ACOUSTICAL LETTER

PLATINUM: A method to extract excitation signals for voice synthesis system

Masanori Morise*

College of Information Science and Engineering, Ritsumeikan University,
1–1–1 Nojihigashi, Kusatsu, 525–8577 Japan

(*Received 12 July 2011, Accepted for publication 28 October 2011*)

Keywords: Voice analysis, Voice synthesis, Excitation signal, Vocoder
PACS number: 43.72.Ar, 43.72.Ja [doi:10.1250/ast.33.123]

1. Introduction

Recently, vocal synthesis is a major topic, and high-quality synthesis systems such as Vocaloid2 [1] were released once computer technology had developed sufficiently. To develop a high-quality vocal synthesizer, a voice synthesis system that can manipulate pitch and timbre without sound quality deterioration is required.

Roughly two types of systems for voice synthesis have been proposed. One is based on the time domain pitch synchronous overlap-add (TD-PSOLA [2]), which synthesizes a voice using the short time waveform directly extracted from the input signal. The other is based on a vocoder [3], which analyzes a voice in terms of its pitch (fundamental frequency; F_0) and timbre (spectral envelope) and synthesizes it with the estimated parameters. TD-PSOLA and vocoders have trade offs. TD-PSOLA synthesizes voice with better quality than vocoders; however, vocoders can manipulate pitch and voice timbre independently.

We propose a high-quality voice synthesis system that uses our excitation signal extraction method. A subjective evaluation experiment showed that the proposed system can synthesize a voice more naturally than with conventional vocoder-based systems.

2. Vocoder-based voice synthesis system

2.1. Background

We approximate voiced speech $y(t)$ to the convolution of periodic pulses $x(t)$ and vocal tract response $h(t)$:

$$y(t) = h(t) * x(t), \tag{1}$$

$$x(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_0), \tag{2}$$

where $*$ represents the convolution and T_0 represents the fundamental period that is the inverse value of F_0. The spectra of these signals are given by

$$Y(\omega) = H(\omega)X(\omega). \tag{3}$$

Traditional vocoder-based systems analyze a voice in terms of its fundamental frequency (F_0) and spectral envelope and synthesize the voice with the estimated parameters. Since sound quality of the synthesized voice depends on the estimation performance, many methods for accurately estimating each parameter have been proposed for high-quality voice synthesis [4–6].

Although vocoder-based systems can independently manipulate the F_0 and spectral envelope, low sound quality is a major problem. The STRAIGHT [7] and TANDEM-STRAIGHT [8] have been proposed to solve this problem. They use the pitch synchronous analysis [9] to improve the estimation performance of the spectral envelope. Furthermore, aperiodicity is used as the parameter to represent not only the periodic signal but also the aperiodic signal.

2.2. Problems with aperiodicity

Since the human voice does not have perfect periodicity, high-quality voice synthesis systems generally use the mixed excitation signal to represent the aperiodic signal. In STRAIGHT and TANDEM-STRAIGHT, the aperiodicity is defined as the spectrum to synthesize both periodic and aperiodic signals. The periodic and the aperiodic spectra are calculated using the spectral envelope and aperiodicity, and the periodic and aperiodic signals are individually calculated.

This approach cannot represent the phase of the input voice because the periodic signal is calculated as the minimum phase response, and the vocal tract response $h(t)$ generally includes not only minimum phase response but also maximum phase response. To accurately synthesize a voice, it is essential to extract the phase of the input voice. We used a waveform-based parameter as a new parameter instead of aperiodicity.

3. Proposed method and voice synthesis system

3.1. Algorithm

The proposed method, named PLATINUM (PLATform INference by removing Underlying Material), is used to synthesize natural voice. PLATINUM extracts the waveform-based parameter to reconstruct the input voice. Figure 1 illustrates the overview of the proposed system that uses the excitation signal. The proposed system equals vocoder-based systems except that it uses the excitation signal instead of aperiodicity, which therefore suggests that it is possible for the proposed system to independently manipulate the F_0 and spectral envelope like vocoder-based systems.

The observed spectrum $Y(\omega)$ is defined as the product of the spectral envelope $H(\omega)$ and target spectrum $X(\omega)$ for reconstructing the waveform. The target spectrum $X(\omega)$ is given by

*e-mail: morise@fc.ritsumei.ac.jp

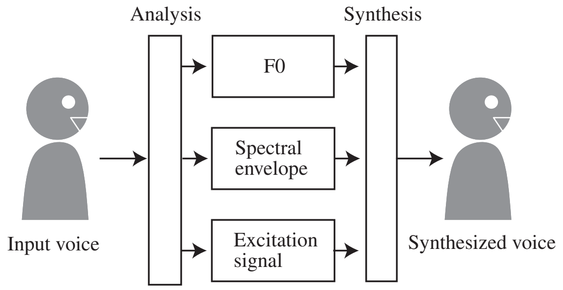

Fig. 1 Overview of the proposed system.

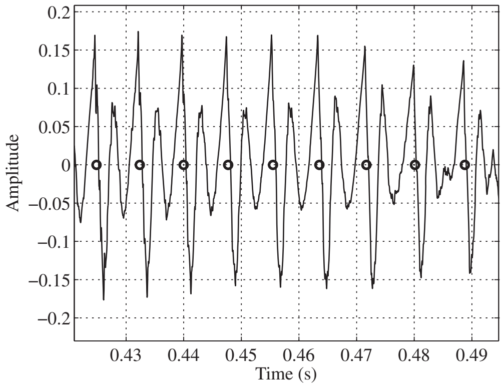

Fig. 2 Waveform (line) and temporal positions marked using PLATINUM (circles).

$$X(\omega) = \frac{Y(\omega)}{H(\omega)}. \tag{4}$$

Since the phase of $H(\omega)$ for vocoder-based systems is generally the minimum phase, the maximum phase of the input voice is included in $X(\omega)$. The power of $X(\omega)$ is nearly flat, provided that the spectral envelope is accurately estimated. If $H(\omega)$ does not include any zeros, the inverse spectrum $1/H(\omega)$ can be calculated reliably.

3.2. Pitch marking

To estimate $X(\omega)$, determining the temporal positions for windowing is an important problem. PLATINUM uses the F_0 contour and waveform. First, the voiced section is estimated based on the F_0 contour, and the temporal position with maximum value of $y(t)^2$ is then extracted as the basic temporal position. The other positions are automatically calculated based on the basic position and F_0 contour.

Figure 2 illustrates a waveform (line) and the calculated positions by using PLATINUM (circles). Similar to TD-PSOLA, the waveform is windowed by a Hanning window with a length of $2T_0$. The spectrum $Y(\omega)$ in Eq. (4) is calculated using the windowed waveform and used to calculate $X(\omega)$ of each frame.

4. Evaluation and discussion

To verify the effectiveness of the proposed system, a MUSHRA-based evaluation [10] was carried out. The STRAIGHT (STR), TANDEM-STRAIGHT (TAN), Cepstrum-based (CEP) systems, and the proposed system (PROP) were used for evaluation. The proposed system used DIO [11] as the F_0 estimation method and STAR [12] as the spectral envelope estimation method. The spectral envelope estimated using STAR does not contain any zeros [12].

In this experiment, not only the synthesized voice but also the F_0-modified voices ($F_0 \pm 25\%$) and the spectral-stretched voices ($\pm 15\%$) were used for demonstrating the robustness against their modification. The voices used for the evaluation were of three males and three females. The sampling was 44,100 Hz/16 bit, and a 32-dB (A weighted) room was used. Five subjects with normal hearing ability participated in the evaluation.

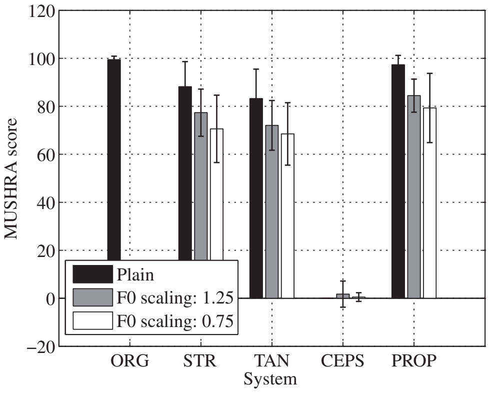

Fig. 3 Experimental results. Plain and F_0-modified voices are plotted.

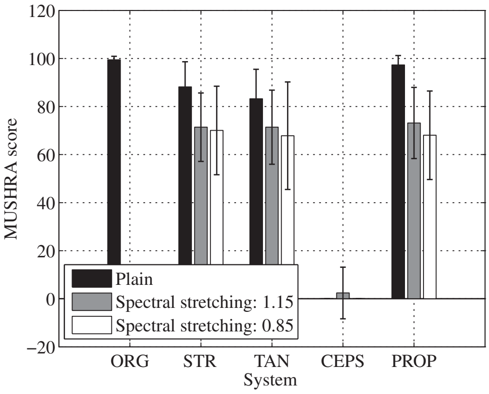

Fig. 4 Experimental results. Plain and spectral-envelope-modified voices are plotted.

Figures 3 and 4 illustrate the evaluation results. The vertical axis represents the MUSHRA score, which equals the sound quality. Results showed the proposed system could synthesize voice more naturally then other systems, and the sound quality of the F_0-modified voices was higher than that with the other systems. However, the sound quality of spectral-stretched voices based on the proposed system was the same as that with STRAIGHT and TANDEM-STRAIGHT.

The proposed system is the same as TD-PSOLA, provided that voice is synthesized without spectral modification. On the other hand, it can modify the spectral envelope and synthesize natural voice. These results suggest that the proposed system has advantages over both STRAIGHT and TD-PSOLA.

5. Concluding remarks

We proposed a voice synthesis system with our excitation signal extraction method for high-quality voice synthesis. The excitation signal extraction method improves the sound quality of synthesized voice. Subjective evaluation of sound quality was carried out by comparing the proposed system with conventional systems. Voice synthesized with the proposed system is superior to the conventional systems in sound quality. We also confirmed that the proposed system is robust against F_0 and spectral envelope modification.

Modern voice conversion techniques such as voice morphing [13] are a crucial topic; therefore, voice morphing with the proposed system will be examined in future work.

Acknowledgment

This research was partly supported by Ono Acoustics Research Fund and Grant-in-Aid for Scientific Research funded by MEXT, Japan.

References

[1] H. Kenmochi and H. Ohshita, "VOCALOID — Commercial singing synthesizer based on sample concatenation," *Proc. Interspeech2007*, pp. 4009–4010 (2007).

[2] C. Hamon, E. Moulines and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," *Proc. ICASSP89*, Vol. 1, pp. 238–241 (1989).

[3] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, **11**, 169–177 (1939).

[4] W. Hess, *Pitch Determination of Speech Signals* (Springer-Verlag, Berlin, 1983).

[5] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal pitch detection," *J. Acoust. Soc. Am.*, **36**, 269–302 (1964).

[6] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, **50**, 637–655 (1971).

[7] H. Kawahara, I. Masuda-Katsuse and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, **27**, 187–207 (1999).

[8] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," *Proc. ICASSP2008*, pp. 3933–3936 (2008).

[9] M. V. Mathews, J. E. Miller and E. E. David, "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Am.*, **33**, 179–185 (1961).

[10] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems" (2003).

[11] M. Morise, H. Kawahara and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," *Proc. AES 35th Int. Conf.*, CD-ROM (2009).

[12] M. Morise, T. Matsubara, K. Nakano and T. Nishiura, "A rapid spectrum envelope estimation technique of vowel for high-quality speech synthesis," *Trans. IEICE Jpn.*, **J94-D**, 1079–1987 (2011) (in Japanese).

[13] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP2009*, pp. 3905–3908 (2009).