Development of an F0 Control Model Based on F0 Dynamic Characteristics for Singing-Voice Synthesis

Takeshi Saitou, Masashi Unoki, and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan

Abstract

A fundamental frequency (F0) control model, which can cope with F0 dynamic characteristics related to singing-voice perception, is required to construct natural singing-voice synthesis systems. This paper discusses importance of F0 dynamic characteristics in singing-voices and demonstrates how strongly they influence singing-voice perception through psychoacoustic experiments. This paper, then, proposes an F0 control model that can generate F0 contours of singing-voices based on these considerations, and a singing-voice synthesis system. The results show that several types of F0 fluctuation – overshoot, vibrato, preparation, and fine-fluctuation – affect the perception and quality of a singing-voice, and that overshoot has the greatest effect. Moreover, the results show that the proposed F0 control model can control F0 fluctuations, and generate F0-contours of singing-voices, and can be applied to natural singing-voice synthesis.

Key words: F0 fluctuation, singing-voice perception, F0 control model, singing-voice synthesis

1 Introduction

Singing and speaking are important ways in human communications to express both linguistic and nonlinguistic information. Speaking- and singing-voice synthesis methods are important in the subject of speech signal processing, and various speech synthesis systems have been proposed. However, most of these systems were proposed for speaking-voice synthesis rather than singing-voice synthesis. To construct a singing-voice synthesis system, it is important to reveal the acoustic features affecting singing-voice perception, and to learn how to control these features for natural singing-voice synthesis.

When singing a song, one tries to express lyrics by changing notes according to a melody without changing individuality of one and one's singing-style. Singers can also sing more artistic songs. Therefore, the characteristics of a singing-voice are more dynamic and complicated than those of a speaking-voice, and these characteristics significantly affect the quality of singing-voices. Especially important is that the naturalness of voices depends on the F0 contour. For example, the F0 contours of singing-voices correspond to the melody of a song, and these contours include several types of fluctuation such as overshoot, vibrato, and fine fluctuations. Overshoot means that F0 transitionally exceeds the target note just after a note change (Mori et al , 2004; de Krom and Bloothooft , 1995). Vibrato is the quasi-periodic modulation of F0 observed in professional singers' voices (Seashore , 1938). These fluctuations are important characteristics regarding the quality of singing-voices (Myers and Michel , 1987; Hakes et al , 1987; Horii , 1989). Moreover, there are fine fluctuations that affect the naturalness of singing-voices in F0 contours (Akagi and Kitakaze , 2000). However, a quantitative assessment of the perceptual influence of fluctuations in singing-voice F0 contours has not been done, even though such fluctuations are likely to be important factors affecting the production of high-quality synthesized speech.

On the other hand, it is important to learn how to generate F0 contours for speech synthesis. Most speech synthesis methods are based on the source-filter model, so F0s related to source information and formant information related to filter characteristics are separately used in the model. Consequently, for speech synthesis, F0 contours are important components, and many methods for controlling them have been proposed, for example, the Fujisaki-model (Fujisaki and Tatsumi , 1981) that is a useful F0 control model for a speaking-voice and this model can be used to control the F0 contours of speaking-voices. However, the F0 control models proposed so far cannot be used to control and generate F0 contours of singing-voices. This is because the dynamic range of the F0 contours in singing-voices is wider than those of speaking-voices and the F0 fluctuations in singing-voices are larger and more rapid than those of speaking-voices. To devise a singing-voice synthesis method, we have to develop a way of controlling the acoustic parameters, especially F0 contours including fluctuations that affect singing-voice perception.

To construct a high-quality singing-voice synthesis system, we need first to investigate F0 fluctuations affecting singing-voice perception and then construct an F0 control model for a singing-voice. This paper reveals F0 fluctuations as significant characteristics of a singing-voice by investigating how strongly they affect singing-voice perception and quality through psychoacoustical experiments. Moreover, this paper develops an F0 control model that can generate an F0 contour of a singing-voice by controlling F0 fluctuations which affect singing-voice perception, as well as a singing-voice synthesis method based on our F0 control model.

The rest of this paper is organized as follows. Section 2 describes our analysis of the singing-voices F0 contours and the F0 fluctuations that are peculiar to a singing-voice. Section 3 describes how strongly F0 fluctuations influence the quality and perception of singing-voices by carrying out pychoacoustical experiments. Section 4 describes an F0 control model that can control these fluctuations and generate F0 contours of singing-voices. It also describes the singing-voice synthesis method using the proposed F0 control model. Section 5 is a brief conclusion.

2 Analysis of F0 fluctuations

The F0 contours of voices are used to express linguistic, para-linguistic, and nonlinguistic information. The F0s of speech contain slow and large fluctuations related to prosodic information, and rapid and fine fluctuations related to the naturalness of speech (Fujisaki et al , 2000). The F0 characteristics of a singing-voice differ from those of a speaking-voice, and three particular characteristics are known to appear in the F0 contours of singing-voices (Nakayama and Kobayashi , 1996):

(a) The dynamic range of F0 contours is wider than that for a speaking-voice;
(b) The F0 contour corresponds to a note and tends to be stable in a note. The note changes of the F0 contours correspond to the melody of a song; and
(c) There are many F0 fluctuations that are observed only in singing-voices.

These characteristics are peculiar to singing-voices. Characteristics (a) and (b) are static characteristics related to melody. Characteristic (c) is related to fluctuations such as overshoot and vibrato, and many researchers have suggested that these fluctuations are important singing-voice characteristics. To test the validity of these suggestions, this paper extracted F0 fluctuations from the estimated F0 contours of several singing-voices and investigated the relationship between F0 fluctuations and singing-voice perception.

2.1 Singing-voice data

The singing-voices of professional singers in the Western classical style have been extensively investigated. To extract common F0 fluctuations found in all kinds of singing-voices, though, it is important to analyze the singing-voices of amateur as well as professional singers. An investigation of singing-voices

for various vocalisms is also needed. Therefore, this paper used two kinds of singing-voice data set.

One singing-voice data set was obtained from recordings of three female adults singing a Japanese children's song,“ Nanatsunoko ”. They were amateur singers who were taking voice lessons. Each singer was asked to sing the song with only the Japanese vowel /a/, to simplify the experimental conditions. Each song was divided into four parts, and the total number of data was 12. The songs were recorded on DAT with 48-kHz sampling and 16-bit accuracy, and were down-sampled to 20 kHz. This singing-voice dataset is referred to as DATA-1 in this paper.

The other data set was obtained from recordings of various professional singers singing a common lyric “ Kaedeiroduku-Yamanoasawa ”with a free melody (Nakayama , 2004). The singing-voice data selected from the data set were from four sopranos (18 data), two mezzo-sopranos (7 data), one alto (4 data), three tenors (13 data), three baritones (14 data), two pop musicians (6 data), and one enka (Japanese ballad) singer (2 data). The total number of data was 64. This dataset is referred to as DATA-2 in this paper.

2.2 F0 estimation method

To analyze F0 contours that include dynamic and complicated characteristics, especially the F0 contours of singing-voices, an F0 extraction method that can accurately estimate F0 contours was needed. For this, we chose TEMPO in STRAIGHT (Kawahara et al , 1999), having confirmed beforehand that TEMPO can more accurately extract fine fluctuations in the F0 contours than other methods can. Note that it was reported that TEMPO is one of the most useful methods for estimating F0 contours (de Cheveigne and Kawahara , 2001).

2.3 F0 fluctuations in singing-voice

Figure 1 shows an estimated F0-contour of a song in DATA-1. The ordinate indicates log-frequency. Figure 1(a) shows a melody component that represents note changes in the extracted F0. Figure 1(b) shows four F0 fluctuations observed in the F0 contours:

(1) Overshoot: Deflection exceeding the target note after note changes;
(2) Vibrato: Quasi-periodic frequency modulation (5 - 8 Hz);

(3) Fine-fluctuation: Irregular fine fluctuations regarding the modulation frequency and modulation amplitude. The modulation frequency is higher than about 10 Hz, and the modulation amplitude is about 1.2 % of the average F0 value; and
(4) Preparation: Deflection in the opposite direction of a note change observed just before the note changes.

Since F0 changes in a singing-voice are more rapid than in a speaking-voice, vocal cord control in singing is thought to differ from that in speaking. Therefore, some researchers have focused on the F0 characteristics during the note transition. Sundberg reported that the response time in F0 risings was longer than that in F0 fallings by measuring the F0 speed changes (Sundberg , 1979). Kasuya et al. suggested that the F0 of a singing-voice exceeds the target note just after a note change, and called this phenomenon overshoot. de Krom et al. measured the extent of overshoot, and Mori et al. suggested that the extent of overshoot affects the quality of a synthesized singing-voice (Mori et al , 2004; de Krom and Bloothooft , 1995). In our present analysis, we consistently observed not only overshoot but also preparation, which is a deflection of F0 just before a note change, in DATA-1 and DATA-2. Therefore, this paper considers preparation to be a new type of F0 fluctuation.

There are many reports on vibrato as a singing-voice characteristic. Many researchers have studied the occurrence of vibrato in Western classical singing, and reported that the artistic quality of singing was frequently judged according to the presence of vibrato in the sound. Vibrato corresponds to quasi-periodic F0 modulation and is characterized by two parameters: rate and extent (Sundberg , 1987). The vibrato rate specifies the number of undulations per second, and the extent describes how far F0 rises and falls during a vibrato cycle. Seashore reported that the vibrato rate is about 5 - 8 Hz, and that good vibrato is regular in its interval and rate and is consistent around the F0 contour (Seashore , 1938). In our study, vibrato was typically observed not only in Western classical singing, but also in Japanese traditional singing. In addition, the vibrato characteristics of the professional singers tended to be more constant than those of the amateur singers. Based on these results, this paper considered vibrato to be an important singing-voice characteristic.

Akagi and Kitakaze analyzed fine fluctuations in the F0 contours of singing-voice (Akagi et al , 1998; Akagi and Kitakaze , 2000). They reported that fine fluctuations involve the modulation frequency (MF) containing frequency components of up to 20 Hz and modulation amplitudes (MAs) that were 20 cent on average and 100 cent at maximum, which are one-fifth of and equal to the half-tone musical scale, respectively. Moreover, they suggested that these characteristics affect the quality of a singing-voice. Thus, this paper also

focused on fine-fluctuations in singing-voice F0 contours.

3 Importance of F0 fluctuations

Section 2 described four types of F0 fluctuation in singing-voice F0 contours. The importance of each type of F0 fluctuation concerning singing-voice quality has been reported in past studies. However, these studies did not provide results sufficient for a quantitative assessment of the perceptual influence of F0 fluctuations in singing-voice F0 contours. For example, there has been no consideration of what is the most effective F0 fluctuation for singing-voice perception or investigation of the differences in the effect of each F0 fluctuation.

It is important to clarify the effect of F0 fluctuations on singing-voice perception not only for learning how to control singing-voice F0 contours but also for synthesizing natural singing-voices. Thus, this paper attempted to show F0 fluctuations as significant singing-voice characteristics by investigating how strongly each F0 fluctuation influences perception. First, we synthesized singing-voices that are removed each F0 fluctuation from real F0 contours. Second, we investigated the naturalness of synthesized singing-voices through psychoacoustical experiments. We assessed the differences in the effect of each F0 fluctuation and determined which F0 fluctuation had the greatest impact on singing-voice perception.

3.1 Stimuli

Figure 2 shows the singing-voice synthesis method for the experiment. Since the experiment focused on only the effect of individual F0 fluctuations, we synthesized singing-voices by using the Klatt formant synthesizer (Klatt , 1980). This was because the Klatt formant synthesizer could synthesize singing-voices in which F0 information can be modified while the spectrum information and power level are fixed. First, F0 contours were extracted from real singing-voices by using TEMPO. Next, each F0 fluctuation was removed from F0 contours of real singing-voices and re-synthesized the singing-voices using the modified F0s. These synthesized singing-voices were then used as stimuli in psychoacoustical experiments to investigate the effect of each F0-fluctuation on singing-voice perception.

The synthesized singing-voices were

NORMAL: Singing-voice synthesized using the extracted F0 from a real

singing-voice;
NO-OS: Singing-voice with F0 contour removed by replacing the overshoot F0s with the average F0 of the target note;
NO-VIB: Singing-voice with F0 contour removed by replacing the vibrato F0s and with the average F0 during the vibrato period;
NO-PRE: Singing-voice with F0 contour removed by replacing the preparation F0s with the average F0 of the target note; and
SMS: Singing-voice whose F0 was smoothed by an FIR lowpass-filter (cut-off frequency: 5Hz). These stimuli contained only the melody component with all F0 fluctuations excluded.

In all, twenty stimuli were synthesized using DATA-1. Stimuli were paired randomly to make a paired-comparison psychoacoustical experiment. The total number of paired stimuli was 80.

These stimuli were synthesized by setting the formant frequency of the Japanese vowel /a/ to 800, 1200, 2500, 3500, 4500, and 5500 Hz, and each bandwidth was set to 10 % of the corresponding formant frequency. The power level was set to 80 dB. The excitation impulse trains were made as follows. Let us assume the F0 transition with fluctuations is $f_m(t)$. If the pulse is set at time t_n, the next pulse must be set at

$$t_{n+1} = t_n + 1/f_m(t_n), \tag{1}$$

The generated pulse train was filtered to modify each pulse into a Rosenberg wave. The synthesized voices were made by convoluting the response of the synthesizer with the excitation impulse trains.

3.2 Psychoacoustical experiment

Scheffe's method of paired comparison was used to evaluate the naturalness of stimuli described in Sec 3.1. In this experiment, subjects evaluated which stimulus was a more natural singing-voice according to a seven-grade evaluation measure (Figure 3). The " naturalness "of a singing-voice has a multidimensional meaning. However, since this experiment defined a " natural singing-voice " as a synthesized singing-voice that could be perceived as the singing by a human, we defined " naturalness " as the difference between the synthesized singing-voice and real singing-voice sung by a human. If subject perceived stimulus A was closer than stimulus B to human singing, stimulus A was regarded as the more natural singing-voice. Therefore, this paper dealt with the " naturalness " in 1-dimensional. The pair-wise stimuli were presented through binaural headphones at a comfortable sound pressure level. Each paired stimulus was randomly presented to each subject. The subjects

were six graduate students with normal hearing ability. The naturalness of the synthesized singing-voices under each condition was calculated as a function of population.

3.3 Results and discussion

Figure 4 shows the experimental results. The numbers under the horizontal axis indicate the degree of naturalness of a synthesized singing-voice. The results of the F-test confirmed that there were significant differences between all stimuli at a 5 % critical rate. The results indicate that three types of F0 fluctuation – overshoot, vibrato, and preparation – strongly affect singing-voice perception, and overshoot has the greatest effect. This result suggests that overshoot is the most important fluctuation with regard to singing-voice perception. In this experiment, we did not use a synthesized singing-voice with only fine-fluctuations removed; however, it is also an important fluctuation because the naturalness of SMS is lowest. The NO-VIB score may seem to be somewhat high, but no fluctuations can be removed from F0-contours while maintaining the naturalness of the NORMAL voices. Therefore, we dealt with all four types of fluctuation when we constructed our F0 control model.

4 F0 Control Model for Singing-Voices

There are several F0 control models for speaking-voices (Ishizaka and Flanagan , 1972; Fujisaki and Tatsumi , 1981; Moriyama et al , 1996). The Fujisaki model can precisely control and generate F0 contours of a speaking-voice by setting a few control parameters. Since this model represents an F0 contour with a critically damped second-order linear system, it is difficult to control F0 fluctuations, especially overshoot, and generate F0 contours of a singing-voice. Mori et al. suggested a method for controlling overshoot in the F0 transition of singing. However, this model did not consider control of vibrato, preparation, and or fine fluctuations. This paper therefore developed a method that can generate F0 contours of singing-voices by considering how to control four F0 fluctuations affecting singing-voice perception.

4.1 Schematic graph of F0 control model

To construct an F0 control model, we considered the following.

(I) The F0 control model should deal with four types of F0 fluctuation – over-

shoot, vibrato, preparation, and fine-fluctuation – based on melody components;
(II) The F0 control model should control F0 fluctuations by determining a few control parameters; and
(III) The F0 control model should be applicable to natural singing-voice synthesis.

Figure 5 shows a schematic graph of the proposed F0 control model. The model input is the melody component described by a sum of step functions. The model generates F0 contours by adding the four types of F0 fluctuation to melody component. Each fluctuation is represented as follows.

Overshoot: Second-order damping model.
Vibrato: Second-order oscillation model (no-loss).
Preparation: Second-order damping model.
Fine-fluctuation: Irregular rapid oscillation with the modulation frequency of higher than 10 Hz and the modulation amplitude of 5 Hz at maximum.

The transfer function $H(s)$ of the second-order system is represented as

$$H(s) = \frac{K}{s^2 + 2\zeta\Omega s + \Omega^2}, \tag{2}$$

where s is the Laplace operator, Ω is a natural frequency, ζ is a damping coefficient, and K is the proportional gain. Here, the impulse response $h(t)$ of $H(s)$ can be obtained as

$$h(t) = \begin{cases} \frac{K}{2\sqrt{\zeta^2-1}}(\exp(\lambda_1\Omega t) - \exp(\lambda_2\Omega t)), & |\zeta| > 1 \ (a) \\ \frac{K}{\sqrt{\zeta^2-1}}\exp(-\zeta\Omega t)\sin(\sqrt{1-\zeta^2}\Omega t), & |\zeta| < 1 \ (b) \\ Kt\exp(-\Omega t), & |\zeta| = 1 \ (c) \\ \frac{K}{\Omega}\sin(\Omega t), & |\zeta| = 0 \ (d) \end{cases} \tag{3}$$

where ,$\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$ and $\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$. Equation (3)-(a) represents a solution to the second-order exponential damping model, Eq. (3)-(b) represents a solution to the second-order damping model, Eq. (3)-(c) represents a solution to the second-order critical oscillation model, and Eq. (3)-(d) represents a solution to the second-order oscillation (no-loss) model. Thus, Eq. (3)-(b) is used for the overshoot and preparation, and these are controlled just after and before F0 changes respectively. Vibrato is controlled using Eq. (3)-(d) when note stable. These F0 fluctuations were controlled by determining control parameters Ω, ζ, and K.

Fine-fluctuation is generated by a lowpass filtering and normalizing of white noise, and then is added into whole of F0 contour.

4.2 Optimal control parameter values

To correctly generate the F0 contour of a singing-voice, it is important to determine the optimal parameter values. Therefore, we have determined adequate control parameter values for each type of F0 fluctuation by analyzing the singing-voice data sets described in Sec 2.1.

In particular, a nonlinear least-squared-error method (Press , 1988) was used to minimize the error E between the real F0 contour $x(t)$ and the generated F0 contour $y(t)$ by the proposed F0 control model. The equation used to determine the optimal parameter values was

$$E = \sqrt{\frac{a}{N} \sum_{m=M}^{M+N} (x(mT) - y(mT))^2}, \tag{4}$$

where, $T = 1/fs$ (fs is sampling frequency). M and N are start time and duration for determining optimal control parameter values of each F0 fluctuations. In addition, when optimizing the control parameter values for overshoot and preparation, we determined that the control parameter values of K should equal those of Ω. The obtained parameter values are shown in Table 1.

With regard to controlling overshoot and preparation, parameter Ω controls the speed of the F0 transition and the duration of the overshoot, while parameter ζ controls the extent of the overshoot and preparation. As shown in Table 1, the overshoot extent is larger than that of preparation, and the overshoot duration is shorter than that of preparation.

As mentioned in Sec 2.3, vibrato is characterized by two parameters: the rate and extent. In the proposed F0 control model, Ω controls the vibrato rate and K controls the vibrato extent. Thus, we determined the optimal parameter values by analyzing these two parameters using DATA-1 and -2. The analysis results show that the vibrato rate is about 5.5 Hz and the extent is about 5.2 % (the ratio of modulation depth vs. the average F0 in the vibrato cycle).

To control fine-fluctuation, the cutoff frequency and damping rate of the lowpass filter is 10 Hz and -20 dB/oct, and normalizing of amplitude is 5 Hz.

These values were based on the considerations explained in Sec 2.3. From the analysis results, we then determined the optimal parameters for controlling F0 fluctuations as shown in Table 1.

5 Evaluation of the F0 control model

Figure 6 shows F0 contours generated by the proposed F0 control model from melody components obtained from the F0 contour in Fig. 1. The control model clearly can control each form of F0 fluctuation and generate an F0 contour of a singing-voice. For the evaluation of proposed F0 control model, we applied the model to two types of synthesis method. First, a singing-voice synthesis using the Klatt formant synthesizer was constructed for evaluating the F0 contour generated by the proposed F0 control model. Second, a singing-voice synthesis system was constructed using STRAIGHT (Kawahara et al , 1999) for confirming applicability of the proposed model to natural singing-voice synthesis system.

5.1 Singing-voice synthesis using Klatt formant synthesizer

As mentioned in Sec 3.1, the Klatt formant synthesizer could synthesize singing-voices in which only the differences in F0 contours were reflected. Therefore, we firstly synthesized singing-voices using the Klatt formant synthesizer shown in Fig. 7 for evaluating F0 contour generated by the model. Basically, the synthesis procedure was the as described in Sec 3.1, except for using F0 contours generated by the F0 control model.

5.1.1 Psychoacoustical experiments

To investigate the quality of synthesized singing-voices using generated F0 contour, we carried out psychoacoustical experiments according to the following procedure. We added each F0 fluctuation to the melody component by using the F0 control model and synthesized singing-voices using those F0s. We presented them to subjects who judged their naturalness.

Six stimuli were used in the experiment:

NORMAL: Synthesized singing-voice using extracted F0 of a real song.
SYN-All: Synthesized singing-voice using F0 with all F0 fluctuations added to the melody component by using the F0 control model.

SYN-OS: Synthesized singing-voice using F0 with only overshoot added to the melody component.
SYN-PRE: Synthesized singing-voice using F0 with only preparation added to the melody component.
SYN-VB: Synthesized singing-voice using F0 with vibrato and fine fluctuations added to the melody component.
SYN-BASE: Synthesized singing-voice using only the melody component.

Table 1 lists control parameters values for all F0 fluctuations. The song was the Japanese children 's ballad " Nanatsunoko, " as mentioned in Sec. 2.1. These stimuli were synthesized under the same conditions as described in Sec 3.1. To evaluate the naturalness of these stimuli, we used Scheffe 's paired comparison with the same procedure and conditions as in Sec. 3.

5.1.2 Results and Discussion

Figure 8 shows the relationship of the naturalness of each stimulus, and significant differences between all stimuli were confirmed by F-test (critical rate is 5 %). The results show that the naturalness of the synthesized singing-voice using only the melody component was the lowest. This was because the evaluation value of the BASE was the smallest in all stimuli. However, the naturalness of the synthesized singing-voice was clearly increased by adding each type of F0 fluctuation to the melody component, and the quality of SYN-ALL is almost the same as that of NORMAL. Therefore, it is clear that the proposed F0 control model can deal with each F0 fluctuation and generate F0 contours adequately.

Figure 4 indicates that overshoot has the greatest effect on singing-voice perception, and that the effect of preparation was slightly larger than that of vibrato. Figure 10 shows a similar result that the effect of each F0 fluctuation is larger in the order of overshoot, preparation, and vibrato. Although it is impossible to compare Figs. 4 and 8 directly because the synthesized singing-voices represented in Fig. 4 contained characteristics that cannot be eliminated from real F0 contour or cannot be controlled by the proposed F0 control model, it is clear that the effect of each F0 fluctuation on singing-voice perception as shown in Fig. 4 is reconfirmed by the experimental results in Fig. 10.

However, since it was difficult for the Klatt formant synthesizer to control spectrum information, there were two problems in this experiment. One is that the qualities of all stimuli were not good, and even the naturalness of NORMAL was far from that of a real singing-voice. Therefore, it was impossible to evaluate F0 contours generated by the proposed F0 control model by com-

paring synthesized singing-voices and real singing-voices. The other problem is that it was difficult for this synthesis method to synthesize voices singing lyrics. Therefore, we improved the singing-voice synthesis method by using a high-quality vocoder, STRAIGHT (Kawahara et al , 1999).

5.2 Singing-voice synthesis using STRAIGHT

Figure 9 shows the singing-voice synthesis using STRAIGHT. This improvement was aimed at evaluating the generated F0 contour by comparing synthesized singing voices against real singing-voices, and at extending the proposed F0 control model so that it could deal with lyrics. The method consisted of two blocks: the F0 control model and STRAIGHT, instead of the Klatt synthesizer.

STRAIGHT consists of TEMPO, STRAIGHT-core, and SPIKES. TEMPO is the F0 estimation block, and SPIKES is the excitation-pulse generator for the source information using the F0 contour. STRAIGHT-core estimates the spectrum envelope by using F0-adaptive time-frequency smoothing to eliminate periodic interference, and it synthesizes sound by using the spectrum envelope and excitation pulses. The spectrum envelope is not manipulated in the synthesis process.

5.2.1 Psychoacoustical experiments

The psychoacoustical experiments were carried out to evaluate singing-voices that were synthesized by improved synthesis method. The procedure and conditions of experiments were almost the same as those of the previous experiment.

The control parameters values and chosen song were the same as before. The stimuli were also basically the same. However, since the spectrum information of all stimuli was identically the real song spectrum information, the quality of NORMAL was almost the same as that of the real song. Therefore, a proper comparison of the synthesized singing-voices (SYN-ALL, SYN-OS, SYN-PRE, SYN-VB, and SYN-BASE) against real singing-voices (NORMAL) could be done.

5.2.2 Results and discussion

The experimental result is shown in Fig. 10, and the significant differences between all stimuli were also confirmed by F-test (critical rate is 5 %). From the figure, it is clear that the naturalness of synthesized singing-voice increases as F0 fluctuations are added one by one, and that the effect of overshoot is the largest in all F0 fluctuations. Moreover, the quality of SYN-ALL is close to that of real singing-voice NORMAL.

This result is almost the same as the one in Fig. 8, and it is also consistent with the result in Fig. 4. However, the naturalness of SYN-ALL is slightly less than that of NORMAL, wheras the naturalnesses of SYN-ALL and NORMAL are almost the same as in Fig. 8. This is because there were deteriorations in synthesized singing-voice quality caused by a slight divergence between the generated F0 contour and the real singing-voice spectrum sequence. Although the quality of SYN-ALL also suffers from the same phenomenon, the difference in naturalness between SYN-ALL and NORMAL is quite small. Therefore, the proposed F0 control model seems to be adequate for natural singing-voice synthesis. Moreover, it is expected that a higher-quality singing-voice synthesis system can be constructed by considering a spectrum control corresponding to F0 contours. These results demonstrate not only that F0 fluctuations are important for singing-voice perception, but also that the proposed F0 control model can generate F0 contours with sufficiently accurate F0 fluctuations for synthesizing natural singing-voices.

6 Conclusion

In this paper, we have reconsidered the significance of F0 fluctuations in singing-voice perception as characteristics for natural singing synthesis. We have also considered how to control F0 fluctuations to generate the F0 contour of a singing-voice and how to synthesize a natural singing-voice.

Our analysis focused on four types of F0 fluctuation – overshoot, vibrato, fine-fluctuations, and preparation – as the dynamic characteristics of singing-voices, and we investigated how strongly these F0 fluctuations affect singing-voice perception through psychoacoustical experiments. In the experiments, these characteristics were observed in both Western classical and Japanese traditional singing styles of professional and amateur singers, and we found that the naturalness of a singing-voice decreased when we removed any of these F0 fluctuations from the F0 contour. Therefore, these F0 fluctuations are important factors affecting singing-voice perception. Moreover, the effect

of overshoot is especially great, and the effect of preparation is slightly larger than that of vibrato.

This paper then developed an F0 control model that can control F0 fluctuations affecting singing-voice perception. This model can generate an F0 contour by adding individual F0 fluctuations to the melody component. Since this model can control the characteristics of each F0 fluctuation by determining optimal control parameter values, the generated F0 contours contain F0 fluctuations whose characteristics are similar to a real one.

This paper finally applied our F0 control model to a singing-voice synthesis system using the Klatt formant synthesizer and STRAIGHT to evaluate the model's effectiveness. The evaluation results showed that the naturalness of a synthesized singing-voice was increased by adding each type of F0 fluctuation, and that the quality of a synthesized singing-voice including all F0 fluctuations was close to that of a real singing -voice. Again, overshoot strongly affected the quality of the synthesized singing-voice, and the ranking of each F0 fluctuation effect on singing-voice perception was almost the same as when removing each F0 fluctuation from the real singing-voice F0 contour. From these results, we concluded that F0 fluctuations are important factors for singing-voice perception and synthesis, and that the effect of each F0 fluctuation is larger in the order of overshoot, preparation, and vibrato. Moreover, it is conclude that the proposed F0 control model can generate F0 contours including adequate F0 fluctuations, and can be applied into natural singing-voice synthesis.

Besides F0 control, spectral sequence control is important for singing-voice synthesis, and we confirmed the possibility that a singing-voice synthesis method using our F0 control model could synthesize high-quality singing-voice by considering spectrum control. Therefore, spectral characteristics affecting singing-voice perception and spectral sequence control should be studied in future work.

Acknowledgements

We would like to thank Ken-Ichi Sakakibara for many useful comments and advices. This work was supported by a grant-in-aid for scientific research from the JSPS (No. 13610079).

References

Akagi, M., Iwaki, M. and Minakawa, T., 1998. Fundamental frequency fluctuation in continuous vowel utterance and its perception. In: Proc. ICSLP98, Sydney, vol. 4, pp. 1519-1522.

Akagi, M. and Kitakaze, H., 2000. Perception of synthesized singing-voices with fine-fluctuations in their fundamental frequency fluctuations. In: Proc. ICSLP2000, vol. 3, pp. 458-461

de Cheveigne, A. and Kawahara, H., 2001. Comparative evaluation of F0 estimation algorithms. In: Proc. Eurospeech2001, pp. 2451-2454.

de Krom, G. and Bloothooft, G., 1995. Timing and accuracy of fundamental frequency changes in singing. In: Proc. ICPhS'95, vol. 1, pp. 206-209.

Fujisaki, H. and Tatsumi, M., 1981 Analysis control in singing. Vocal fold physiology, UNIVERSITY OF TOKYO PRESS, pp. 347-363.

Fujisaki, H., Ohno, S., and Narusawa, S., 2000. Physiological mechanisms and biomechanical modeling of fundamental frequency control for the common Japanese and the standard Chinese. In: Proc. 5th Seminar on Speech Production 2000, pp. 145-148.

Hakes, J., Shipp, T., and Doherty, T., 1987. Acoustic characteristics of vocal oscillations: vibrato, exaggerated vibrato, trill, and trillo. J. Voice, vol. 1, pp. 326-321.

Horii, Y., 1989. Acoustic analysis of vocal vibrato: a theoretical interpretation of data. In J. Voice, vol. 3, pp. 151-159.

Ishizaka, K. and Flanagan, J. L., 1972. Synthesis of voiced sounds from two-mass model of the vocal cords. In Bell Syst. Tech. J., 51, 6, pp. 1233-1268.

Kawahara, H. Katayose, A. Patterson, R.D. and de Cheveigne, A., 1999. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In: Proc. Eurospeech'99, pp. 2781-2784.

Kawahara, H., Masuda- Katsuse, I., and de Cheveigne, A., 1999. Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous- frequency based on F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication, Vol. 27, pp. 187-207.

Klatt, D., 1980. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, volume 67, number 3, pp. 971-995.

Mori, H., Odagiri, W., and Kasuya, H., 2004. Transition Characteristics of Fundamental Frequency in Singing. In: Proc.ICA2004, Mo 5. C1.3, pp. I499-500.

Moriyama, T., Ogawa, H., and Tenpaku, S., 1996. A new control model based on rising and falling fundamental frequency. In: Proc. of ASA and ASJ Third Joint Meeting, pp. 1171-1176.

Myers, D. and Michel, J., 1987. Vibrato and pitch transitions. J. Voice, vol.1, pp. 157-161.

Nakayama, I. and Kobayashi, N. 1996. Singing voice : Charm and troubles on voice quality. The Journal of the Acoustic Society of Japan, vol. 52, no.5, pp. 383-388.

Nakayama, I., 2004. Comparative Studies on Vocal Expression in Japanese Traditional and Western Classical-style Singing, Using a Common Verse. In: Proc. ICA2004, Mo4. C1.1, pp. I295-296.

Press, W. H., Flannery, B. P., Teukolsky, S., and Vetterling, W, T., 1988. Numerical Recipes in C, Cambridge University Press, Cambridge.

Seashore, C., 1938. Studies in the Psychology of Music vol. 1: The vibrato. University of Iowa City.

Sundberg, J., 1979. Maximum speed of pitch changes in singers and untrained subjects. J. Phonetics, vol. 7, no.2. pp. 71-79.

Sundberg, J., 1987. The science of the singing-voices: A rhapsody on perception, pp.163-165, Northern Illinois University Press.

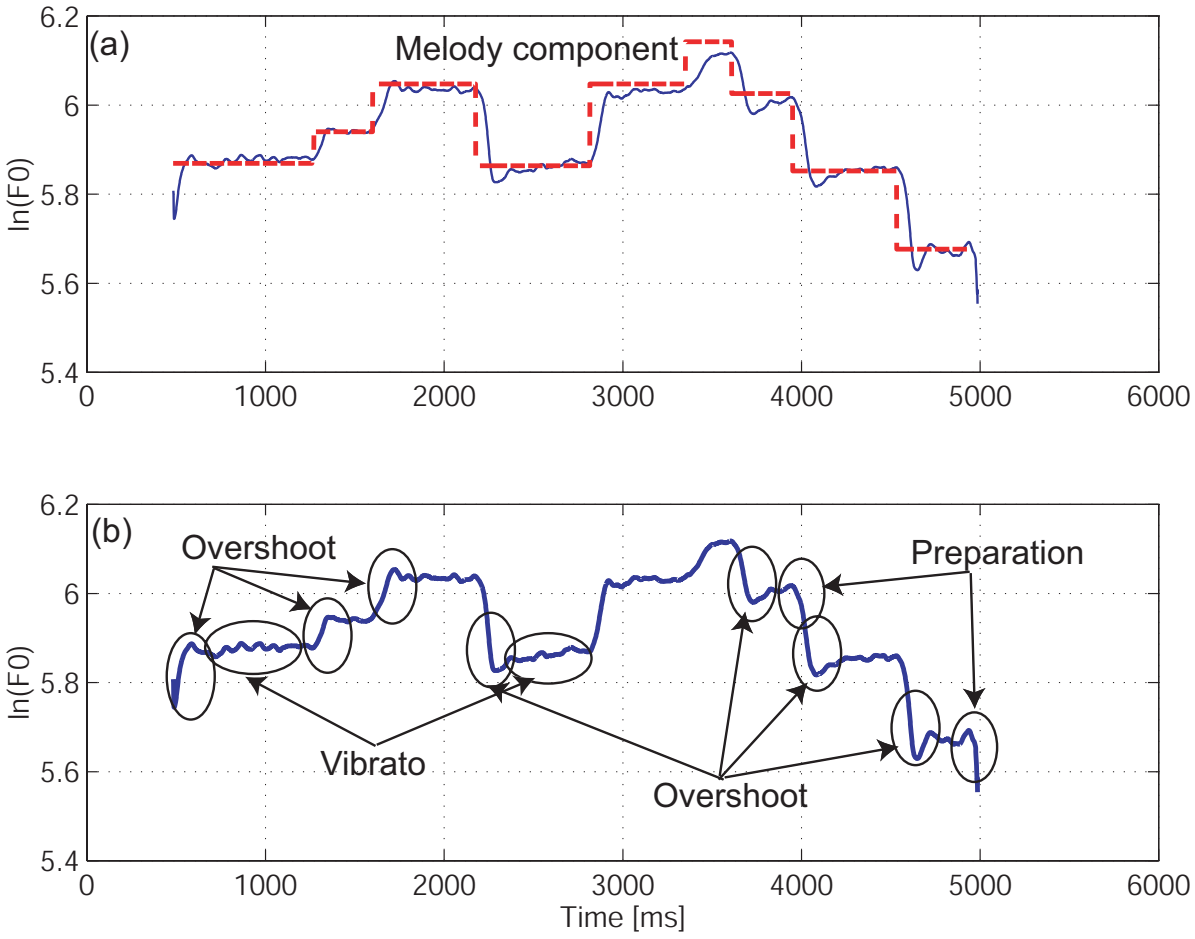

Fig. 1. F0 contour and F0 fluctuations of singing-voice. F0 was extracted using TEMPO. (a) Melody component, (b) F0 fluctuations: overshoot, vibrato, and preparation. Fine fluctuation along entire contour.

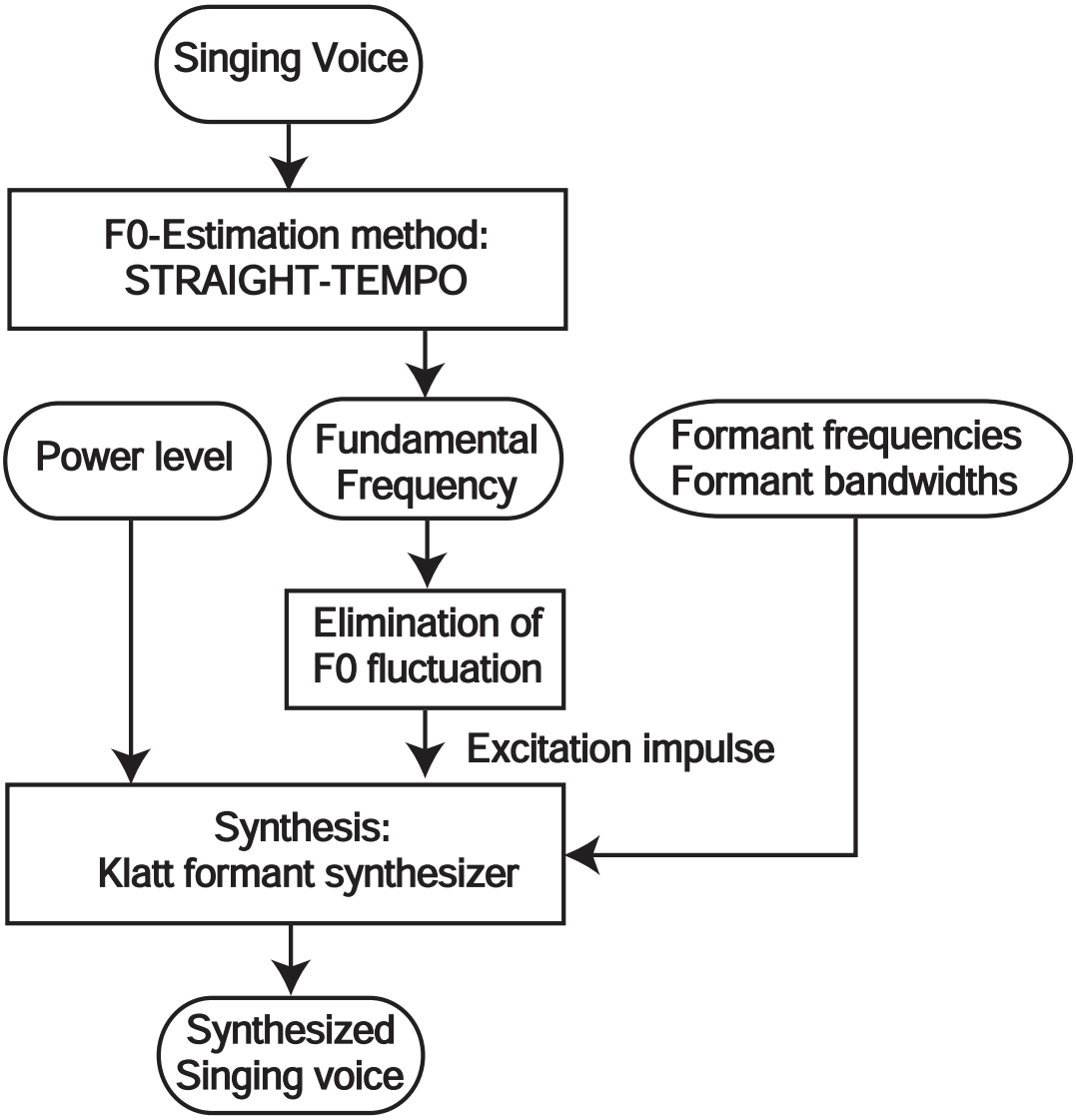

Fig. 2. Singing-voice synthesis using the Klatt formant synthesizer for synthesizing a singing-voice where F0 fluctuations are eliminated from the real F0 contour of the singing-voice.

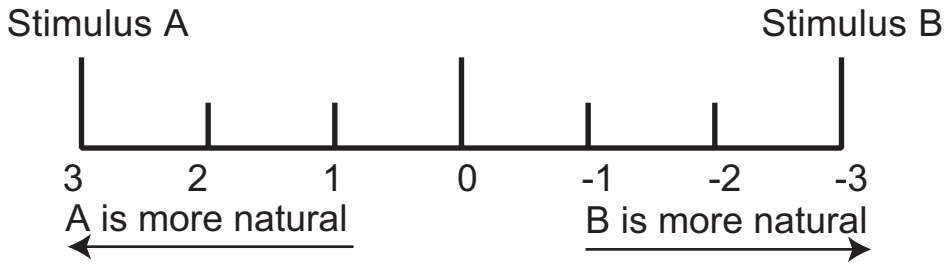

Fig. 3. Evaluation measure of Scheffe's paired comparison (seven grades: 3, 2, 1, 0, -1, -2, and -3).

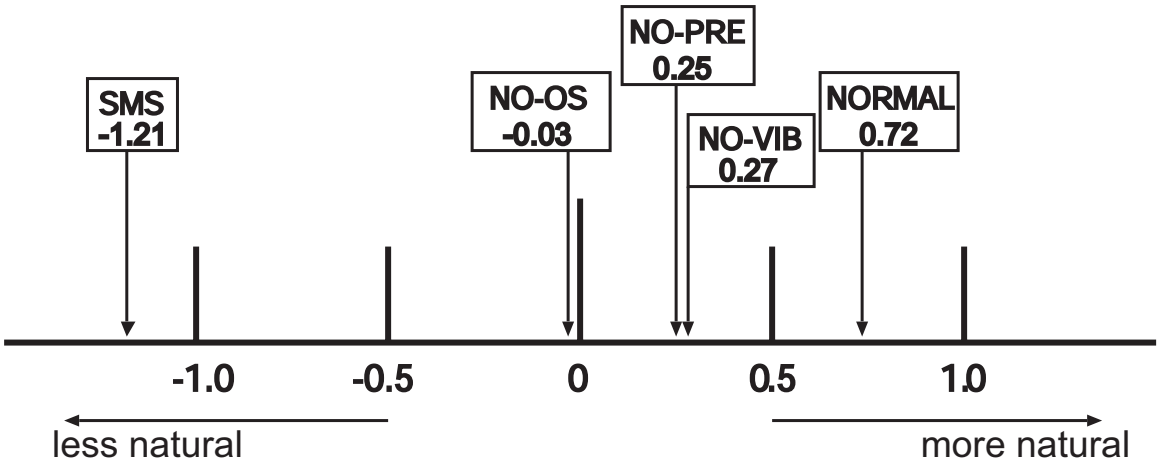

Fig. 4. Experimental results: Importance of overshoot, vibrato, and preparation. NORMAL: Synthesized singing-voices using real F0 contours. NO-VIB, NO-PRE, and NO-OS: Synthesized singing-voices using an F0 contour where vibrato, preparation, and overshoot have been respectively eliminated. SMS: Synthesized singing-voice using an F0 contour smoothed by lowpass filtering.

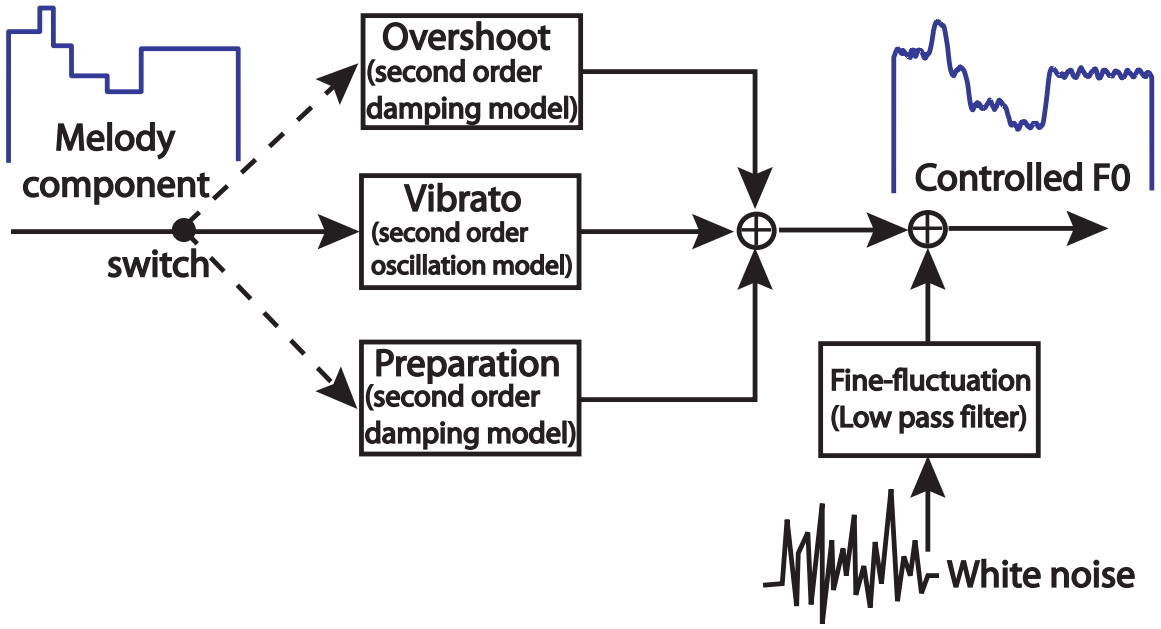

Fig. 5. Schematic of F0 control model. This model can generate the F0 contour of a singing-voice by adding each type of F0 fluctuation into the melody component that is the input component of this model.

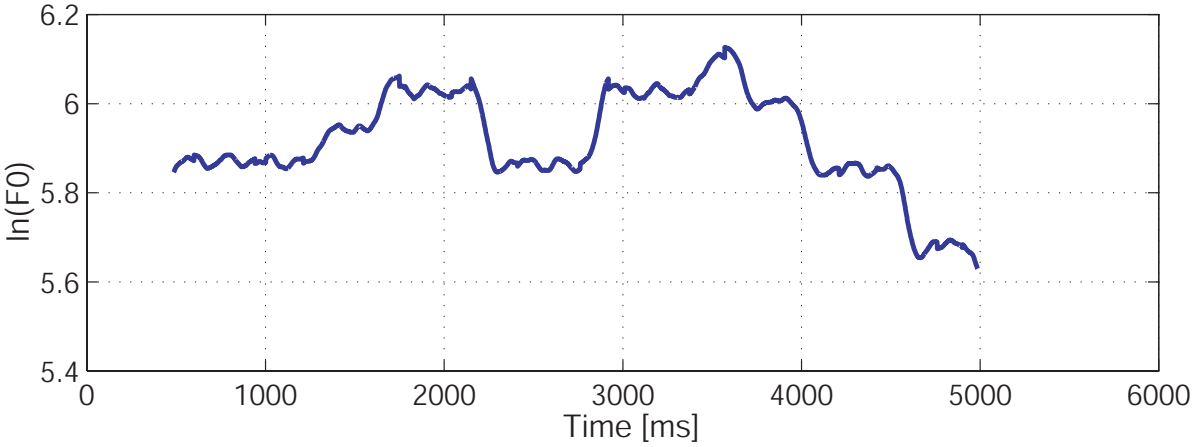

Fig. 6. Generated F0 contour (same portion as shown in Fig. 1(b)).

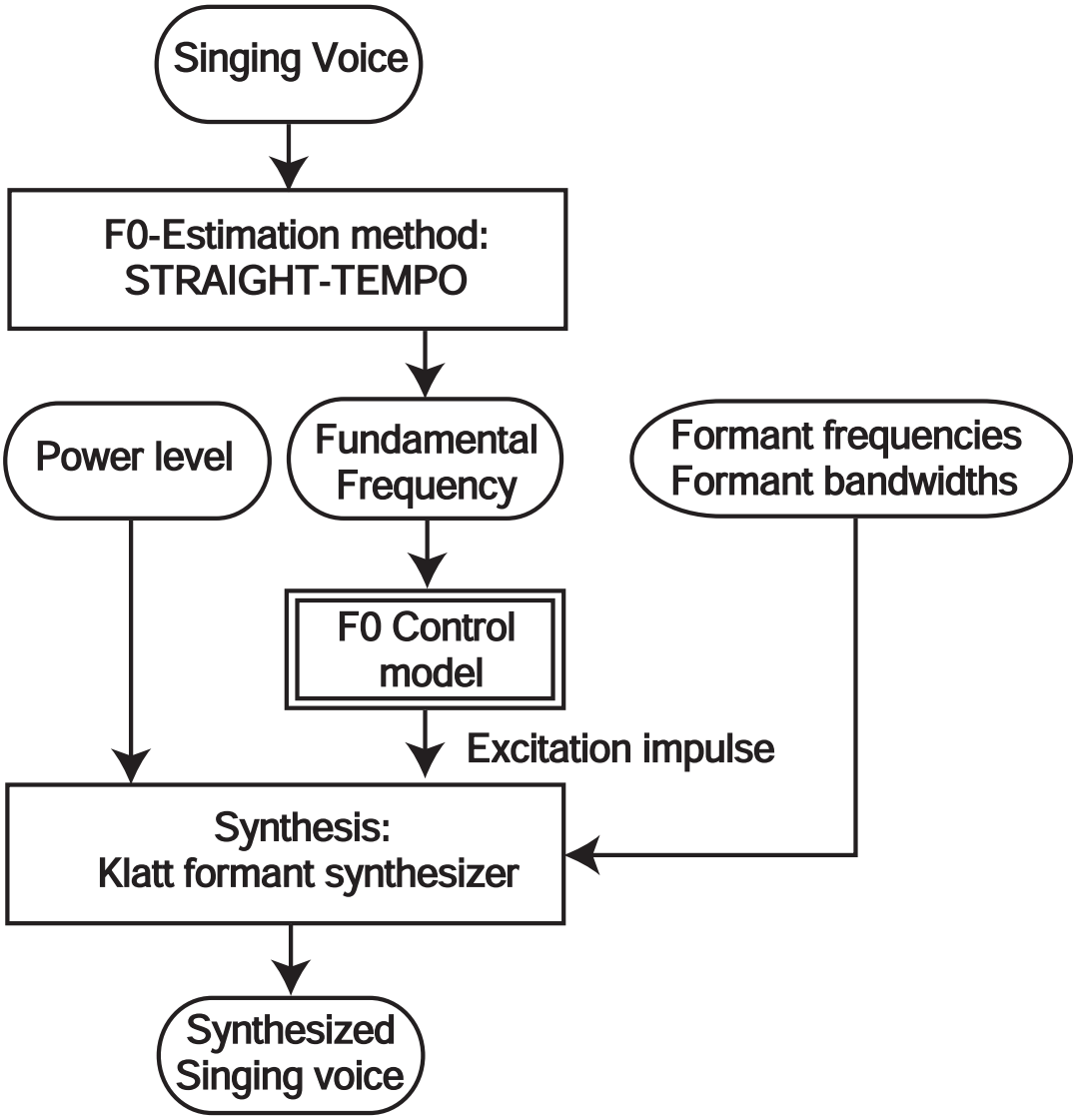

Fig. 7. Singing-voice synthesis system using Klatt formant synhtesizer and the proposed F0 control model to evaluate the F0 control model.

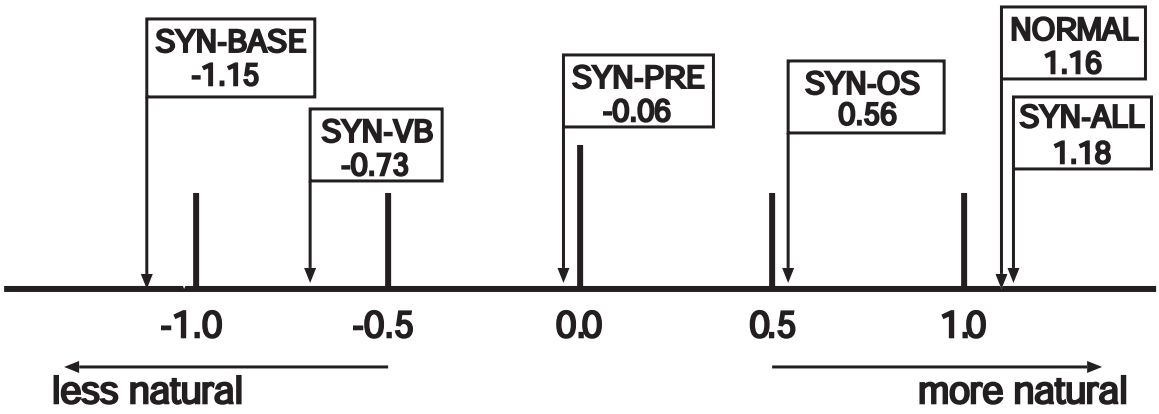

Fig. 8. Experimental results of synthesized singing-voice evaluation (synthesis method: Klatt formant synthesizer). NORMAL: Synthesized singing-voices using real F0 contours. SYN-ALL: Synthesized singing-voice using a generated F0 contour including all F0 fluctuations. NO-VIB, NO-PRE, and NO-OS: Synthesized singing-voices using generated F0 contours including vibrato, preparation, and overshoot, respectively. SYN-BASE: Synthesized singing-voice using the melody component only.

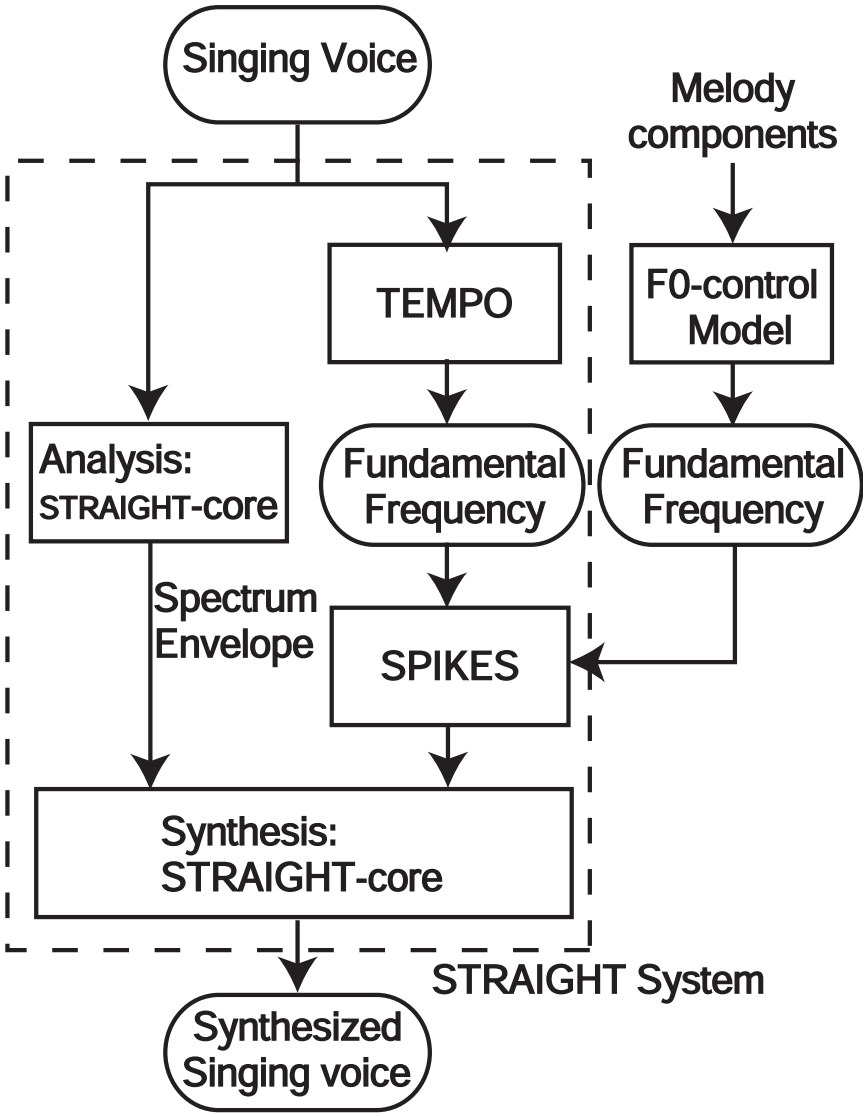

Fig. 9. Singing-voice synthesis system using STRAIGHT and the proposed F0 control model to evaluate the F0 control model.

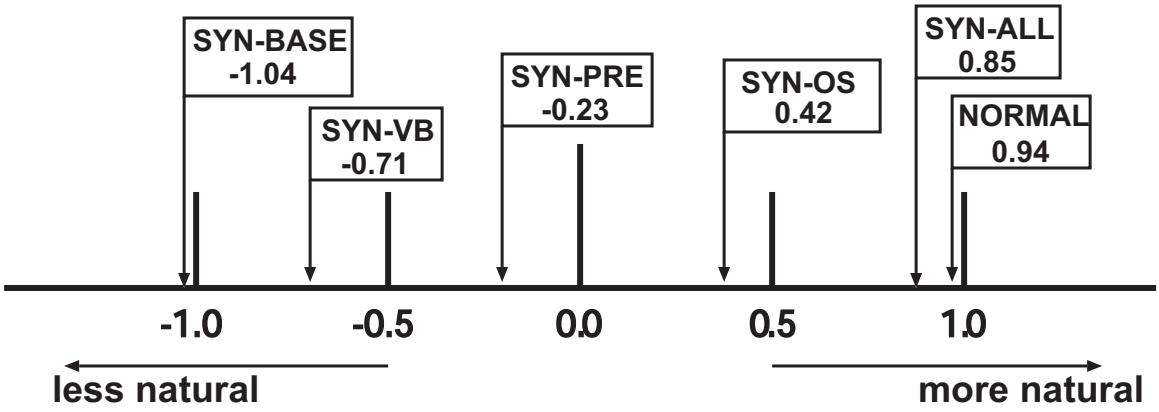

Fig. 10. Experimental results of synthesized singing-voice evaluation (synthesis method: STRAIGHT). NORMAL: Synthesized singing-voices using real F0 contours. SYN-ALL: Synthesized singing-voice using a generated F0 contour including all F0 fluctuations. NO-VIB, NO-PRE, and NO-OS: Synthesized singing-voices using generated F0 contours including vibrato, preparation, and overshoot, respectively. SYN-BASE: Synthesized singing-voice using the melody component only.

Table 1. Optimized parameter values in the F0 control model.

F0 fluctuation	Ω [rad/ms]	ζ	K
overshoot	0.0348	0.5422	0.0348
vibrato	0.0345	—	0.0018
preparation	0.0292	0.6681	0.0292