Development of the F0 Control Model for Singing-Voices Synthesis

Takeshi Saitou, Masashi Unoki, and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
{t-saitou; unoki; akagi}@jaist.ac.jp

Abstract

Fundamental frequency (F0) control models for singing voices are required to construct singing-voice synthesis systems that can generate natural singing-voices. This paper describes the development of an F0 control model for singing-voices synthesis. F0 fluctuations are revealed as characteristics that need to control the F0 contour of singing-voices by investigating how much they influence singing-voices perception through psycho-acoustical experiments. These fluctuations have wider dynamic range and more complicated changes rather than in speaking-voices. The F0 control model is developed so that it can control important F0 fluctuations for the purpose of singing-voice perception. The singing-voice synthesis method using the F0 control model is proposed to synthesize natural singing-voices. Results of these experiments show that the F0 fluctuations are significant factors for singing-voices perception; the F0 control model can generate F0 contours of singing-voices and can be applied to synthesize natural singing-voices.

1. Introduction

Singing and speaking are important ways in human communications to express linguistic and emotional information. Propositions of speech (speaking voices and singing voices) synthesis methods are important issues in speech signal processing, and therefore various speech synthesis methods have been proposed. However, most of these methods were proposed not for singing-voice synthesis but for speaking-voice synthesis.

Singing-voices have more dynamic and complicated characteristics than those of speaking voices, and these characteristics are significant factors in the naturalness of singing voices. In particular, it is well known that there are the following three characteristics in the F0 contours of singing voices [1].

(a) The dynamic range of the F0 contours is wider than that of speaking voices.

(b) A steady state of an F0 contour corresponds to a Note. The note changes of the F0 contours correspond to Melody.

(c) There are many F0 fluctuations that are observed only in singing voices.

These characteristics are peculiar to singing voices. (a) and (b) are static characteristics related to melody. (c) is related to fluctuations such as overshoot, vibrato, and fine fluctuation. These are reported as important F0 fluctuations for singing-voice perceptions [1, 2, 3].

Most speech synthesis methods are based on the source-filter model so that F0s related to source information and formant information related to filter characteristics are separately used in the model. Therefore, F0 contours are important components of speech synthesis, and methods for controlling F0 contours are required to propose speech synthesis methods. The usual F0 control models can generate F0 contours of speaking voices. However, they cannot control F0 fluctuations and cannot generate F0 contours of singing voices, because the characteristics of singing voices such as (a) – (c) are different from those of speaking voices. Thus, in order to propose a singing-voice synthesis method, we have to develop an F0 control model that can control the characteristics of F0 contours in singing voices.

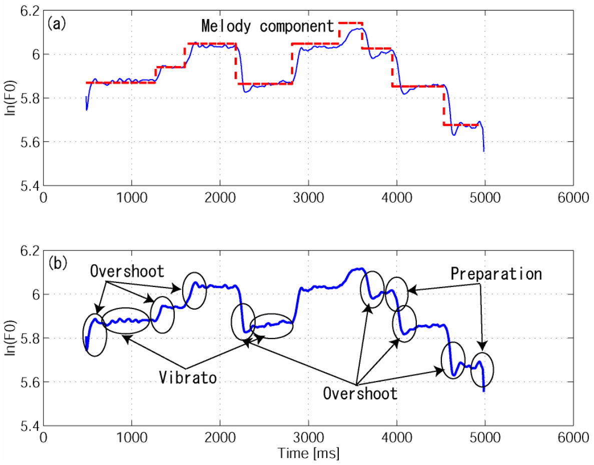

Figure 1: *Extracted F0 using TEMPO. (a) melody component, (b) F0 fluctuations: overshoot, vibrato, and preparation. Fine-fluctuation in whole contour.*

This paper reveals F0 fluctuations as significant characteristics of singing voices by investigating how much they influence singing-voice perception through psycho-acoustical experiments. Moreover, this paper presents an F0 control model that can control F0 fluctuations, and proposes a singing-voices synthesis method using the F0 control model.

2. Analysis of the F0 fluctuations

It is well known that the F0 fluctuations are significant factors in singing-voice perception. In this paper, we reconsider F0 fluctuations of F0 contours in singing voices, and then show their importance for singing-voice perception.

2.1. Singing-voice data

The singing-voice data used for this experiment 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. This aims to deal with only F0 fluctuations while spectrum information is fixed. The songs were recorded on a DAT with 48-kHz sampling and 16-bit accuracy, and then down-sampled to 20 kHz.

2.2. Analysis of F0 contour

The F0s were estimated using the F0 extraction method, TEMPO in STRAIGHT [4, 5]. Figure 1 shows an estimated F0 contour on a log-frequency axis. Figure 1 (a) shows a melody component that represents note change in the extracted F0. Figure 1 (b) shows four F0 fluctuations that are found in F0 contours. These fluctuations are as follows.

- **Overshoot:** deflection exceeding the target note after the note changes.
- **Vibrato:** periodic frequency modulation (4 - 7 Hz).
- **Fine-fluctuation:** irregularly fine fluctuation higher than 10 Hz.
- **Preparation:** deflection in the opposite direction of note change observed just before note changes.

The first three fluctuations have already been reported as important fluctuations, which are peculiar to singing voices [1, 2, 3]. This paper deals with preparation as a new fluctuation as well as the three previous fluctuations.

2.3. Importance of F0 fluctuations

We removed each F0 fluctuation from the F0 contours and re-synthesized the singing voices using the F0s. These synthesized singing voices were used as stimuli of psycho-acoustical experiments to investigate the importance of each F0 fluctuation for singing-voice perception.

- **NORMAL:** singing voices synthesized using the extracted F0 from a real song.
- **NO OS:** Singing voices whose overshoot was removed.
- **NO VIB:** Singing voices whose vibrato was removed.
- **NO PRE:** Singing voices whose preparation was removed.
- **SMS:** Singing voices whose F0 was smoothed by an FIR lowpass filter (cut-off frequency was 5Hz).

These were synthesized singing voices of a Japanese vowel /a/ using the Klatt formant synthesizer to represent F0 fluctuations, as shown in Fig. 2. The formant frequencies 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. All stimuli were randomly paired for the psychoacoustical experiment of paired comparison. The number of paired stimuli was 20 for each singing-voice data.

Scheffe's method of paired comparison was used to evaluate the naturalness of singing voices (The seven grades of evaluation are −3, −2, −1, 0, 1, 2, and 3). 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 voice at each condition was a calculated function based on the population.

2.4. Results and Discussion

Figure 3 shows the experimental results. The numerals under the horizontal axis indicate the degree of naturalness of a synthesized singing voice. The results indicate that the effects of three F0 fluctuations, Overshoot, Vibrato, and Preparation on singing voices perception are large, and the effect of Overshoot is the largest. This result suggests that Overshoot is the most important fluctuation for singing-voice perception. In this result, there is no condition for removing only fine-fluctuations, however, it is also an important fluctuation, compared with SMS. The score of NO-VIB may seem to be somewhat high, but no fluctuations can be removed from the F0 contours while the naturalness of Normal is maintained. Therefore, we have to deal with all four fluctuations in the F0 control model.

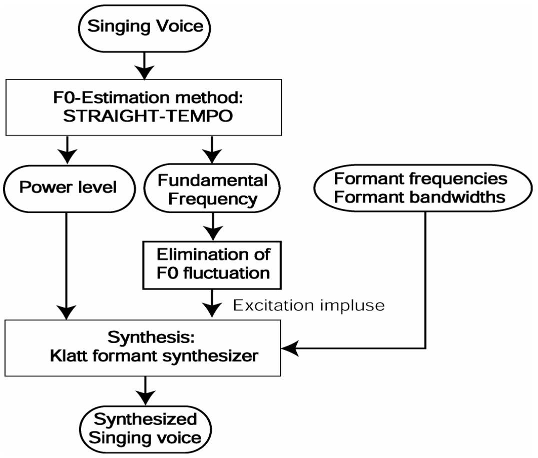

Figure 2: *Singing-voice synthesis using Klatt formant synthesizer.*

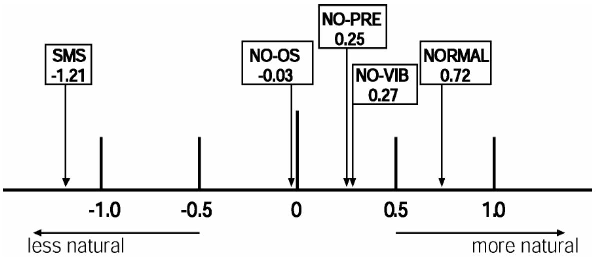

Figure 3: *Importance of Overshoot, Vibrato, and Preparation for singing-voice perception.*

3. F0 control model for singing-voices

We propose a new method that can control four F0 fluctuations and generate F0 contours of singing voices.

3.1. F0 control model

A schematic graph of the proposed F0 control model is shown in Figure 4. This model generates F0 contours adding four fluctuations: Overshoot, Vibrato, Preparation, and Fine-fluctuation into the Melody component.

The input for the model is Melody components described as a sum of a step function (Input component). Overshoot, Vibrato and Preparation are controlled using the transfer function of a second-order system represented as

$$H(s) = \frac{K}{s^2 + 2\zeta\Omega s + \Omega^2}, \qquad (1)$$

in which Ω is natural frequency, ζ is damping coefficient, and K is proportional gain. Here, the impulse response of $H(s)$ can be obtained as

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

in which $\lambda_1,\lambda_2=(-\zeta\pm\sqrt{\zeta^2-1})\Omega$, and Eqs. (2a), (2b), (2c) and (2d) are solutions to second-order exponential damping, second-order damping, second-order critical oscillation, and second-order oscillation (no-loss) models, respectively. In this paper, the control parameters are used for

- **Overshoot:** Second-order damping model [Eq. (2a)].
- **Vibrato:** Second-order oscillation model (no-loss) [Eq. (2d)].
- **Preparation:** Second-order damping model. [Eq. (2a)].

These F0 fluctuations were controlled by the determination of control parameters Ω, ζ, and K. Control of Fine-fluctuation is done by lowpass filtering and normalizing white noise.

3.2. Optimal control parameters

It is important to determine the optimal parameter for synthesizing natural singing voices. Therefore, we determine adequate control parameters for each F0 fluctuation by analyzing various singing-voices data.

Singing-voice datasets [7], obtained from various recordings of the popular song: "Kaedeiroduku-Yamanoasawa" with free melody, were used to determine the optimal parameters. The data selected from the datasets are 4 sopranos (18 data), 2 mezzo-sopranos (7 data), 1 alto (4 data), 3 tenors (13 data), 3 baritones (14 data), 2 pop-musicians (6 data), and 1 enka (Japanese ballad) singer (2 data). The control parameters for each F0 fluctuation were fitted to all F0 fluctuations (total 64 data) individually.

In order to determine the optimal control parameters of each F0 fluctuation, a nonlinear least-squared-error method was used to minimize the error between the extracted and the controlled F0s. In addition, when optimizing the Overshoot and Preparation control parameters, K was determined to be the same control parameter as Ω.

Vibrato is characterized by two components: the rate and period [8]. The rate specifies the period of F0 oscillation, and the amplitude specifies the rate of value of the F0's average. In the proposed F0 control model, Ω controls the period, and K controls the amplitude. Thus, we determine the optimal parameters by analyzing these two parameters. The analyzed results show that the period of oscillation is about 182 [ms] and the amplitude is about 5.2%.

Controlling Fine-fluctuation is done by the method mentioned in Sec. 3.1. The cutoff frequency of the lowpass filter is 10 Hz and the amplitude is normalized to 5 Hz [3].

From these results, we determined the optimal parameters for controlling F0 fluctuations as shown in Table.1

Table 1: *Optimized parameter sets*

Fluctuation	Ω [rad/ms]	ζ	K
Overshoot	0.0348	0.5422	0.0348
Preparation	0.0292	0.6681	0.0292
Vibrato	0.0345	0	0.0018

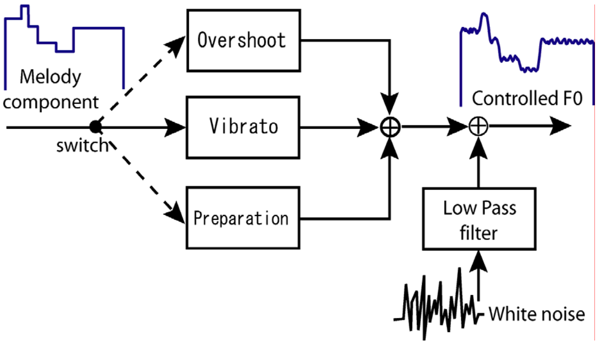

Figure 4: *Schematic of F0 control model.*

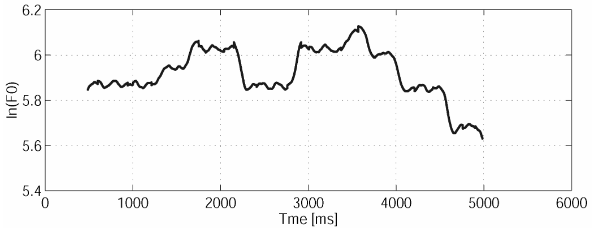

Figure 5: *Generation of F0 contour by F0 control mode. (the same portion as that shown in Figure 1.)*

4. Evaluation of F0 control model

In order to verify that the proposed F0 control model can control the F0 fluctuations of singing voices, we applied the proposed model to singing-voice synthesis and investigated the naturalness of these synthesized singing voices.

4.1. Singing-voices synthesis

We synthesized singing voices using a synthesis method as shown in Figure 6. This method consists of two blocks: the F0 control model and STRAIGHT [5] instead of the Klatt synthesizer as shown in Figure 2. The aim of this improvement is to extend the proposed method for singing-voices synthesis so that it can deal with singing voices, including lyrics because it is difficult for the Klatt formant synthesizer to control the spectrum. STRAIGHT consists of TEMPO, STRAIGHT-core, and SPIKES. TEMPO [4] is the F0 estimation block and SPIKES is the excitation-pulse generator for source information using the F0 contour. STRAIGHT-core estimates the spectrum envelope using F0-adaptive time-frequency smoothing to eliminate periodic interferences, and synthesizes sound using the spectrum envelope and excitation pulses. In the synthesis process, the spectrum envelopes are not manipulated.

4.2. Psychoacoustical experiments

In order to investigate whether the proposed F0 control model can control the F0 contour of singing voices, we carried out psychoacoustical experiments in the following manner. We added each fluctuation onto the F0 contour of the melody component using the F0 control model, and synthesized singing voices using those F0s. We presented them to subjects to judge their naturalness.

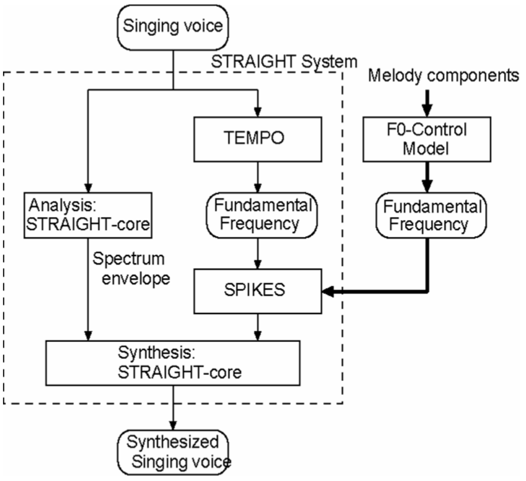

Figure 6: *Singing-voices synthesis method using STRAIGHT and F0 control model*

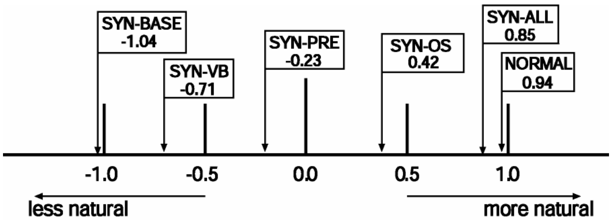

Figure 7: *Results of synthesized singing-voices evaluation*

Six stimuli were used in the experiment as follows.

- **NORMAL:** using extracted F0 (reference)
- **SYN-All:** using all fluctuations
- **SYN-OS:** using Overshoot
- **SYN-PRE:** using Preparation
- **SYN-VB:** using Vibrato and Fine-fluctuation
- **SYN-BASE:** adding no fluctuations to the Melody component.

Control parameters for all F0 fluctuations are shown in Table 1. The title of the Japanese children's ballad is "Nanatsunoko" as in Sec. 2. The spectrum of synthesized singing-voices was identical for all stimuli. The experiments were carried out on the same procedure and the same conditions as described in Sec. 3.

4.3. Results and Discussion

The results in Figure 7 show that the naturalness of synthesized singing-voices increases by adding each F0 fluctuation onto the F0 contours, and the quality of the SYN-ALL is almost the same as that of a real singing voice. The results also show that Overshoot is the most effective F0 fluctuation for singing-voice perception. These results indicate not only that the F0 fluctuations are important for singing-voices perception but also that the proposed F0 control model can be applied to natural singing-voice synthesis.

5. Conclusion

In this paper, we have reconsidered significant F0 fluctuations in singing-voices perception as characteristics that need to control the F0 contour of singing-voices. We have also developed an F0 control model that can control F0 fluctuations and then demonstrated whether it can be applied to the singing-voice synthesis method that can produce natural singing-voice sounds. The results are as follows.

1. The naturalness of singing-voices decreases when removing each F0 fluctuation from F0 contours.
2. The naturalness of synthesized singing voices increases by adding each F0 fluctuation into the Melody component.
3. The quality of synthesized singing voices is almost the same as that of real singing voices.

These results show that F0 fluctuations, especially overshoot, vibrato, fine-fluctuation, and preparation are important factors in singing-voices perception. The F0 control model can generate F0 contours including all F0 fluctuations by determining optimal parameters for each F0 fluctuation. Moreover, the singing-voice synthesis method can produce natural singing voices.

Singing-voice synthesis using the proposed F0 control model could generate natural synthesized singing voices, but they were only the sound of a Japanese vowel. Thus, in order to generate synthesized singing voices with lyrics, we will consider that it is necessary to construct a method that can control the spectrum envelope of singing voices.

6. Acknowledgment

This work was supported by a grant-in-aid for scientific research from the JSPS (No. 13610079) and conducted as a program for the "Fostering Talent in Emergent Research Fields" in Special Coordination Found for Promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology.

7. References

[1] Yatabe, M.; Kasuya, H., 1998. Dynamic characteristics of fundamental frequency in singing. *Proc. Autumn Meeting of the Acoustical Society of Japan*. 3-8-6, 545-546.

[2] Odagiri, W.; Kasuya, H., 1999. Study of analysis, synthesis and perception of vocal vibrato. *Proc. Autumn Meeting of the Acoustical Society of Japan*. 1-7-5, 383-384.

[3] Akagi, M.; Kitakaze, H., 2000. Perception of synthesized singing-voices with Fine-fluctuations in their fundamental frequency fluctuations. *Proc. ICSLP2000*, Beijing, vol. III, 458-461.

[4] Kawahara, H.; Katayose, A..; Patterson, R.D.; de Cheveigné, A., 1999. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. *Proc. Eurospeech99*. 2781-2784.

[5] Kawahara, H.; Masuda- Katsuse, I.; 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, 187-207.

[6] Fujisaki, H.; Tatsumi, M., 1981. Analysis of pitch control in singing. *Vocal fold psychology*, University of Tokyo Press, 347-363.

[7] Nakayama, I., 2002. Nihongo Wo Uta Uta Utau. *singing-voices database*. No. 17, 18.

[8] Sundberg, J., 1987. *The Science of the Singing-Voices.* Northern Illinois University Press.