EXTRACTION OF F0 DYNAMIC CHARACTERISTICS AND DEVELOPMENT OF F0 CONTROL MODEL IN SINGING VOICE

Takeshi Saitou, Masashi Unoki, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi-machi, Nomi-gun, Ishikawa 923-1292 Japan
{t-saitou, unoki, akagi}@jaist.ac.jp

ABSTRACT

Fundamental frequency (F0) control models, which can cope with F0 dynamic characteristics related to singing-voice perception, are required to construct natural singing-voice synthesis systems. This paper discusses the importance of F0 dynamic characteristics in singing voices and demonstrates how much it influence on singing voice perception through psychoacoustic experiments. This paper, then, proposes an F0 control model that can generate F0 fluctuations in singing voices, and a singing-voice synthesis method. The results show that F0 contour including fluctuations: Overshoot, Vibrato, Preparation, and Fine-fluctuation, affects singing voice perception, and the proposed synthesis method can generate natural singing voices by controlling these F0 fluctuations.

1. INTRODUCTION

Singing voice has more dynamic and complicated characteristics than speaking voice for speech perception. Especially naturalness of voices are contained in the fundamental frequency (F0) contour. There are, for example, many F0 fluctuations in the contour such as Overshoot, Vibrato, etc. However, a quantitative assessment of the perceptual influence of fluctuations in F0 contours has not been investigated deeply, although it is known that the fluctuations in F0 contours may be important factors in producing high-quality synthesized speech.

On the one hand, various speech synthesis method have been proposed for speaking-voice. Many of these methods are based on the source-filter model so that F0 and formant information can be separately used in the model. However the proposed F0-control models cannot generate F0 contours of singing voices, although they can deal with speaking-voice synthesis. Because dynamic range of the F0 contours in singing voices is wider than that in speaking voices and F0 fluctuations in singing voices are larger and more rapid than those in speaking voices, as mentioned in earlier. So, in order to construct a singing-voice synthesis method, we have to clarify F0 fluctuation characteristics and then develop a method for controlling these fluctuations in F0 contours.

This paper discusses the importance of using dynamic characteristics in F0 contours in singing voices through psychoacoustic experiments, and then proposes a F0 control model that can cope with dynamic characteristics in F0 contour in singing voice.

2. ANALYSIS OF F0 DYNAMIC CHARACTERISTICS

F0s of speech contain slow and large fluctuations related to prosodic information, and rapid and fine fluctuations related to the naturalness of speech. In singing voices, it was known that there are three characteristics in F0 contours of singing voices [1].

(a) A particular value of F0 corresponds to a particular Note.

(b) F0 fluctuations in any one note is stable.

(c) There are many F0 fluctuations which are only observed in singing voice.

(a) and (b) are static characteristics related to melody. (c) is related to dynamic characteristics such as Overshoot, Vibrato, etc. In this section, we deeply investigate F0 dynamic characteristics of (c).

2.1. Singing voice data

The singing-voice data used for our experiments were obtained from recordings of three adults singing a Japanese children's song "Nanatsunoko". The singers were asked to sing it with Japanese vowel /a/ only, to simplify the experimental conditions. The songs were recorded on a DAT with 48-kHz sampling and 16-bit accuracy, and then were down-sampled to 20 kHz.

2.2. F0 estimation method

The F0s were estimated using the F0 extraction method, TEMPO in STRAIGHT [2, 3]. We confirmed beforehand that TEMPO could accurately extract Fine fluctuations in F0 contours. It can extract modulation frequencies with a precision of up to about one-fifth of the F0.

2.3. Analysis of F0 contour

Figure 1 shows an estimated F0 contour along the logarithmic axis. Fig. 1 (a) shows **Melody component** that represents the note change of the extracted F0. We extracted four F0 dynamic characteristics as follows.

Overshoot: Deflection exceeding the target note after note changes.

Vibrato: Periodic frequency modulation (4 - 7 Hz) .

Preparation: Deflection of the opposite direction of note change observed just before note changes.

Fine-fluctuation: Irregularly fine fluctuation higher than 10 Hz.

It is reported that Overshoot, Vibrato and Fine-fluctuation in F0 contour are the dynamic characteristics, which are peculiar to singing voices [1, 4, 6]. In this paper, we also extracted Preparation as a new dynamic characteristic.

The analyzed F0s for Fine-fluctuations show that the modulation frequency (MFs) contained frequency component up to 20 Hz and that the modulation amplitudes (MAs) were 20 cent on average and 100 cent at maximum, which are one-fifth of and the same as the half-tone musical scale, respectively. The extracted MF and MA were $[\mathrm{MF(Hz)}, \mathrm{MA/F0(\%)}] = (20\mathrm{Hz}, 1.2\%)$, where $F0$

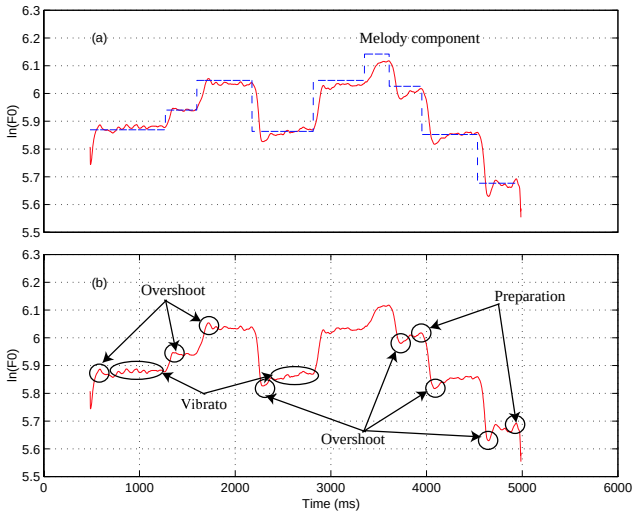

Figure 1: Extracted fundamental frequency (F0) using TEMPO in STRAIGHT; portion of "Nanatsunoko," /kawaii-nanatsuno/ and dynamic characteristics: (a) Melody component related to the estimated F0; (b) Overshoot, Vibrato, and Preparations. Fine-Fluctuation is in whole contour.

was 125 Hz. About the detection thresholds, the previous works reported that the Fine fluctuations in the F0 contour of singing voices affect the perception of quality and that the magnitude of this effect depends on the MF and MA [5, 6]. These reports suggested that (1) the Fine-fluctuations having a lower MF or a larger MA were more detectable, and (2) the Fine-fluctuations were more detectable when the F0 rose. From these findings, we concluded that humans might be able to perceive the MF and MA of Fine-fluctuation components in singing voices. In this paper, we use these to control the Fine-fluctuation in the generated F0 contour.

3. IMPORTANCE OF F0 FLUCTUATIONS

We extracted four types of F0 fluctuations from the observed F0 contours. These fluctuations may affect perception of singing voices. We carried out psychoacoustic experiments to demonstrate how much the F0 dynamic characteristics influence on singing voice perception.

3.1. Stimuli and Synthesis

We eliminated each F0 dynamic characteristic from F0 contours and re-synthesized the singing voices using the modified F0s.

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

NO-OS: Singing-voice set removed Overshoot component.

NO-VIB: Singing-voice set removed Vibrato component.

NO-PRE: Singing-voice set removed Preparation component.

SMS: Singing-voice set whose F0 was smoothed by an FIR low-pass filter (cut-off frequency was 5Hz).

All stimuli were paired and recorded randomly. The number of paired stimuli was 20.

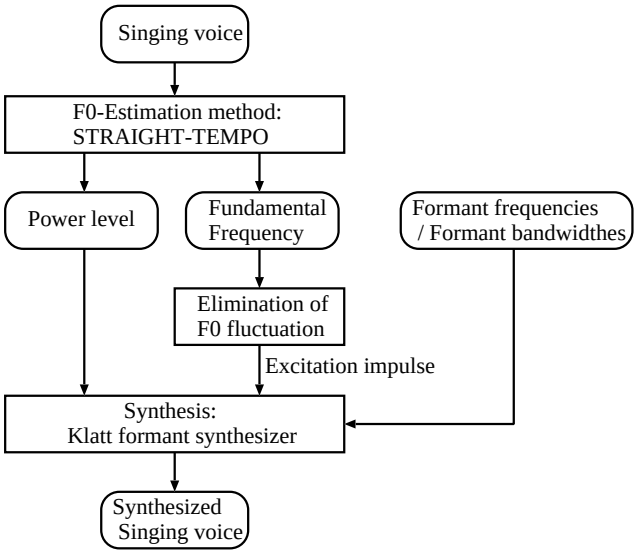

Figure 2: Singing-voice synthesis using the Klatt formant synthesizer.

The stimuli were synthesized voices of the vowel /a/ using the Klatt formant synthesizer to reflect F0 fluctuations, as shown in Fig. 2. 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

$$t_{n+1} = t_n + 1/f_m(t_n). \quad (1)$$

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. The formants frequencies of the Japanese vowel /a/ were set to be 800, 1200, 2500, 3500, 4500, and 5500 Hz, and each bandwidth was set to be 10 % of the corresponding formant frequency.

3.2. Procedure

The paired stimuli were presented through binaural earphones at a comfortable loudness level. Each paired stimulus was randomly presented to each subject once. The subjects were six graduate students having normal hearing ability. Scheffe's method of paired comparison was used to evaluate the naturalness of singing voices (Seven-grade evaluation measure: $-3, -2, -1, 0, 1, 2, 3$). Then, naturalness of the synthesized singing voice at each condition is calculated as parameter of population.

3.3. Result and Discussion

Figure 3 shows the experimental results. The numerals below the horizontal axis indicate the degree of naturalness of a singing voice. The results indicate that the effects of three F0 dynamic characteristics, Overshoot, Vibrato, and Preparation, on singing-voice perception are large, and the effect of Overshoot is the largest. It is confirmed that Preparation as a new fluctuation is one the of important fluctuations for singing voice perception. In addition, these indicate that lack of these F0 fluctuations, labeled SMS, affects singing voice perception strongly. Hence, we have to consider these F0 fluctuations in the contour to construct an F0 control model.

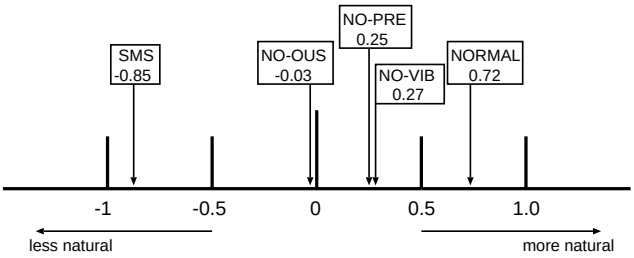

Figure 3: Experimental result: Importance of Overshoot, Vibrato, and Preparation.

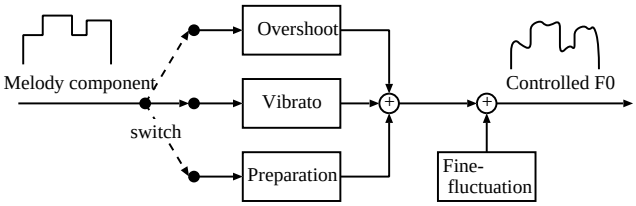

Figure 4: Schematic graph of F0 control model.

4. THE F0 CONTROL MODEL FOR SINGING VOICE

For F0 control models [7, 8] for speaking voices, it is difficult to control and generate F0 contours including dynamic characteristics of singing voices. In this paper, we develop a new method which can control F0 dynamic characteristics and generate F0 contours of singing voices.

4.1. Schematic graph of F0 control model

Figure 4 shows a schematic graph of the proposed F0 control model. Input of the model is Melody component in F0 contour. This is described by sum of step functions. This model generates controlled F0 contours adding four fluctuations; Overshoot, Vibrato, Preparation, and Fine-fluctuation as follows, into Melody component.

Overshoot: Second-order damping model.

Vibrato: Second-order oscillation model (no-loss).

Preparation: Second-order damping model.

Fine-fluctuation: Irregularity rapid oscillation with higher than MF of 10-Hz and MA of 5-Hz.

The transfer function of first three fluctuations are described as

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

where Ω and ζ are system parameters. Here, the impulse response of $H(s)$ can be obtained as

$$h(t) = \begin{cases} \frac{\Omega}{2\sqrt{\zeta^2-1}}\left(\exp(\lambda_1\Omega t) - \exp(\lambda_2\Omega t)\right), & |\zeta| > 1 \\ \frac{\Omega}{\sqrt{1-\zeta^2}}\exp(-\zeta\Omega t)\sin(\sqrt{1-\zeta^2}\Omega t), & |\zeta| < 1 \\ \Omega t\exp(-\Omega t), & |\zeta| = 1 \\ \sin(\omega t). & |\zeta| = 0 \end{cases} \tag{3}$$

Table 1: Optimized parameter sets in the F0 control model.

Fluctuation	ω [rad/ms]	ζ
Overshoot	0.031	0.52
Vibrato	0.033	—
Preparation	0.028	0.72

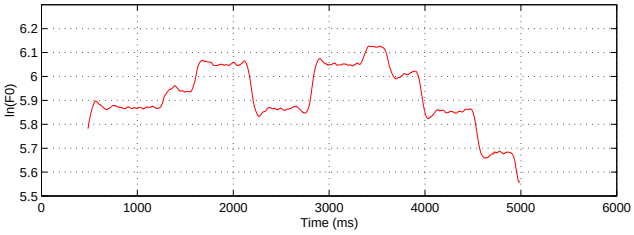

Figure 5: Generated F0 contour (same portion as that shown in Fig. 1 (b)).

First equation means a solution to second-order exponential damping model, second means a solution to second-order damping model, third means a solution to second-oder critical oscillation model, and fourth means a solution to second-order oscillation model (no-loss). So, Overshoot and Preparation used for second equation, and Vibrato used for fourth equation. A nonlinear least-squared-error method was applied to minimize the error between the extracted and the controlled F0s within parts corresponding to conditions, in order to determine the optimized parameters of Ω and ζ at each condition. The parameters optimized by minimizing the error for all stimuli are shown in Table 1.

Controlling of Fine-fluctuation consists of a lowpass filtering of white noise with the cutoff frequency of 10 Hz and a normalizing of amplitude of 5 Hz. These values were set based on the considerations in Sec. 2.3.

This F0 control model can generate the controlled F0 contours which includes each dynamic characteristic, by determining the optimal control parameters for ζ and Ω. F0 contours generated by the model from Melody components obtained from F0 contour in Fig. 1 is shown in Fig. 5.

4.2. Singing voice synthesis

We synthesized singing-voices using a synthesis method as shown in Fig. 6. This method consists of two blocks: the F0 control model and STRAIGHT [3] instead of the Klatt synthesizer as shown in Fig. 2. The aim of this improvement is to extend the proposed method for natural singing-voices synthesis that can deal with singing-speech, including lyrics. Because it is difficult to do it via the Klatt synthesizer.

STRAIGHT consists of TEMPO, STRAIGHT-core, and SPIKES. TEMPO [2] is the F0 estimation block and SPIKES is the excitation-pulses generator for source information using the F0 contour. STRAIGHT-core is the spectrum envelope estimation using F0-adaptive time-frequency smoothing to eliminate periodicity interferences, and is the synthesizer using the spectrum envelope and excitation pulses. In this method, the F0 control model is incorporated with STRAIGHT instead of TEMPO block. For synthesis process, the spectrum envelope is not manipulated.

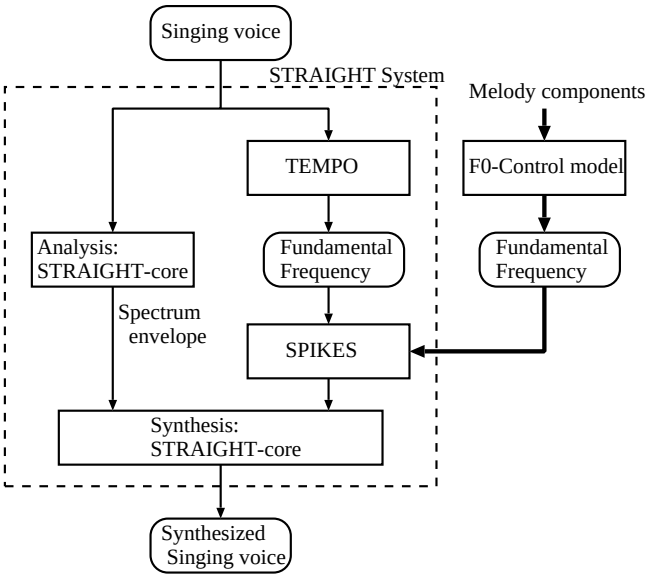

Figure 6: Singing-voice synthesis method using STRAIGHT and F0-control model.

4.3. Demonstrations

In order to investigate how much these fluctuations influence sining-voice quality, we added each fluctuation into the F0 contour of the melody component, synthesized singing voices using those F0s, and presented them to subjects to judge their naturalness.

Six stimuli were used in the experiment as follows.

NORMAL: using the extracted F0 (no control)

SYN-All: using all the fluctuations

SYN-OS: using the Overshoot

SYN-PRE: using the Preparation

SYN-VB: using the Vibrato and Fine-fluctuation

SYN-BASE: adding no fluctuations to the Melody component.

The spectrum of synthesized singing voices was identical in all stimuli. The psychoacoustic experiment was carried out on the same procedure and conditions as utilized in the experiment in Sec. 3.

4.4. Results and Discussion

The results in Figure 7 show that the F0 control model can generate F0 contours including F0 fluctuations related to singing voices perception, and the proposed synthesis method can produce singing voices that sound as natural as NORMAL voices, when all the fluctuations are added to the Melody component. These results indicate that the F0 fluctuations are important to the naturalness of singing voices.

5. CONCLUSION

In this paper, we have discussed some F0 dynamic characteristics in singing voices and demonstrated how much the F0 dynamic characteristics influence on singing voice perception, through psychoacoustic experiments. The results show that F0 fluctuations,

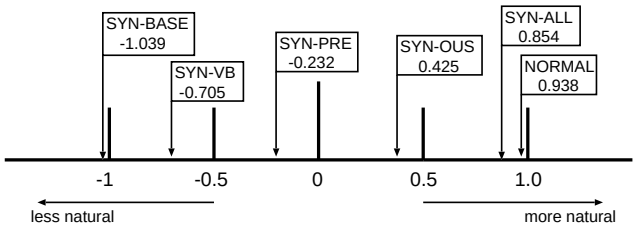

Figure 7: Result of synthesized singing-voice evaluation.

especially Overshoot, Vibrato, Fine-fluctuation, and Preparation affect singing voice perception. Then, we have also proposed an F0 control model that can control F0 fluctuation characteristics and generate the controlled F0 contours for singing-voices. The F0 control model can be applied to synthesizing natural singing voices.

6. ACKNOWLEDGMENT

This work was supported by CREST of JST and by a grant-in-aid for scientific research from the Ministry of Education (No. 13610079).

7. REFERENCES

[1] Yatabe, M. and Kasuya, H., "Dynamic characteristics of fundamental frequency in singing," Proc. Autumn Meeting of the acoustical society of Japan, 3-8-6, 1998.

[2] Kawahara, H. *et al*., "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," Proc. Eurospeech99, pp.2781-2784, Sept. 1999.

[3] Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A., "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, 1999.

[4] Odagiri, W. and Kasuya, H., "Study of analysis, synthesis and perception of vocal vibrato," Proc. Autumn Meeting of the acoustical society of Japan, 1-7-5, 1999.

[5] Akagi, M., Iwaki, M. and Minakawa, T., "Fundamental frequency fluctuation in continuous vowel utterance and its perception", ICSLP98, Sydney, Vol.4, pp. 1519-1522, Dec. 1998.

[6] Akagi, M. and Kitakaze, H., "Perception of synthesized singing voices with fine fluctuations in their fundamental frequency fluctuations," Proc. ICSLP2000, Beijing, vol. III, pp. 458-461, Oct. 2000.

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

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