A LOW-BAND SPECTRUM ENVELOPE MODELING FOR HIGH QUALITY PITCH MODIFICATION

Ryo MOCHIZUKI†‡ *and Tetsunori KOBAYASHI*†

†Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
‡Matsushita Electric Industrial Co., Ltd.
600 Saedo-cho, Tsuzuki-ku, Yokohama 224-8539, Japan

ABSTRACT

A low-band spectrum envelope reconstruction method was tested to see if it could improve the sound quality of speech modified by the PSOLA(Pitch Synchronous OverLap Add) method. In the conventional TD(Time Domain)-PSOLA method, the spectrum envelope extracted using a Hanning window with a two-pitch-period length had no reliable information in the band of frequencies lower than original F_0. This problem causes the sound degradation of the F_0 modified speech. In the proposed method, the low-band spectrum envelope was properly modified according to the F_0 modification rate. The amplitude of the F_0 harmonic components in the low-band was reproduced based on the spectral tilt of the spectrum envelope. Subjective listening test results suggest this proposed method yields better sound quality than the conventional TD-PSOLA method when the downward modification rate exceeds 0.4 octave.

1. INTRODUCTION

Several spectrum envelope representation methods have been proposed to achieve high quality F_0 modification for speech synthesis[1]-[4]. The spectrum envelope representation technique plays an important role in obtaining a fine spectrum envelope free from interference by an F_0 and its harmonic components. In the TD-PSOLA method[1], a short-term waveform is extracted using a Hanning window with a two-pitch-period length (this short-term waveform is hereafter called the "unit waveform"). To preserve the original sound quality and personal characteristics of the original waveform in the synthesized sound, an extracted unit waveform should represent the original spectrum envelope as closely as possible after unit waveform extraction. For the TD-PSOLA, the spectrum envelope between each F_0 harmonic component is interpolated smoothly using the amplitude spectrum of the window function used to extract the unit waveform; consequently, the fine spectrum envelope is retained in the band above the F_0. However, it is difficult to represent the spectrum envelope below the F_0 because the original unit waveform has no reliable information about that band. This causes the sound degradation of the F_0 modified speech, especially when the F_0 is modified downward.

In this paper, we propose a PSOLA-based F_0 modification method in which the low-band spectrum envelope of the unit waveform is reconstructed. In this reconstruction process, the magnitude of the low-band spectrum peak is estimated using the spectral tilt. As a result of the subjective listening test, the effectiveness of the proposed method was shown.

2. SPECTRUM ENVELOPE MODELING

2.1. Spectrum envelope of the unit waveform

Fig.1 shows an illustration of a spectrum envelope which is constructed by convoluting F_0 harmonic components with the frequency characteristics of the Hanning window. To make the problem easier, the harmonics in this illustration have the same amplitude. For this system, a flat-shaped envelope is formed in the band of frequencies higher than F_0 because the spectrum envelope between each F_0 harmonic component is interpolated by the influence of the amplitude spectrum of the window function. In contrast, the spectrum envelope diminishes abruptly below the F_0. It is thought that there is no reliable spectral information below the original F_0. The low-band spectrum envelope cannot be reproduced by the conventional TD-PSOLA method when the F_0 is shifted downward. It is also thought that the original magnitude of the low-band spectrum might cause the sound degradation when the F_0 is modified upward. Consequently, the spectrum envelope modification along with the F_0 modification rate is expected to reduce these problems.

2.2. Spectrum envelope reconstruction procedure

For our system, a spectrum envelope in the target F_0 band is generated, and the F_0 modification is performed upon it

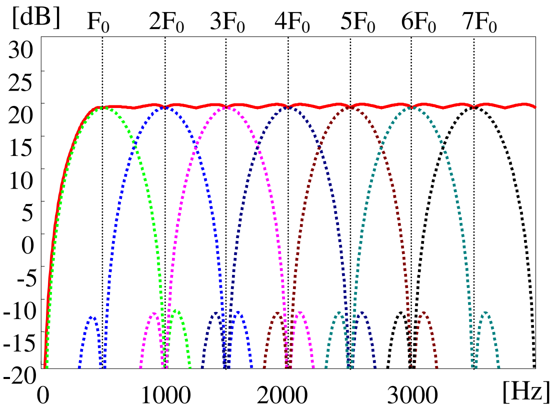

Fig. 1. Spectrum envelope constructed by convoluting F_0 harmonic components with the frequency characteristics of the Hanning window.

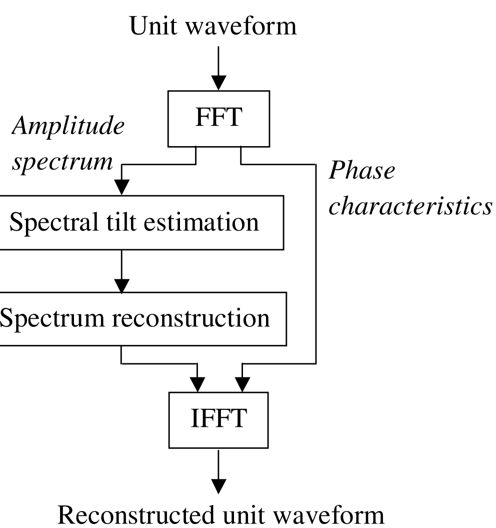

Fig. 2. Spectrum envelope reconstruction procedure.

using the PSOLA method. The low-band spectrum envelope is modified at a rate equal to the given F_0 modification rate. The process of spectrum reconstruction is performed on each unit waveform. The unit waveforms are extracted by referring to pitch marks determined by a residual correlation method[5]. Fig.2 illustrates a procedure of this proposed method, whose steps are explained in the following subsections.

2.2.1. Spectrum envelope (FFT)

The amplitude spectrum of the unit waveform is calculated using short-time FFT. Phase characteristics are stored to use in the process of the unit waveform representation.

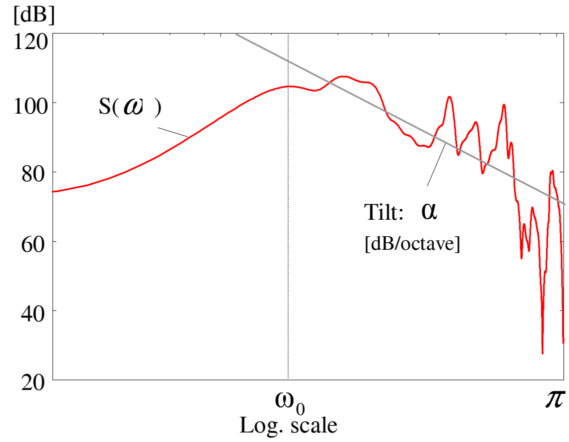

Fig. 3. Estimation of the spectral tilt.

2.2.2. Spectral tilt estimation

As illustrated in Fig.3, the spectral tilt coefficient α [dB/oct.] is estimated using an LMS(Least Mean Square) algorithm. Since the spectrum envelope information below the F_0 is unreliable, the spectrum envelope between F_0 and $F_s/2$ is used to estimate the tilt(F_s is the sampling frequency). The spectral tilt coefficient α is defined as follows:

$$\alpha = \frac{\sum_{\omega_i \in \Omega} \log_2 \frac{\omega_i}{\overline{\omega}} \cdot \ln \frac{|S(\omega_i)|}{\overline{|S(\omega)|}}}{\sum_{\omega_i \in \Omega} (\log_2 \frac{\omega_i}{\overline{\omega}})^2} \tag{1}$$

$$\Omega = \{\omega_i | \omega_0 < \omega_i < \pi\}$$

where ω_0 is the angular frequency corresponding to the original F_0, $\overline{\omega}$ is the mean angular frequency, and $\overline{S(\omega)}$ is the mean spectral amplitude.

2.2.3. Spectrum reconstruction

Fig.4 (a) and (b) show the original spectrum envelopes $S(\omega)$ and low-band reconstructed spectrum envelopes $S'(\omega)$ when the F_0 was shifted downward and upward, respectively. The spectrum envelope $S'(\omega)$ reconstructed by convoluting line spectra with the frequency characteristics of the Hanning window is defined as follows:

$$S'(\omega) = \begin{cases} \sum_{i=1}^{N} A_i \cdot \frac{|W_i(\omega)|}{Wmax_i} & (\omega < \omega'_0 N) \\ S(\omega) & (\omega'_0 N \le \omega) \end{cases} \tag{2}$$

$$Wmax_i = max|W_i(\omega)|, \qquad N = \left\lceil \frac{\omega_0}{\omega'_0} \right\rceil + 1$$

where ω'_0 is the angular frequency corresponding to the target F_0, i is the harmonic number of ω'_0, A_i is the amplitude of the i-th target line spectrum, $W_i(\omega)$ is the frequency characteristics of the window function, and $\lceil x \rceil$ denotes the maximum integer that does not exceed x.

Fig.5 shows the distribution of the spectral tilt values of the unit waveforms extracted from voiced sounds in natural utterances; these sounds cover several F_0's (F_0 range is between 150 and 450 Hz). There is no mutual dependence between the F_0's and the spectral tilt values. Consequently, the spectral tilt for each unit waveform should be kept at its original value independent of the F_0 modification process. Taking into account these considerations, the target line spectrum A_i is given by following equation.

$$A_i = \begin{cases} \exp\{\alpha \log_2(i \cdot \omega_0'/\omega_0)\} \cdot S(\omega_0) \\ \qquad\qquad (i < \omega_0/\omega_0') \\ S(i \cdot \omega_0') \qquad (\omega_0/\omega_0' \leq i) \end{cases} \tag{3}$$

In Eq.(3), the line spectra below the original F_0 are calculated using the spectral tilt coefficient α and $S(\omega_0)$. This process is only required when the F_0 is shifted downward because all the required line spectrum information above the original F_0 can be obtained from the original spectrum envelope $S(\omega)$. The frequency characteristics of the analysis window $W_i(\omega)$ are designed in time-domain as follows:

$$W_i(\omega) = F\big[w_i(t)\big] \tag{4}$$

$$w_i(t) = w_{han}(t, T_0) \cdot \cos(2\pi i t/T_0) \tag{5}$$

where T_0 is pitch period, $F[\cdot]$ denotes FFT, and $w_{han}\,(t, \tau)$ is the Hanning window given by the following equation.

$$w_{han}(t, \tau) = 0.5(1.0 + \cos(\pi t/\tau))(|t| < \tau) \tag{6}$$

2.2.4. *Reconstructed unit waveform (IFFT)*

Finally, each new unit waveform is represented by an IFFT of the reconstructed spectrum envelope $S'(\omega)$ that possesses the original phase characteristics. After rearranging the renewed unit waveforms using the target pitch period, F_0 modification is performed.

3. LISTENING TEST

A listening test was conducted to evaluate the natural quality of the F_0 modified speech whose F_0 contour was shifted by constant rates. The stimuli produced by the proposed method were compared to those produced by the conventional TD-PSOLA method. Because the results for a prior test indicated that there was no perceptible difference between either method at any modification rate for the F_0 upward modification, the listening test in this section was aimed at the downward modification. Six sentences uttered by a female speaker were used for this test (the F_0 of these sentences ranged between 170 and 350 Hz). The F_0 shifting rates ranged between 0.0 and 1.0 octave downward in 0.2 octave steps. Eight subjects were asked to select their preferred sample from order-randomized pairs.

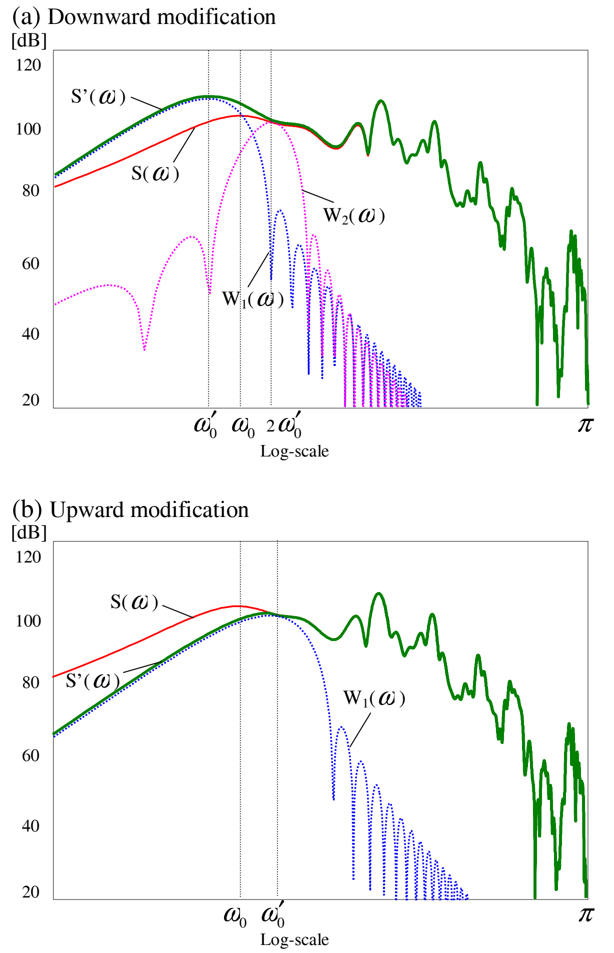

Fig. 4. Reconstracted spectrum envelopes $S'(\omega)$ for (a) downward and (b) upward modifications.

The preference scores for the listening test are shown in Fig.6. These results suggest that the proposed method was preferred to the conventional TD-PSOLA method as the modification rate increases downward. The subjects commented that the samples modified by the proposed method were less hoarse than those modified by the TD-PSOLA method. This subjective evaluation thus suggests the proposed method produces better sound quality than the conventional TD-PSOLA when the downward modification rate is exceeds 0.4 octave.

4. DISCUSSION

Fig.7 panels (a) and (b) show the spectrum envelopes extracted from the downward and the upward F_0 modified waveforms, respectively. Fig.7 (a) shows that a spectrum difference between the proposed and the conventional meth-

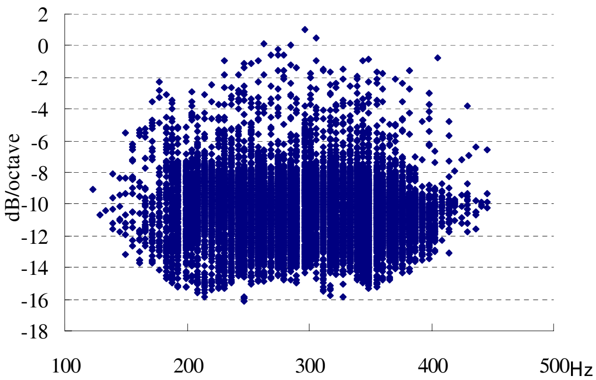

Fig. 5. Distribution of spectral tilts of several unit waveforms extracted from voiced parts of speech.

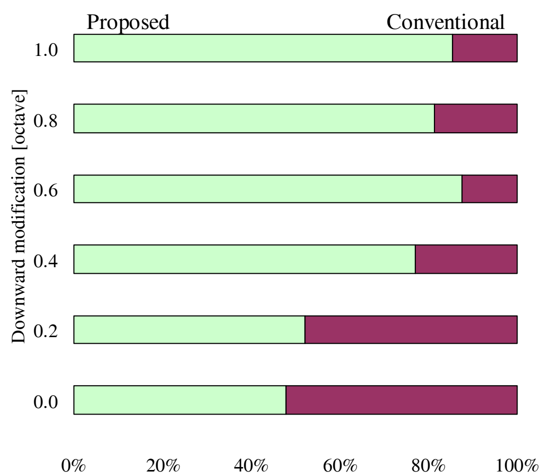

Fig. 6. Result of subjective listening tests. Stimuli modified downward using the proposed method were compared with those modified using the conventional TD-PSOLA method.

ods is present in the band below the F_0. It also shows that the enhanced shape of the low-band spectrum which is modified using the proposed method remains after downward F_0 modification; this enhancement is thought to provide the advantage in sound quality.

In contrast, there is only a slight difference in both low-band spectrum envelopes observed in Fig.7 (b). Shortening the pitch period by the OLA using either method results in similar spectrum envelopes. Consequently, it is thought that the amplitude spectrum below the target F_0 is canceled even if the conventional TD-PSOLA method is used. This results in no perceptible difference between two methods when the F_0 is shifted upward.

5. CONCLUSION

We proposed a new PSOLA-based F_0 modification method in which the spectrum envelope below the F_0 is estimated

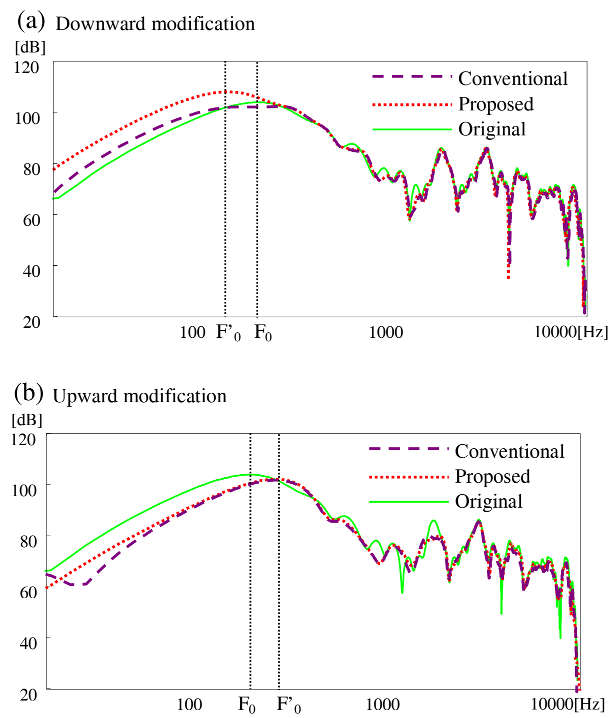

Fig. 7. Spectrum envelope extracted from F_0 modified speech. (a) downward modification, and (b) upward modification.

based on the spectral tilt. In conclusion, the results of a subjective listening test suggest that a proposed F_0 modification method offers better sound quality than conventional TD-PSOLA when the downward modification rate exceeds 0.4 octave.

6. REFERENCES

[1] E.Moulines and F.Charpentier, ”Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Comm., Vol.9, pp.453-467 (1990).

[2] Y.Stylianou, J.Laroche, and, E.Moulines, ”High-quality speech modification based on a Harmonic + Noise Model,” Proc. EUROSPEECH, pp.451-454 (1995).

[3] H.Kawahara, I.Masuda-Katsuse, and Alain de Cheveigne, ”Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Comm., Vol.27, pp.187-207 (1999).

[4] S.Takano, K.Tanaka, H.Mizuno, M.Abe, and S.Nakajima, ”A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction,” IEEE Trans. Speech Audio Process. Vol.9, 3-10 (2001).

[5] Y.Arai, R.Mochizuki, H.Nishimura, and T.Honda, ”An excitation synchronous pitch waveform extraction method and its application to the VCV-concatenation synthesis of Japanese spoken words,” Proc. ICSLP96, Vol.3, pp.1437-1440 (1996).