FUNDAMENTAL FREQUENCY FLUCTUATION IN CONTINUOUS VOWEL UTTERANCE AND ITS PERCEPTION

Masato Akagi, Mamoru Iwaki and Tomoya Minakawa

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

ABSTRACT

This paper reports how rapid fluctuations of fundamental frequencies in continuously uttered vowels influence vowel quality and shows that vowel qualities with various fundamental frequency fluctuations can be discriminated perceptually. For this purpose, electroglottographs (EGGs) of vowels uttered by nine males were obtained using Laryngograph, and fundamental frequencies with rapid fluctuations were estimated from them. Analyzing forty-five estimated fundamental frequencies, they can be classified into four groups. Moreover, psychoacoustic experiments, with five subjects, evaluating voice quality by multidimensional scaling (MDS) showed that voice quality of the synthesized speech using the fundamental frequencies of the groups was completely discriminable and there was a distinctive frequency band of fundamental frequency fluctuation for specifying each group perceptually.

Keywords: fundamental frequency fluctuation, electroglottograph, multidimensional scaling, voice quality

1. INTRODUCTION

Even though speakers try to keep pitch frequency fixed when uttering vowels, periods of glottal vibration are not uniform and fluctuate finely in time; this fluctuation affects voice quality [1][2][3]. Moreover, when a vowel waveform is synthesized using an uniform, non-fluctuating pulse train, its quality seems to be unnatural, like a buzzer. However, vowel quality is improved by varying the pulse intervals. This paper focuses on the relationship between the perception of vowel quality and fine fluctuations of the fundamental frequency (F0), especially how rapid fluctuations of fundamental frequencies in continuously uttered vowels influence vowel quality and whether vowel qualities with various fundamental frequency fluctuations can be discriminated perceptually.

For this purpose, first, electroglottographs (EGGs) of the five vowels of Japanese uttered by nine males were measured using Laryngograph, and F0s with fine fluctuations were estimated from the closing points of the glottis obtained from the EGGs. Using coefficients of variation of the F0s, the forty-five estimated F0s could be classified into four groups: (A) not including any particular fluctuation, (B) including relatively rapid fluctuation, (C) including relatively slow fluctuation, and (D) including both rapid and slow fluctuations.

Next, experiments on discriminating vowels synthesized by the Klatt formant synthesizer using typical F0s of groups (A), (B), and (C) with fine fluctuations were performed, to determine whether subjects (this time five) could discriminate between F0 fluctuations of the groups. Only F0 fluctuation differed among the synthesized vowels. The results suggest that the qualities of vowels of the three groups were clearly different from each other and could be discriminated almost completely.

Additionally, experiments comparing voice qualities were performed using vowels synthesized using averaged F0s, fluctuation-reduced F0s obtained using high- or low-pass filters, and the original F0s with fine fluctuations, to determine which frequency bands of the F0 fluctuations were significant for perceiving voice quality of each group. The experiment results were analyzed using a multidimensional scaling (MDS) technique. They suggest that each group had a distinctive F0 fluctuation frequency that enabled it to be discriminated perceptually.

2. ESTIMATION OF F0

In order to estimate F0s of vowels with fine fluctuations, EGG waves L were recorded for about 10 s using Laryngograph with 48-kHz sampling and 16-bit accuracy, from nine male speakers. The data 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 F0s as reciprocals of the time intervals between adjacent changing points. At first, the recorded EGG waves L were filtered using a low-pass filter (LPF) whose cut-off frequency was 2 kHz to eliminate high-frequency noise components. Next, derivatives of the filtered L were calculated to specify the closing points of the glottis, which are indicated by peaks in the derivatives. Then, F0s at any time could be estimated by linearly interpolating reciprocals of time intervals between peaks. Although unusual peaks and intervals were often found in the derivatives, they were corrected by hand.

Figure 1(a) shows an F0 wave of vowel /a/ estimated by the above mentioned method and Fig. 1(b) is a comparable F0 wave of the same vowel estimated by the auto-correlation of the LPC residual. Parameters of the auto-correlation were frame = 30 ms, Hamming, and frame period = 5 ms. The figure shows that the F0 contour estimated from the EGG wave (a) had finer fluctuations than (b).

3. ANALYSIS OF F0 FLUCTUATION

The analyzed data were F0s with fine fluctuations estimated from the EGG waves; they were sections 2 s long manually cut out from stable portions of vowels. The number of the data was 45 (5 vowels x 9 speakers). The means (M) of F0 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, independent of speakers and vowels.

In order to classify F0 fluctuations into groups, we calculated power spectra of F0s by a 96,000-point Fourier transform (FT) (48-kHz sampling, 2-s duration). Figures 2(a), (b), and (c) show log power spectra of three typical F0 waves shown in Fig. 3(a), (b), and (c), respectively. The panels in Fig. 2 show interest-

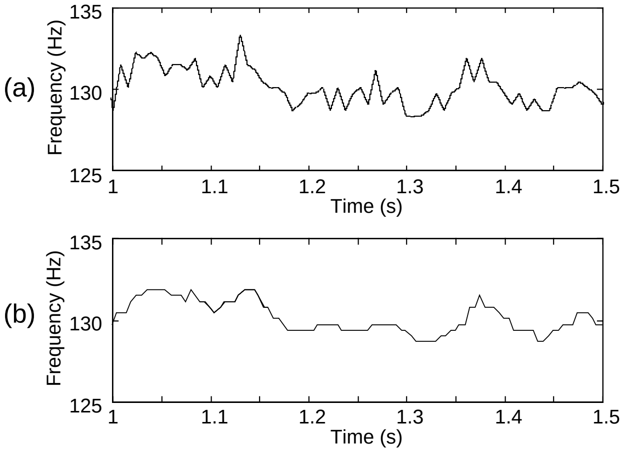

Fig. 1 Estimated F0 waves of vowel /a/ by (a) the proposed method and by (b) auto-correlation of LPC residual.

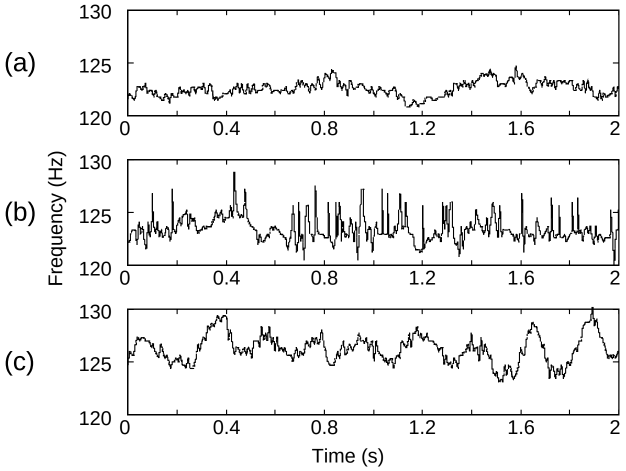

Fig. 3 Three typical F0 waves (a), (b), and (c) classified into Groups (A), (B), and (C), respectively.

ing features: panel (a) shows small magnitudes in the whole frequency region, panel (b) shows large magnitudes in the high frequency region, and panel (c) shows large magnitudes in the low frequency region. The large magnitude in the low frequency region indicates that F0 fluctuated slowly and the large magnitude in the high frequency region indicates that F0 fluctuated rapidly. The boundary of the frequency regions was about 10 Hz.

The F0 waves were low- or high-pass filtered with a 10-Hz cut-off frequency and analyzed statistically. The filtering procedure was as follows: the F0 waves were translated by the 96,000-point FT, components lower or higher than 10 Hz were replaced by zero, and then they were reproduced by the 96,000-point inverse FT. The low-pass filtered F0s had slow fluctuation lower than 10 Hz and the high-pass filtered F0s had rapid fluctuation higher than 10 Hz. Figures 4 and 5 illustrate the slow and rapid fluctuations of the F0 waves shown in Fig. 3(b) and (c).

F0 fluctuations were classified into four groups based on the coefficients of variation (CV) of the low- and high-pass filtered waves. The CV is a statistical criterion,

$$CV = SD/M, \qquad (1)$$

where SD is the standard deviation and M is the mean. When the CV of the low-passed wave (CVl) was larger than 0.0075, the wave was considered as having slow fluctuation and when the CV of the high-passed wave (CVh) was larger than 0.0045, the wave was considered as having rapid fluctuation.

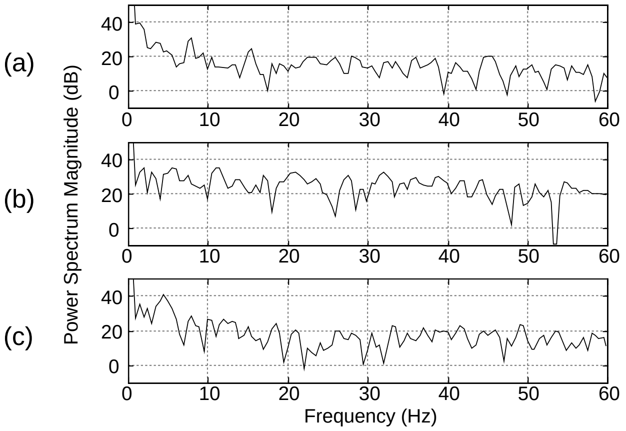

Fig. 2 Log power spectra of three typical F0 waves, which are shown in Fig. 3. The symbols (a), (b), and (c) indicate the same wave data.

The four groups were as follows:

(A) not including any particular fluctuation, CVl < 0.0075 and CVh < 0.0045,
(B) including relatively rapid fluctuation, CVl < 0.0075 and CVh > 0.0045,
(C) including relatively slow fluctuation, CVl > 0.0075 and CVh < 0.0045, and
(D) including both rapid and slow fluctuations, CVl > 0.0075 and CVh > 0.0045.

The F0 waves illustrated in Fig. 3(a), (b), and (c) are typical waves in Groups (A), (B), and (C), respectively. The numbers of vowels in Groups (A), (B), (C), and (D) were 24, 3, 18, and 0, respectively. No waves were classified into Group (D) in the data used for this work, probably because the number of the data was small or F0 fluctuations classified into Group (D) may have been abnormal. This should be checked in future.

4. PSYCHOACOUSTIC EXPERIMENTS

Psychoacoustic experiments were carried out to determine whether subjects could discriminate between F0 fluctuations of the groups (Experiment 1) and which frequency bands of the F0 fluctuations were significant for perceiving voice quality of each group (Experiment 2).

4.1 Experiment 1

A. Stimuli

The stimuli were three re-synthesized speech waves of vowel /a/ coming from Groups (A), (B), and (C) using the Klatt formant synthesizer to reflect fine F0 fluctuation.

To synthesize the stimuli with their own F0 fluctuations, the excitation waves were made as follows:

Let us assume F0 transition

$$F_0(t) = \overline{F}_0 + F_f(t), \qquad (2)$$

where $\overline{F}_0$ is averaged F0 and $F_f(t)$ is F0 fluctuation. If a pulse is set at time t_n, the next pulse must be set at

$$t_{n+1} = t_n + 1/F_0(t_n). \qquad (3)$$

The generated pulse train was filtered to modify each pulse into

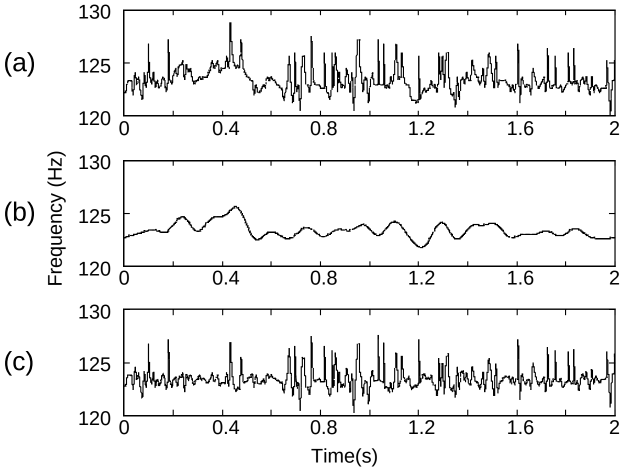

Fig. 4 Low- and high-pass filtered waves of the F0 wave shown in Fig. 3 (b): (a) original wave, (b) low-pass filtered wave, and (c) high-pass filtered wave with DC component.

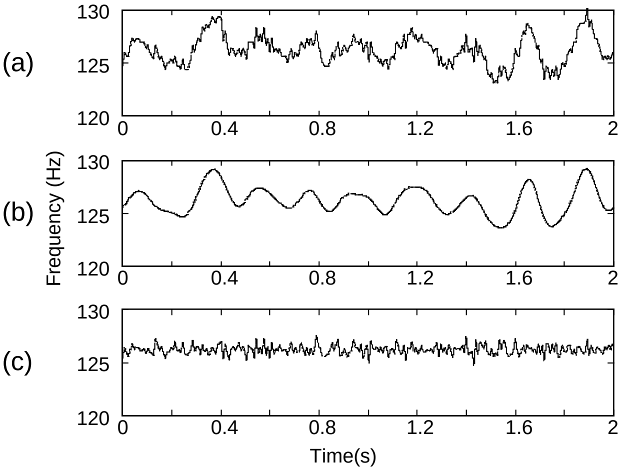

Fig. 5 Low- and high-pass filtered waves of the F0 wave shown in Fig. 3 (c): (a) original wave, (b) low-pass filtered wave, and (c) high-pass filtered wave with DC component.

the Rosenberg wave. The F0 fluctuation $F_f(t)$ came form each group and the averaged F0 $\overline{F}_0$ of Group (A) was used for $F_0(t)$. Formant frequencies and bandwidths were also fixed in the case of Group (A). In the experiments, the number of formants was three and all formant frequencies and bandwidths were measured from sound spectrograms of three speech waves.

The duration of each stimulus was 2 s and both ends were tapered over a period of 50 ms by a sine function. The amplitude was normalized. The stimuli were formed into pairs with a 1-s interval. The number of paired stimuli was nine.

B. Procedure

The paired stimuli were presented through binaural earphones (Sennheiser HDA-200) at a comfortable loudness level. Each paired stimulus was randomly presented to each subject five times. The subjects were five male graduate students very familiar with the characteristics of the speakers' voices. The task was to judge whether two synthesized vowels were the same or not.

C. Results and Discussion

The number of correct answers was 224, out of a total of 225 stimuli (9 types x 5 times x 5 subjects). This indicates that the subjects could discriminate between F0 fluctuations of Groups (A), (B), and (C), and the differences in F0 fluctuations do influence vowel quality.

4.2 Experiment 2

A. Stimuli

To determine which frequency bands of the F0 fluctuations are significant for perceiving voice quality of each group, we filtered the F0 fluctuations $F_f(t)$ using the 96,000-point FT; during low-pass filtering, components higher than the cut-off frequency were replaced by zero and they were translated by the 96,000-point inverse FT. The filtered $F_f(t)$ denotes $\hat{F}_f(t)$.

The pule train for the excitation wave was generated from

$$\hat{F}_0(t) = \overline{F}_0 + \hat{F}_f(t), \qquad (4)$$

using the same procedure as in Eq. (3). The generated pulse train was filtered to modify each pulse into the Rosenberg wave. All parameters (F0 fluctuation $F_f(t)$, averaged F0 $\overline{F}_0$, formant frequencies and bandwidths) came form each group.

The stimulus types were as follows.

SET-A (low-pass filtered set): LPFs with cut-off frequencies of 10, 30, 60, and 100 Hz were used for filtering. Additionally, original F0 $F_0(t)$ and averaged F0 $\overline{F}_0$ were also used for the excitation waves. Thus, the number of stimuli was six (1 original F0 + 4 LPFs + 1 averaged F0).

SET-B (high-pass filtered set): HPFs with cut-off frequencies of 10, 30, 60, and 100 Hz were used for filtering. Others were the same as *SET-A*. Thus, the number of the stimuli was also six.

The duration of each stimulus was 2 s and both ends were tapered over 50 ms by a sine function. The amplitude was normalized. The stimuli were paired in each stimulus set with a 1-s interval. The number of paired stimuli in each stimulus set was 36 and the number of stimulus sets was six (2 sets x 3 groups).

B. Procedure

The paired stimuli in each stimulus set were presented through binaural earphones (Sennheiser HDA-200) at a comfortable loudness level. Each paired stimulus was randomly presented to each subject two times. The subjects were the same as in Experiment 1. The task was to evaluate the similarity of two synthesized vowels in terms of five grades: very different (0) - different (1) - neither different nor similar (2) - similar (3) - very similar (4).

C. Results and Discussion

Perceptual distances for MDS were measured based on the evaluation results. Two-dimensional configurations of points of each stimulus set are represented in Fig. 6. The left panels are for *SET-A* and the right panels are for *SET-B* for Groups (A), (B), and (C). In the figures, Lx, M and $L(H): x$ indicate the original F0, the averaged F0, and LPF (HPF) with cut-off frequency of x Hz, respectively.

Distances between Lx and M in two-dimensional configurations were large and the points of other stimuli were distributed around them, suggesting that the subjects could discriminate the stimuli from each other and that the qualities of synthesized vowels using Lx and M were clearly different. Additionally, vowel qualities with filtered F0s were close to those of either vowels synthesized using Lx or M, depending on the filter cut-off frequencies.

Comparing the left and right panels of Group (C) in Fig. 6(c), the critical cut-off frequency was 10 Hz. This is because all points of stimuli with filtered F0s gather at *Lx* in the left panel and at *M* in the right panel. This indicates that significant features of Group (C) exist in the frequency band lower than 10 Hz. Comparing the left and right panels of Group (B) in Fig. 6(b), the critical cut-off frequency is between 10 and 30 Hz for LPF and between 30 and 60 Hz for HPF. This indicates that significant features for Group (b) are distributed around 30 Hz. For Group (A) in Fig. 6(a), on the other hand, the critical cut-off frequency cannot be set for the LPF, although 10 Hz is critical for HPF. This result suggests that features of Group (A) scatter in the frequency band lower than 60 Hz. These findings show that each group had a particular critical cut-off frequency at which perceptual judgment changed and had a distinctive F0 fluctuation frequency enabling its discrimination perceptually. Considering the log power spectra of the F0 fluctuations shown in Fig. 2, significant features exist in the high-power region.

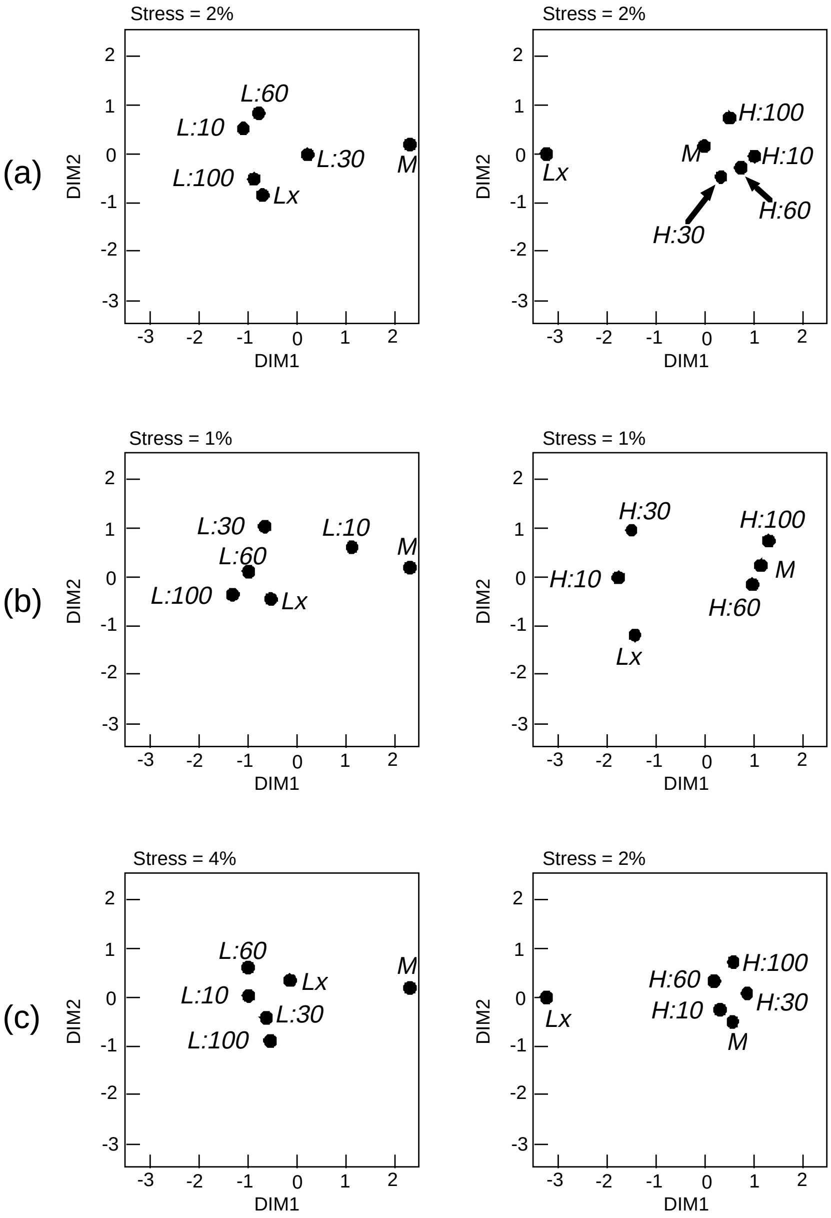

Fig. 6 Two-dimensional configurations of points of each stimulus set. The left panels are *SET A* and the right panels are *SET B* for Groups (A), (B), and (C) (top to bottom).

5. GENERAL DISCUSSION

The results of the two experiments indicate that F0 fluctuation affects voice quality and that different F0 fluctuation causes different voice quality, which can be perceived completely, even though the difference in the cut-off frequency of LPF or HPF to filter F0 fluctuation was less than 20 Hz.

Some F0 fluctuations had rapid fluctuation components. For example, fluctuation frequency components around 30 Hz were distinctive for group (B). If the LPF with 10-Hz cut-off frequency applies to them, voice quality changes into quite a different one. It is doubtful whether well-known pitch prediction methods can extract F0s with fluctuations as rapid as 30 Hz.

The results suggest that new pitch frequency extraction methods to predict fine fluctuations are needed for speech analysis-synthesis techniques and some rules formulating temporal fluctuation of pitch pulses that can represent differences between Groups (A), (B), and (C) are also needed for text-to-speech techniques[4][5].

ACKNOWLEDGMENT

This work was supported by CREST and by Grant-in-Aid for Scientific Research from the Ministry of Education (No. 07680388).

REFERENCES

1. Fourcin, A. F., “Normal and pathological speech: phonetic, acoustic, and laryngographic aspects,” Manual of Laryngograph.
2. Koike, Y., “Application of some acoustic measures for the evaluation of laryngeal dysfunction,” Studia Phonologica (Kyoto Univ.) 7: 17-23, 1973.
3. Kasuya, H., Ogawa, S. and Kikuchi, Y., “An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology,” Speech Comm., 5: 171-181, 1986.
4. Kadambe, S. and Boudreaux-Bartels, G. F., “Application of the wavelet transform for pitch detection of speech signals,” IEEE Trans. Information Theory, 38: 917-924, 1992.
5. Endo, Y. and Kasuya, H., “A speech analysis-conversion-synthesis system taking period-to-period fluctuations into account,” Trans. IEICE, J81-A, 7: 1031-1041, 1998.