mean square (RMS) errors between observed durations and predicted ones are shown in Table 4. The result shows advantages of the decision tree-based method over the HMM-based methods.

Table 4: Result of duration prediction.

Method	Decision tree	HMM (With boundary information)	HMM (Concatenated training)
Error(s)	0.017	0.021	0.028

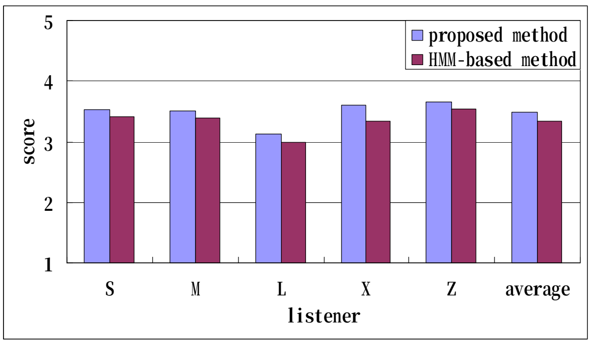

Figure 9: Result of the listening test on synthetic speech quality.

Speech synthesis was conducted for the 30 utterances used for the duration prediction by the two speech synthesis systems: one using the proposed methods of F_0 contour generation, and the other (full HMM-based speech synthesis system) achieved using the speech synthesis toolkit (HTS) [16]. Again the five native speakers of Mandarin were asked to evaluate the naturalness of synthetic speech using the five-point scoring. The result of evaluation shown in Figure 9 clearly indicates that the developed system can generate speech with higher naturalness than the HMM-based one.

5. Word emphasis

Although word emphasis is not handled explicitly in most of current speech synthesis systems, its control becomes important in many situations, such as when the systems are used for generating reply speech in spoken dialogue systems: words conveying key information to the user's question need to be emphasized. Word emphasis associated with narrow focus in speech can be achieved by contrasting the F_0's of the word(s) to be focused from those of neighboring words. This contrast can be realized by placing the word(s) at the phrase component initial, by increasing the accent/tone command amplitudes of the word(s), and by decreasing the accent/tone command amplitudes of the neighboring words. Way of using these three controls maybe different from language to language. Our observation of Mandarin speech indicated the first one being dominant [13]. Since amplitudes of tone commands generally take larger values when they are placed at the phrase command initial, the second and the third controls are somewhat realized automatically.

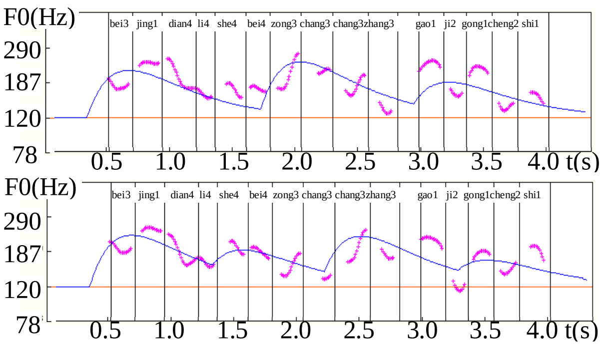

Figure 10: Generated F_0 contours for "bei3 jing1 dian4 li4 she4 bei4 zong3 chang3 chang3 zhang3, gao1 ji2 gong1 cheng2 shi1. ((He is) the director of Beijing Power Equipment Group and senior engineer.)" The first and the second panels show when "zhong2 chang2" and "chang2 zhang3" are emphasized, respectively. Stars indicate generated F_0 contours, while solid curves indicate phrase components.

Table 5: Results of listening test.

Testee	W	Z	S	X	Average
Correctness of focus position	86.7%	83.3%	80.0%	76.7%	81.6%
Score of evaluation	4.30	4.77	4.42	4.31	4.45

Ten sentences, which were different from the 100 sentences used to train the method, were prepared. For each sentence, focuses were placed one of 3 pre-selected words. A phrase command was inserted immediately before the word to be emphasized. After generating other phrase commands by rule, tone commands were predicted by the two-step scheme. By doing so, 3 different F_0 contours were generated for a sentence. TD-PSOLA type speech synthesis was then conducted by substituting the original F_0 contours to the generated ones. Totally, 30 test utterances were synthesized. For the phone durations, we used the original ones extracted from the target speech.

These 30 synthetic utterances were randomly presented to four native speakers of Mandarin, who were asked to mark the word where he/she perceived an emphasis. The marked parts coincided with the original emphasis assignment in 81.6 % on average. This result indicates that an appropriate emphasis control is achieved. Quality of the synthetic speech was also checked in the same way (in 5-rank scoring) as explained in sections 3 and 4. The result in Table 5 again confirms that a good quality is obtainable by the two-step scheme. If we compare F_0 contours shown in Figure 10, it is clear that tone components are generated differently for different phrase components.

Surely, more precise control of F_0 contours can be realized for word emphasis by training the binary decision trees using corpus with word emphasis. However, we should note that focus control in this section is realized without such a corpus. This comes from the ability of "flexible" F_0 contour control of the proposed method.

6. Conclusion

A method was proposed for synthesizing sentence F_0 contours of Mandarin speech. It first generates phrase components in a rule-based way, and then predicts tone components through a corpus-based method. By adding the information of phrase components to the inputs for the predictors of tone components, the both components are time-aligned. A full speech synthesis system was realized using HMM-based speech synthesis. Listening experiments on synthetic speech indicated that a better speech quality was realized by the proposed method as compared to generating F_0 contours by the HMM-based speech synthesis. It was also shown that, by using the 2-step scheme, an empirical control of word emphasis is possible still keeping a good quality in synthetic speech. Future research includes realization of various styles in synthetic speech by the proposed method.

7. Acknowledgement

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

8. References

[1] Lee, L.-S.; Tseng, C.-Y.; Hsieh, C.-J., 1993. Improved tone concatenation rules in a formant-based Chinese text-to-speech system, IEEE Trans. on Speech and Audio Processing, 1(3), 287-294.

[2] Chen, S.; Hwang, S.; Y. Wang, 1998, An RNN-base prosodic information synthesizer for Mandarin text-to-speech, IEEE Trans. on Speech and Audio Processing, 6(3), 226-239.

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

[4] Tokuda, K.; Masuko, T.; Miyazaki, N.; Kobayashi, T., 1997, Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, Proc. IEEE ICASSP, 229-232.
[5] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J. Acoust. Soc. Japan (E), 5(4), 233-242.
[6] Hirose, K.; Fujisaki, H., 1993. A system for the synthesis of high-quality speech from texts on general weather conditions, IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, E76-A(11), 1971-1980.
[7] Gu, W.; Hirose, K.; Fujisaki, H., 2006. A general approach for automatic extraction of tone commands in the command-response model for tone languages. Proc. Speech Prosody, 561-564.
[8] Zhang, J.; Hirose, K., 2004, Tone nucleus modeling for Chinese lexical tone recognition, Speech Communication, 42(3-4), 447-466.
[9] Sun, Q.; Hirose, K.; Gu, W.; Minematsu, N., 2005, Generation of fundamental frequency contours for Mandarin speech synthesis based on tone nucleus model, Proc. Interspeech-Eurospeech, 3265-3268.
[10] Hirose, K.; Sato, K.; Asano, Y.; Minematsu, N., 2005, Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis, Speech Communication, 46, 3-4, pp.385-404.
[11] Sun, Q.; Hirose, K.; Gu, W.; Minematsu, N., 2006, Rule-based generation of phrase components in two-step synthesis of fundamental frequency contours of Mandarin," Proc. International Conference on Speech Prosody, 561-564.
[12] Moulines E.; Charpentier, F., 1990, Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication, 9, pp.453-467.
[13] Sun, Q.; Hirose, K.; Minematsu, N., 2007, Two-step generation of Mandarin F_0 contours based on tone nucleus and superpositional models, Proc. ISCA Workshop on Speech Synthesis (SSW-6), 154-159.
[14] Sun, Q.; Hirose, K.; Minematsu, N., 2008, Improved prediction of tone components for F_0 contour generation of Mandarin speech based on tone nucleus model, Proc. International Conference on Speech Prosody, to be published.
[15] Sun, Q.; Hirose, K.; Minematsu, N., 2008, Improved tone component prediction of tone nucleus for F_0 contour generation of Mandarin speech, Proc. International Workshop on Nonlinear Circuits and Signal Processing, pp.112-115.
[16] HMM-based Speech Synthesis System (HTS): http://hts.sp.nitech.ac.jp/