PERCEPTION OF SYNTHESIZED SINGING VOICES WITH FINE FLUCTUATIONS IN THEIR FUNDAMENTAL FREQUENCY CONTOURS

Masato Akagi and Hironori Kitakaze

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
akagi@jaist.ac.jp

ABSTRACT
This paper demonstrates the importance of fine fluctuations quantitatively by measuring the detection thresholds of fine fluctuations in singing-voice F0s, in which voice quality is particularly important. We analyzed the fine fluctuations left by subtracting the melody and vibrato components from estimated F0s, focusing on the modulation frequency (MF) and modulation amplitude (MA). To test a hypothesis that the fine fluctuations in the F0 of singing voices affect the perception of quality and that the magnitude of this effect depends on the MF and MA, we performed four psychoacoustic experiments using synthesized stimuli. The experimental results indicate that our hypothesis was correct, and suggest that, to produce high-quality synthesized speech, one should extract F0s containing fine fluctuations with an MF of over 7 Hz in the analysis and add not only melody and vibrato but also fine fluctuation components to the F0 contours in the synthesis.

Keywords: singing voice, fundamental frequency, fine fluctuation, and voice quality

1. INTRODUCTION
The dynamics of fundamental frequency (F0) contours is an important factor in producing high-quality synthesized speech and as such has been intensely investigated. The quantitative assessment of the perceptual influence of fine fluctuations left by subtracting phrase and accent components in F0 contours has, however, not been investigated as deeply, even though it is known that fine fluctuations are important to the naturalness of synthesized speech.

The authors have reported how rapid fluctuations of fundamental frequencies in continuously uttered vowels influence vowel quality and have showed that vowel qualities with various fundamental frequency fluctuations can be discriminated perceptually [1]. However, the report did not mention the importance of fine fluctuations quantitatively. In this paper, we quantitatively demonstrate the importance of fine fluctuations by measuring the detection thresholds of fine fluctuations in F0s. This paper is one in a series to investigate the relationship between speech quality and the fine fluctuations.

Our investigation concerns singing voices, in which voice quality is particularly important. The data for our experiments were obtained from recordings of five 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. ANALYSIS OF F0 FLUCTUATIONS
Fine fluctuation components of F0s were extracted from singing voices. The modulation frequency (MF) and modulation amplitude (MA) in the fine fluctuation components were then analyzed.

2.1 Singing voice data
The singing voice data for our experiments were obtained from recordings of five 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 with fine fluctuations were estimated using the F0 extraction method, TEMPO in STRAIGHT [2]. We confirmed beforehand that TEMPO could accurately extract fine fluctuations in F0 contours. It can extract the MFs with a precision of up to about one-fifth of F0.

2.3 Results and Discussion
The extracted F0s were passed through an HPF with a 5-Hz cut-off frequency and transformed into running-spectra. Figure 1 shows (a) an analyzed song wave, (b) estimated F0 and smoothed F0, (c) fine fluctuation component: high-passed F0, (d) running-spectra of the fine fluctuation component and long-term spectrum of the whole song wave, and (e) histogram of the fine fluctuations of up to 20 Hz and that the MAs were 20 cent on average and 100 cent at maximum.

Next, we hypothesized that the fine fluctuations in the F0 of singing voices affect the perception of quality and that the magnitude of this effect depends on the MF and MA. To test this hypothesis, we performed four psychoacoustic experiments using synthesized stimuli.

- Ex. 1: perceptual comparison of synthesized singing voices with and without fine fluctuations
- Ex. 2: measurement of detection thresholds of MF and MA
- Ex. 3: measurement of detection thresholds of two different MFs
- Ex. 4: measurement of detection thresholds when F0s contained many MFs

The experimental results indicate that our hypothesis was correct and suggest that, to produce high-quality synthesized speech, one should extract F0s containing fine fluctuations with an MF of over 7 Hz in the analysis and add not only melody and vibrato but also fine fluctuation components to the F0 contours in the synthesis.
Fundamental frequency and its smoothing

From these findings, we concluded that humans may be able to perceive the MF and MA of fine fluctuation components in singing voices. We hypothesized that the fine fluctuations in the F0 of singing voices affect the perception of quality and that the magnitude of this effect depends on MF and MA.

3. PSYCHOACOUSTIC EXPERIMENTS

Four psychoacoustic experiments were carried out. Experiment 1 examined whether the difference between synthesized singing voices with and without fine fluctuations can be perceived. Experiment 2 tried to determine the detection thresholds of MF and MA. Experiment 3 tried to determine the detection thresholds of two different MFs, and Experiment 4 tried to determine the detection thresholds when F0s contained many MFs.

3.1 Synthesized singing voice

The stimuli were synthesized singing voices with the vowel /a/ using the Klatt formant synthesizer to reflect F0 fluctuations. Formant frequencies of the vowel /a/ were 800, 1200, 2500, 3500, 4500 and 5500 Hz, and each bandwidth was 10 percent of the corresponding formant frequency. The response of the synthesizer with such formant properties was calculated using the Rosenberg wave as input.

The excitation impulse trains were made as follows: Let us assume the F0 transition with fluctuations is $f_{tn}(t)$. If the pulse is set at time $t_n$, the next pulse must be set at $t_{n+1} = t_n + 1/f_{c}(t_n)$.

The synthesized singing voices were made by convoluting the response of the synthesizer with the excitation impulse trains.

3.2 Experiment 1

A. Stimuli

To examine whether we can perceive the difference between synthesized singing voices whose F0s have been extracted and those whose F0s have been smoothed, we filtered the extracted F0 fluctuations using an LPF with a 7-Hz cutoff frequency. The filtered F0 denotes $f_{tmf}(t)$. The pulse trains for the synthesized singing voices were generated from $f_{tm}(t)$ and $f_{tmf}(t)$ using the same procedure as in Eq. (1). The stimuli were given in pairs. The number of the paired stimuli was 20 (4 pairs x 5 singers).

B. 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 13 graduate students. The task was to judge whether the paired synthesized songs were the same or not.

C. Results and Discussion

Since the percentage of correct answers was 78.8%, the subjects can perceive the difference between the synthesized singing voices with fluctuations and those without fine fluctuations. This indicates that differences in F0 fluctuations do influence singing voice quality. Additionally, the subjects said that the voices with the extracted F0s sounded natural.

3.3 Experiment 2

The detection thresholds of MF and MA were determined. Experiment 3 tried to determine the detection thresholds of two different MFs, and Experiment 4 tried to determine the detection thresholds when F0s contained many MFs.

3.4 Experiment 3

The detection thresholds of two different MFs were determined. Experiment 4 tried to determine the detection thresholds when F0s contained many MFs.

3.5 Experiment 4

The detection thresholds when F0s contained many MFs were determined.
3.3 Experiment 2

A. Stimuli

To determine the detection thresholds of MF and MA, we provided two types of F0 transitions; a fixed F0 $f_0$ and a F0 with fluctuation,

$$f_m(t) = (1 + \alpha \sin(2\pi f_t t)) f_0,$$

(2)

where $f_0$ is MF [Hz] and $\alpha$ is MA/F0 [%], and synthesized vowels as the same procedure as in Sec. 3.1. The $f_0$ was fixed at 125 or 250 Hz. The $f_t$ was varied from 2 to 62 Hz (F0: 125 Hz) and from 2 to 90 Hz (F0: 250 Hz) in 2 Hz steps, and the $\alpha$ was varied from 0.1 to 2.0% in 0.1% steps. The stimuli with $f_0$ only (A) and with $f_m(t)$ (B) were given in pairs. The (A)-(A) and (B)-(B) pairs were also mixed as control stimuli.

B. Procedure

The procedure adopted in this experiment is the same as in Experiment 1. The subjects were 15 graduate students. This procedure was also used in Experiments 3 and 4.

C. Results and Discussion

Figure 2 shows a result of Experiment 2 when $f_0 = 125$ Hz and $\alpha = 1.0\%$. The vertical axis is the percentage of correct answers on the (A)-(B) and (B)-(A) stimulus pairs. The figure indicates that the fluctuations can not be perceived when MF exceeds about 15 Hz.

Figure 3 also shows another result, when $f_0 = 125$ Hz and $f_t = 24$ Hz. The vertical axis is the same as in Fig. 2. The figure indicates that the fluctuation can be perceived when $\alpha$ exceeds 1.1 %.

Figure 4 illustrates all the results of Experiment 2 when $f_0 = 125$ Hz. From this figure, the detection thresholds of the fluctuations were $[\text{MF(Hz)}, \text{MA/F0(\%)}] = (6 \text{ Hz}, 0.7 \%), (12 \text{ Hz}, 1.0 \%), (24 \text{ Hz}, 1.1 \%),$ and $(48 \text{ Hz}, 1.4 \%)$. These results indicate that the fine fluctuations with a lower MF or a larger MA were more detectable.

When the F0 was 250 Hz, the detection threshold of MF was 60 Hz ($\alpha=1.0\%$). This suggests that the fluctuations were more detectable when F0 rises.

The magnitudes of the extracted MF and MA exceeded the detection thresholds. Note that the extracted MF and MA were $[\text{MF(Hz)}, \text{MA/F0(\%)}) = (20 \text{ Hz}, 1.2 \%)$ when the F0 was 125 Hz. These results suggest that the magnitude of the effect on the perception of singing-voice quality depended on the MF and MA of the fine fluctuations.

3.4 Experiment 3

A. Stimuli

To determine the detection thresholds of two different MFs, we provided two types of F0 transitions; one was the same as $f_m(t)$ in Experiment 2, and the other is given by

$$f_m(t) = (1 + \alpha \sin(2\pi (f_t + \Delta f_t) t)) f_0,$$

(3)

where $\Delta f_t$ is the MF difference, and vowel synthesis was done using the same procedure as in Sec. 3.1. The $f_0$ and $\alpha$ were fixed at 125 or 250 Hz and 1.0 or 2.0 %, respectively. The $f_t$ was var-
A. Stimuli
To determine the detection thresholds when F0s contained many MFs, we used three F0 transitions; one is the same as $f_a(t)$ in Experiment 2 and the others were

$$f_{a'}(t) = \left[ 1 + \alpha_1 \sin(2\pi f_0 t + \alpha_2 \sin(2\pi f_0 t)) \right] f_0$$

$$f_{a''}(t) = \left[ 1 + \alpha_1 \sin(2\pi f_0 t + \alpha_2 \sin(2\pi f_0 t + \Delta f_z t)) \right] f_0$$

where $f_a(t)$ models a slow fluctuation like vibrato and $f_{a'}(t)$ is the MF difference, and vowel synthesis was done using the same procedure as in Sec. 3.1. The F0 was fixed at 125 or 250 Hz, and the

B. Results and Discussion
Figure 5 shows a result of Experiment 3 when $f_a = 125$ Hz. The vertical axis is percentage of correct answers on the (A)-(B) and (B)-(A) stimulus pairs. The figure indicates that the subjects can discriminate as far as between 5-Hz fluctuation and 6-Hz fluctuation in this condition.

Table 1 shows the results of Experiment 2. The table indicates that the differences in MF were more detectable when $\Delta f_z$ was 1, 5, and 10 Hz. The detection thresholds did not vary, irrespective of the presence of vibrato components. Thus, existing fine fluctuation could have an influence on voice quality, regardless whether melody and/or vibrato components exist or not.

### 3.5 Experiment 4

A. Stimuli
To determine the detection thresholds when F0s contained many MFs, we used three F0 transitions; one is the same as $f_a(t)$ in Experiment 2 and the others were

$$f_{a'}(t) = \left[ 1 + \alpha_1 \sin(2\pi f_0 t + \alpha_2 \sin(2\pi f_0 t)) \right] f_0$$

$$f_{a''}(t) = \left[ 1 + \alpha_1 \sin(2\pi f_0 t + \alpha_2 \sin(2\pi f_0 t + \Delta f_z t)) \right] f_0$$

where $f_a(t)$ models a slow fluctuation like vibrato and $f_{a'}(t)$ is the MF difference, and vowel synthesis was done using the same procedure as in Sec. 3.1. The F0 was fixed at 125 or 250 Hz, and the

B. Results and Discussion
Figure 5 shows a result of Experiment 3 when $f_a = 125$ Hz. The vertical axis is percentage of correct answers on the (A)-(B) and (B)-(A) stimulus pairs. The figure indicates that the subjects can discriminate as far as between 5-Hz fluctuation and 6-Hz fluctuation in this condition.

Table 1 shows the results of Experiment 2. The table indicates that the differences in MF were more detectable when $\Delta f_z$ was 1, 5, and 10 Hz. The detection thresholds did not vary, irrespective of the presence of vibrato components. Thus, existing fine fluctuation could have an influence on voice quality, regardless whether melody and/or vibrato components exist or not.

### 4. GENERAL DISCUSSION

These results indicate that our hypothesis, the fine fluctuations in the F0 of singing voices affect the perception of quality and the magnitude of this effect depends on the MF and MA, was correct. Additionally, the results suggest that, to produce high-quality synthesized speech, one should extract F0s containing fine fluctuations with an MF of over 7 Hz in the analysis and add not only melody and vibrato but also fine fluctuation components to the F0 contours in the synthesis. Thus, STRAIGHT [2] should be one of candidates for speech analysis and synthesis.

### ACKNOWLEDGMENTS

This work was supported by CREST and the JAIST Foundation.

### REFERENCE
