ARCHIVES OF ACOUSTICS
32, 1, 41–50 (2007)

SYNTHESIS OF FUNDAMENTAL FREQUENCY CONTOURS FOR STANDARD CHINESE BASED ON SUPERPOSITIONAL AND TONE NUCLEUS MODELS

Keikichi HIROSE, Qinghua SUN, Nobuaki MINEMATSU

University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
e-mail: {hirose, qinghua, mine}@gavo.t.u-tokyo.ac.jp

(*received October 16, 2006; accepted December 28, 2006*)

A method for generating sentence F_0 contours of Standard Chinese speech is developed. It is based on superposing tone components on phrase components in logarithmic frequency. While tone components are language specific, phrase components are assumed to be more language universal. Taking this situation into account, the method treats two kinds of components differently. The tone components are generated by concatenating F_0 patterns of tone nuclei, which are predicted by a corpus-based scheme, while the phrase components are generated by rules. Experiments on F_0 contour generation were conducted using 100 news utterances by a female speaker. First experiments were conducted on the generation of tone components, with phrase components of the original utterances being used unchanged. The results showed that the method could generate F_0 contours close to those of target speech. Speech synthesis was conducted by substituting original F_0 contours to generated ones by TD-PSOLA. A high score 4.5 in 5-point scale was obtained on average as the result of listening experiments on the quality of synthetic speech. Second experiments were on the generated phrase components, with the tone components extracted from the original utterances. Although the synthetic speech with generated F_0 contours sounded mostly natural, there were occasional "degraded sounds", because of mismatch between the phrase and the tone components. To cope with the mismatch, a two-step method was developed, where information of the phrase contours was used for the prediction of tone components. Validity on the method was shown through perceptual experiments on synthesized speech.

Keywords: speech synthesis, F_0 contour generation, Standard Chinese, superpositional model, tone nucleus model, rule- and corpus-based generation.

1. Introduction

Recently, novel schemes, such as selection-based waveform concatenation method, largely improved the quality of synthetic speech. However, the improvements are mainly on segmental features of speech, and there still remain major problems if the prosodic features are viewed.

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 Chinese speech. As is well known, Standard Chinese 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 Chinese sentences more complicated than non-tone languages like English, Japanese and so on. Therefore, control of F_0 contours becomes an important issue in Chinese 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 Chinese 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 Standard Chinese by introducing negative commands [5]. The Chinese 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, 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. In the case of Japanese, we have already developed such a system, where the training corpus with F_0 model parameters is prepared automatically [6]. Unfortunately, this is currently not the case for Chinese. 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 with 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 speech synthesis of Standard Chinese, 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. The current paper is focused on the generation of tone components.

Phrase components can be generated by a rather simple set of rules 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 Chinese speech synthesis.

2. Tone nucleus model

A syllable of Chinese can be divided 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 Standard Chinese, 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, 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 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 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.

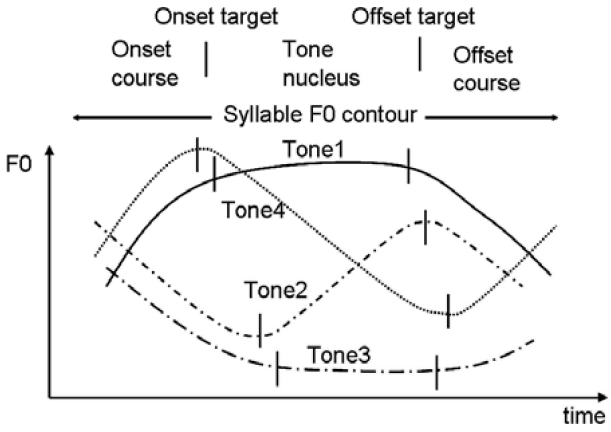

Fig. 1. Tone nuclei for the four lexical tones.

3. Generation of tone components

3.1. Method

As already mentioned in Sec. 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 tone components are generated through 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 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 such as 3-rd 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 with inputs shown in Table 1. The stop threshold, minimum number of samples per leaf node, was set at 20. 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 to 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 Sec. 2 and careful observation of F_0 contours of Standard Chinese speech, the tone nuclei are defined for each tone type as follows; high-level flat F_0 for T1, rising F_0 for T2, low-level flat F_0 for T3, and falling F_0 for T4. 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.

3.2. Experiments

Experiments were conducted on the generation of F_0 contours. The speech data are 100 news utterances by a native female speaker of Standard Chinese. Each utterance consists of about 50 syllables. A smoothing process based on the piecewise 3-rd 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 2.

Table 2. 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 Sec. 3.1, 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 2.

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.

We selected 9 utterances, which consist of syllables for testing only. Their F_0 contours were changed by TD-PSOLA method to those generated by superposing the tone components produced by the proposed method on the original phrase components of the utterances. As clearly shown in Fig. 2, the generated F_0 contour is quite close to the original (observed) F_0 contour of the utterance. In Fig. 2, the RMS error of F_0 in

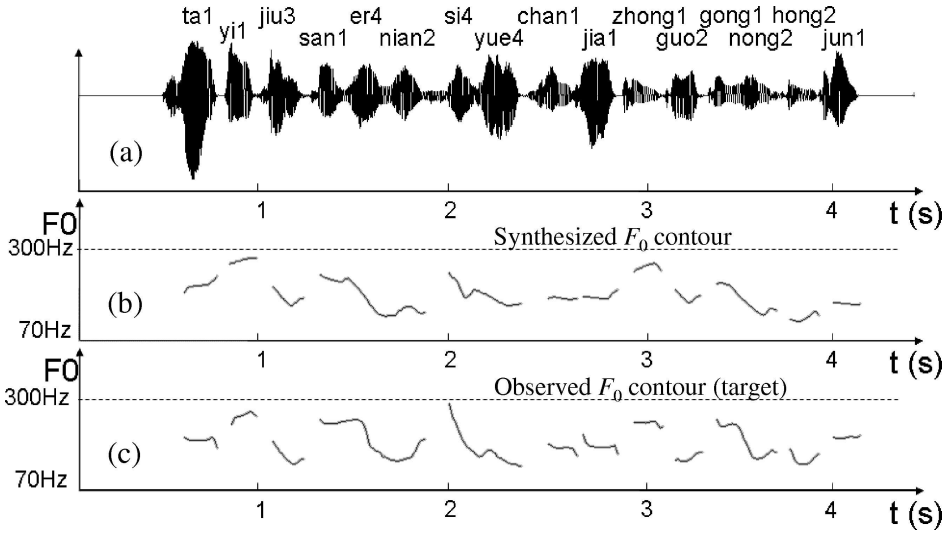

Fig. 2. From top to bottom: waveform of synthesized speech, F_0 contour generated by the proposed method, and observed F_0 contour of target speech. The utterance is "ta1 yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1" (He joined the Chinese Workers' and Peasants' Red Army in April 1932).

log domain between generated speech and original 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 of Chinese. The average score was 4.5, indicating high naturalness of the synthetic speech.

4. Effect of phrase components

Although the phrase components are considered to be mostly language universal, there still are margins of language specific factors. Analysis of F_0 contours of the 100 news utterances in Sec. 3.2 indicates more frequent phrase components as compared to Japanese; average interval between two adjacent phrase commands is 7 syllables for Chinese, while it is 15 syllables for Japanese. In the case of Chinese, phrase components should keep certain values so that an F_0 contour has a margin for downward move corresponding to tone components with negative commands. For the speech data analyzed, the phrase components mostly start from above 150 Hz. Based on these observations, a set of simple rules were constructed for phrase component generation [11].

When tone components extracted from the original utterances are superposed on the generated phrase components, the resulting F_0 contours are in most cases close to the observed ones. However, there are occasional "strange" cases. Figure 3 shows such a case; synthesized waveform together with original F_0 contour (panel (b)) and

synthesized F_0 contour (panel (c)). Although the difference between two contours seems minor, a listening test indicated a considerable degradation in the speech quality.

Reason of this degradation is considered to be due to a mismatch between phrase components and tone components. Since these two types of components are tightly related, such as phrase initial tone components having larger dynamic ranges, they cannot be generated independently. To cope with this problem, a two-step method was developed, where information of generated phrase components was used for the prediction of tone components. Table 3 indicates parameters added to the inputs of the tone component predictors. As indicated in panel (d) of Fig. 3, an F_0 contour closer to original one was obtained by the method.

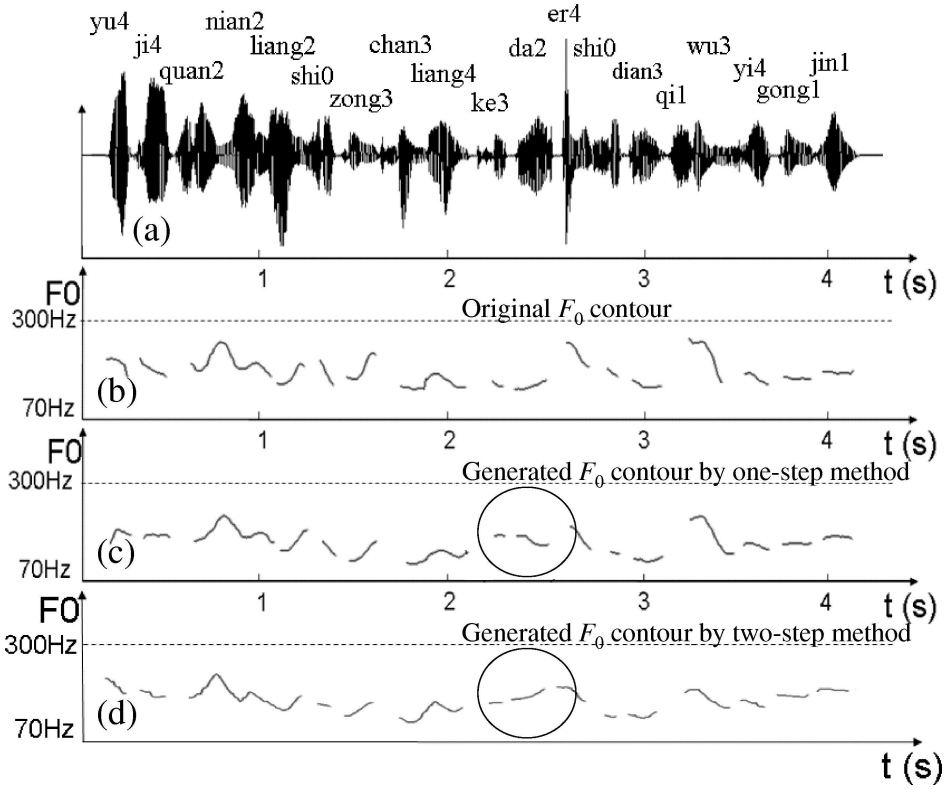

Fig. 3. From top to bottom: waveform of synthesized speech, observed F_0 contour of target speech, F_0 contour generated by the one-step method, and F_0 contour generated by the two-step method. The utterance is "yu4 ji4 quan2 nian2 liang2 shi0 zong3 chan3 liang4 ke3 da2 er4 shi0 dian3 qi1 wu3 yi4 gong1 jin1" (It is estimated that the output of grain can be improved to 2.075 billion kilograms in the whole year).

Speech synthesis was again conducted by changing the original F_0 contours to the F_0 contours generated by the proposed method. The generation was conducted in three cases as shown in Table 4. In the table, "one-step" depicts one-step method, where no information on the phrase components is used for the prediction of tone components.

Table 3. Inputs added to the predictor for the two-step method.

Input	Category
Position of syllable in current phrase	Integer
Number of syllables in current phrase	Integer
Position of phrase in current breath group	Integer
Number of phrases in current breath group	Integer
Position of breath group in sentence	Integer
Current phrase command magnitude	Continuous
Timing of current phrase	Continuous
Mean of phrase F_0	Continuous
Slope of phrase F_0	Continuous

Nine sentences were synthesized, each with 3 versions of F_0 contours generated by the three cases. Total of 27 utterances were randomized and presented to 5 native speakers of Chinese, who were asked to evaluate their quality in the five-point score. Average scores for each evaluator were listed in Fig. 4. Scores above 4 are obtained by the case *r-gt* (generated phrase plus two-step method) for all the 4 listeners. It should be noted that the scores are almost the same as the case *o-gt* using original phrase components. These results indicate that the proposed rule-based method for phrase component generation works quite well under the two-step scheme.

Table 4. Combinations of method of generation for phrase components and tone components.

Code	Phrase component	Tone component
o-gt	Original	Two-step
r-go	Generated	One-step
r-gt	Generated	Two-step

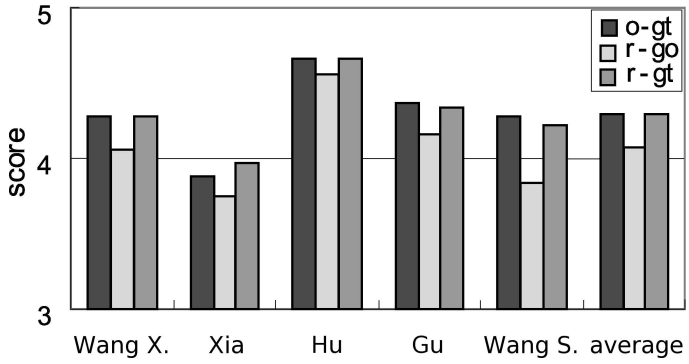

Fig. 4. Result of listening test of synthetic speech.

5. Conclusion

A new method of generating F_0 contours for speech synthesis of Standard Chinese 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. In order to keep timing correspondences between phrase and tone components, a two-step method was developed. Perceptual experiments on the synthetic speech with generated F_0 contours showed that highly natural speech was obtainable by the method. Our next step is to realize a "flexible" control in F_0 contour generation.

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

References

[1] CHEN S., HWANG S., WANG Y., *An RNN-base prosodic information synthesizer for Mandarin Text-to-speech*, IEEE Trans. on Speech and Audio Processing, **6,** 3, 226–239 (1998).

[2] TAO J., CAI L., *Clustering and feature learning based F_0 prediction for Chinese speech synthesis*, Proc. Int. Conference on Speech and Language Processing, pp.2097–200, Denver 2002.

[3] NI J., HIROSE K., *Synthesis of fundamental frequency contours of standard Chinese sentences from tone sandhi and focus conditions*, Proc. Int. Conference on Speech and Language Processing, Beijing, pp. 195–198, 2000.

[4] FUJISAKI H., HIROSE K., *Analysis of voice fundamental frequency contours for declarative sentences of Japanese*, J. Acoust. Soc. Japan (E), **5,** 4, 233–242 (1984).

[5] FUJISAKI H., HIROSE K., HALLE P., LEI H., *Analysis and modelling of tonal features in polysyllabic words and sentences of the standard Chinese*, Proc. Int. Conference on Speech and Language Processing, Kobe, pp. 841–844, 1990-10.

[6] HIROSE K., SATO, K., ASANO, Y., AND MINEMATSU, N., *Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis*, Speech Communication, **46,** 3–4, 385–404 (2005).

[7] GU W., HIROSE K., FUJISAKI H., *Automatic extraction of tone command parameters for the model of F_0 contour generation for Standard Chinese*, IEICE Trans. Information and Systems, **E87-D,** 5, 1079–1085 (2004).

[8] ZHANG J., HIROSE K., *Tone nucleus modeling for Chinese lexical tone recognition*, Speech Communication, **42,** 3–4, 447-466 (2004).

[9] HIROSE K., 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, **E76-A,** 11, 1971–1980 (1993).

[10] NARUSAWA S., MINEMATSU N., HIROSE K., FUJISAKI H., *A method for automatic extraction of model parameters from fundamental frequency contours of speech*, Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing, pp.509–512, Orlando 2002.

[11] SUN Q., HIROSE K., GU W., MINEMATSU N., *Rule-based generation of phrase components in two-step synthesis of fundamental frequency contours of Mandarin*, Proc. International Conf. on Speech Prosody, pp.561–564, Dresden 2006.