PERCEPTION OF FUNDAMENTAL FREQUENCY FLUCTUATION

PACS Reference: 43.66.Jh

AKAGI Masato
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
Tel. +81-761-51-1236, akagi@jaist.ac.jp

ABSTRACT
Fundamental frequencies (F0s) of speech contain not only slow and large fluctuations related to prosodic information but also rapid and fine fluctuations. Especially in singing voices, there are many F0 fluctuations, for example, overshoot, vibrato etc., as well as melody components. This paper introduces the results of an analysis of F0 fluctuations in singing voices and sustained vowels, and discusses how these fluctuations influence the vowel quality and naturalness of singing voices.

1. INTRODUCTION

Fundamental frequencies (F0s) of speech contain slow and large fluctuations related to prosodic information and rapid and fine fluctuations related to the naturalness of speech. Especially in singing voices, there are many F0 fluctuations, for example, overshoot, vibrato, etc., as well as melody components. 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.

This paper discusses the importance of using fluctuations in F0 contours to synthesize natural speech, by summarizing the author's published work [1-3]. The paper first introduces the analyzed results of a study of F0 fluctuations in singing voices and sustained vowels, and then discusses how these fluctuations influence vowel quality and the naturalness of singing voices.

2. ANALYSIS OF F0 FLUCTUATIONS

2.1 Singing voice

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. The F0s were estimated using the F0 extraction method, TEMPO in STRAIGHT [4]. 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.

Figure 1 shows an estimated F0 contour along the logarithmic axis. We extract five F0 fluctuation characteristics as follows.

Melody component: This represents the note change.

Overshoot: Deflection exceeding the target note after note changes.

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

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

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

The analyzed F0s for fine fluctuations show that the modulation frequencies (MFs) contained frequency components 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 [MF(Hz), MA/F0(%)] = (20Hz, 1.2%), when F0 was 125 Hz. From these findings, we concluded that humans might 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 the MF and MA.

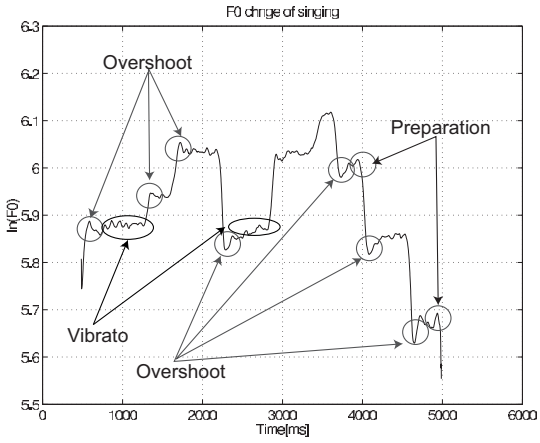

Fig. 1. Estimated F0; portion of "Nanatsunoko," /kawaii nanatsuno/ and fluctuation characteristics.

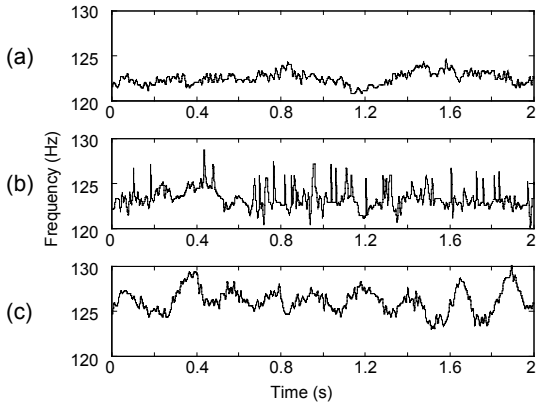

Fig. 2. Three typical F0 contours. (a) not including particular fluctuation, (b) including relatively rapid fluctuation, and (c) including relatively slow fluctuation.

2.2 Sustained vowels

Electroglottograph (EGG) waves L from nine male speakers were recorded for about 10 sec using a Laryngograph having 48-kHz sampling and 16-bit accuracy, to estimate the F0s of vowels. The data obtained were for five Japanese vowels. When uttering the vowels, the speakers monitored a 130-Hz pure-tone through a headphone and tried to keep their fundamental frequencies at 130 Hz. Since L changes rapidly at the closing points of the glottis, we chose the changing points of L and estimated the F0s as reciprocals of the time intervals between adjacent changing points.

The means (M) of the F0s in time were distributed between about 125 Hz and 135 Hz, even though the speakers monitored a 130-Hz pure-tone. The standard deviations (SD) were 0.5 to 3 Hz (0.4 - 2.4% of the mean F0), independent of speakers and vowels. Three typical F0 waves are shown in Fig. 2.

3. PSYCHOACOUSTIC EXPERIMENTS

We extracted five types of F0 fluctuations. These fluctuations may affect perception of voice quality. Four psychoacoustic experiments were carried out to clarify how much these fluctuations influence voice quality. Experiment 1 determined how much overshoot, vibrato, and preparation influenced the naturalness of the singing voices. Experiment 2 examined whether the difference between synthesized singing voices with and without fine fluctuations can be perceived. Experiment 3 tried to determine the detection thresholds of MF and MA. Experiment 4 tried

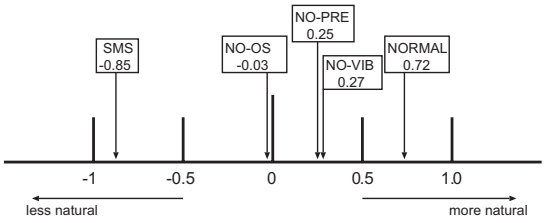

Fig. 3. Result of Experiment 1: Importance of overshoot, vibrato, and preparation.

to determine the detection thresholds of two different MFs.

3.1 Stimuli synthesize

The stimuli were synthesized voices of the vowel /a/ using the Klatt formant synthesizer to reflect F0 fluctuations. 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.

3.2 Experiment 1: Importance of Overshoot, Vibrato, and Preparation

A. Stimuli

We eliminated each the F0 fluctuation from F0 contours and re-synthesized the singing voices using the modified F0s. The same procedure as in Sec. 3.1 was used to synthesize the voices.

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.

B. Procedure

Scheffe's method of paired comparison was used to evaluate the naturalness of singing voices (Seven-grade evaluation measure (-3 to 3) were used. The subjects were six graduate students having normal hearing ability.

C. 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 fluctuations, overshoot, vibrato, and preparation, on singing-voice perception are large, and the effect of overshoot is the largest.

3.3 Experiment 2: Perceptibility of fine fluctuation

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 from the singing-voice data, using an LPF with a 7-Hz cutoff frequency.

The filtered F0 denotes $f_{mf}(t)$. The pulse trains for the synthesized singing voices were generated from $f_m(t)$ and $f_{mf}(t)$ using the same procedure as in Eq. (1). The formant frequencies of the vowel /a/ were 800, 1200, 2500, 3500, 4500, and 5500 Hz, and each bandwidth was 10% of the corresponding formant frequency.

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 answers that were correct was 78.8%, the subjects could 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.

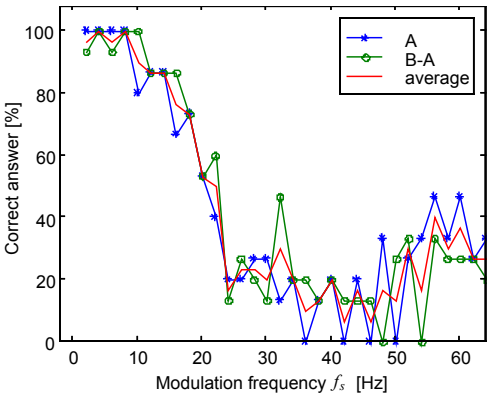

Fig. 4. Result of Experiment 2 when the F0 was 125 Hz and $a = 1.0\%$. The vertical axis is the percentage of correct answers on the A-B and B-A stimulus pairs.

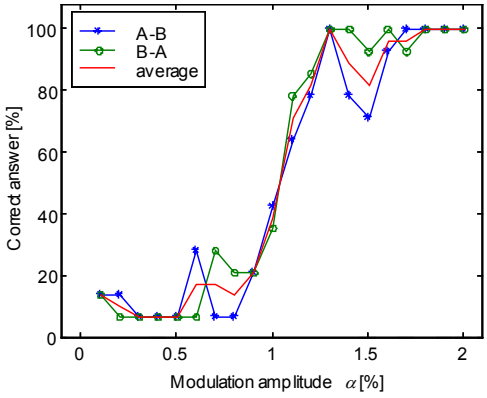

Fig. 5. Result of Experiment 2 when the F0 was 125 Hz and $f_s = 24Hz$. The vertical axis is the same as that in Fig. 4.

3.4 Experiment 3: Detection threshold of MF and MA in fine fluctuation

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,

$$f_m(t) = (1 + a\sin(2pf_s t))f_0, \qquad (2)$$

where f_s is MF [Hz] and a is MA/F0 [%], and the synthesized vowels are the same procedure as in Sec. 3.1. The f_0 was fixed at 125 or 250 Hz. The f_s 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 a 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 used in this experiment was the same as that in Experiment 2. The subjects were 15 graduate students. This procedure was also used in Experiment 4.

C. Results and Discussion

Figure 4 shows the result of Experiment 3 when $f_0 = 125Hz$ and $a = 1.0\%$. The vertical axis is the percentage of correct answers for the (A)-(B) and (B)-(A) stimulus pairs. The figure indicates that the fluctuations cannot be perceived when MF exceeds approximately $15Hz$. Figure 5 shows the result when $f_0 = 125Hz$ and $f_s = 24Hz$. The vertical axis is the same as that in Fig. 4. The figure indicates that a fluctuation can be perceived when a exceeds 1.1%.

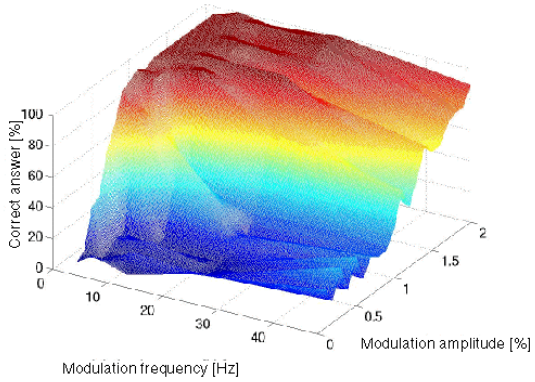

Fig. 6. Result of Experiment 2 when the F0 was 125 Hz.

Figure 6 illustrates all the results obtained for Experiment 3 when $f_0 = 125Hz$. From this figure, the detection thresholds of the fluctuations were [MF(Hz), MA/F0(%)] = (6 Hz, 0.7%), (12 Hz, 1.0%), (24 Hz, 1.1%), and (48 Hz, 1.4%). These results indicate that the fine fluctuations having a lower MF or a larger MA were more detectable.

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

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

Table 1. Results of Experiment 4.

MF difference: Df_s	F0: 125 Hz		F0: 250 Hz	
	a : 1.0 %	a : 2.0 %	a : 1.0 %	a : 2.0 %
1 Hz	5 Hz	7 Hz	6 Hz	8 Hz
5 Hz	9 Hz	11 Hz	11 Hz	13 Hz
10 Hz	12 Hz	14 Hz	17 Hz	over 20 Hz

3.5 Experiment 4: Detection threshold of two different MFs in fine fluctuation

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 3, and the other was given by

$$f_{md}(t) = [1 + a \sin(2p(f_s + Df_s)t)]f_0, \quad (3)$$

where Δf_s is the MF difference, and vowel synthesis was done using the same procedure as in Sec. 3.1. The f_0 and a were fixed at 125 or 250 Hz and 1.0 or 2.0%, respectively. The f_s was varied from 2 to 62 Hz (F0: 125 Hz) and 2 to 90 Hz (F0: 250 Hz) in 2 Hz steps and the Df_s was 1, 5, and 10 Hz. The stimuli with $f_m(t)$ (A) and with $f_{md}(t)$ (B) were given in pairs. The (A)-(A) and (B)-(B) pairs were also mixed as control stimuli.

B. Results and Discussion

Figure 7 shows the result of Experiment 4 when $f_0 = 125Hz$, $a = 1.0\%$, and $Df_s = 1Hz$. The vertical axis is the percentage of correct answers for the (A)-(B) and (B)-(A) stimulus pairs. The figure indicates that the subjects could discriminate between a 5-Hz fluctuation and a 6-Hz fluctuation in this condition.

Table 1 shows the results of Experiment 4. The table indicates that the differences in MF were more detectable when Df_s, MA, and F0 became large. The Df_s, MA, and F0 deviated in the songs and their fluctuations over time could be perceived when a fluctuation exceeded the detection thresholds.

4. SYNTHESIS OF F0 FLUCTUATION

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

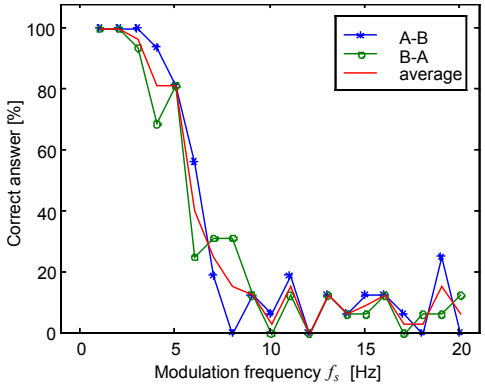

Fig. 7. Result of Experiment 3 when the F0 was 125 Hz, $a = 1.0\%$, and $Df_s = 1Hz$.

4.1 F0 control model for singing voice

Figure 8 shows a schematic graph of the proposed F0 control model. This model generates F0 contours using five components; melody, overshoot, vibrato, preparation, and fine-fluctuation components. The overshoot and preparation components are controlled using a 2nd order damping model and the vibrato component is expressed with a 2nd order oscillation model (no-loss). The fine-fluctuation component is higher than 10-Hz MF and up to 5-Hz MA.

4.2 Singing-voice synthesis

Singing voices were synthesized with the generated F0 contours using STRAIGHT [4]. The spectral data were the same as the analyzed singing-voice data in Sec. 2.1. The following six stimuli were synthesized and presented to the subjects; **NORMAL**: using the extracted F0, **SYN-All**: using all the fluctuation components, **SYN-OS**: using the overshoot component, **SYN-PRE**: using the preparation component, **SYN-VB**: using the vibrato and fine-fluctuation components, and **SYN-BASE**: adding no components to the melody. Figure 9 shows a generated F0 contour having all fluctuation components. The experimental procedure used was the same as that of Experiment 1 in Sec. 3.2.

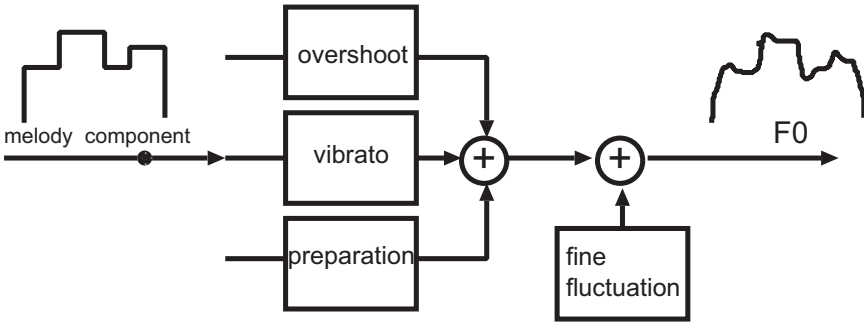

Fig. 8. Schematic graph of F0 control model

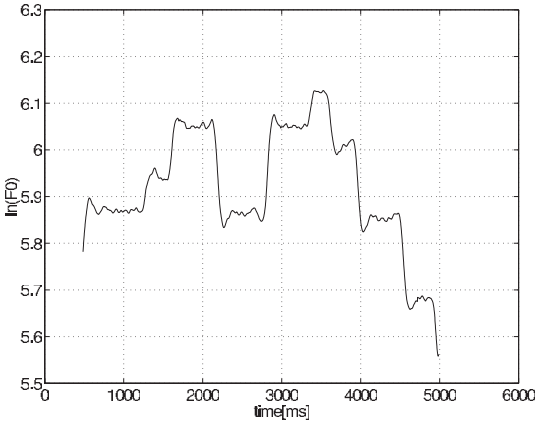

Fig. 9. Generated F0 contour (same portion as that shown in Fig. 1).

4.3. Results and Discussion

The results in Fig. 9 show that the F0 control model can produce singing voices that sound as natural as NORMAL voices, when all the fluctuation components are added to the melody component. This result indicates that the F0 fluctuations are important to the naturalness of singing voices.

5. SUMMARY

This paper reported the results of psychoacoustic experiments where we extracted some F0 fluctuations in singing voices and sustained vowels and demonstrated how much the F0 fluctuations influence voice quality. The results show that F0 fluctuations, especially overshoot, vibrato, preparation, and fine-fluctuation, affect the naturalness of singing voices, and that the fine-fluctuation influences the quality of sustained vowels. These fluctuations are important to the synthesis of natural speech.

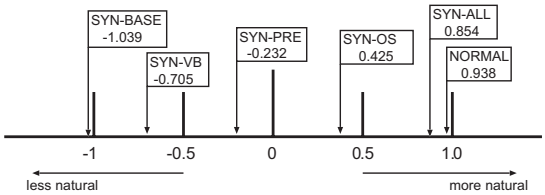

Fig. 10. Result of synthesized singing-voice evaluation.

ACKNOWLEDGEMENT

This work was supported by CREST of JST and by a grant-in-aid for scientific research from the Ministry of Education (No. 13610079), and was done in collaboration with Mamoru Iwaki, Tomoya Minakawa, Hironori Kitakaze, Tsuyoshi Saitou, and Masashi Unoki.

REFERENCES

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

[2] Akagi, M. and Kitakaze, H. (2000). "Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours," Proc. ICSLP2000, Beijing, III-458-461.

[3] Saitou, T., Unoki, M., and Akagi, M. (2002). "Extraction of F0 dynamic characteristics and development of F0 control model in singing voice," ICAD2002, Kyoto (to appear).

[4] 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 F0 extraction: Possible role of a repetitive structure in sounds," Speech Comm., 27, 187-207.