INTERSPEECH 2009 BRIGHTON

Speech synthesis based on the plural unit selection and fusion method using FWF model

Ryo Morinaka, Masatsune Tamura, Masahiro Morita, Takehiko Kagoshima

Corporate Research and Development Center, Toshiba Corporation, Japan

ryo.morinaka@toshiba.co.jp, masatsune.tamura@toshiba.co.jp,
masahiro.morita@toshiba.co.jp, takehiko.kagoshima@toshiba.co.jp

10.21437/Interspeech.2009-597

Abstract

For speech synthesizers, enhanced diversity and improved quality of synthesized speech are required. Speaker interpolation and voice conversion are the techniques that enhance diversity. The PUSF (plural unit selection and fusion) method, which we have proposed, generates synthesized waveforms using pitch-cycle waveforms. However, it is difficult to modify its spectral features while keeping naturalness of synthesized speech. In the present work, we investigated how best to represent speech waveforms. Firstly, we introduce a method that decomposes a pitch waveform in a voiced portion into a periodic component, which is excited by vocal sound source, and an aperiodic component, which is excited by noise source. Moreover, we introduce the FWF (formant waveform) model to represent the periodic component. Because the FWF model represents the pitch waveform in accordance with formant parameters, it can control the formant parameters independently. We realized a method that can easily be applied to the diversity-enhancing techniques in the PUSF-based method because this model is based on vocal tract features.

Index Terms: plural unit selection and fusion method, periodic/aperiodic component, FWF model, formant parameter

1. Introduction

For a speech synthesizer, two principal features are required. One is generation of natural synthesized speech. The other is enhancement of diversity by introducing techniques such as speaker interpolation [1] and voice conversion [2]. The PUSF method [3] can generate natural synthesized speech while keeping stability. This method is a hybrid method of a unit-training-based method [4] and a unit-selection-based method [5]. In addition, this method is characterized by selecting plural speech units and fusing them. The PUSF method, however, cannot easily modify spectral features of speech units because selected plural speech units are fused in the time domain.

The FWF model [6] represents a periodic component waveform in a voiced portion as the sum of formant waveforms. Each formant waveform is modeled by formant parameters such as formant frequency and its spectral shape. The FWF model can easily modify spectral features of speech units because formant parameters are controlled. The FWF model, however, cannot properly represent aperiodic components. Threfore, firstly, we refer to a method that decomposes pitch-cycle waveforms into periodic and aperiodic components and properly fuses them in the time domain. In this paper, we call this method the Per/Aper (Periodic/Aperiodic) PUSF method. Then, we propose a method that introduces the FWF model to the Per/Aper PUSF method.

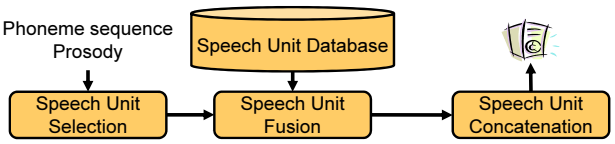

Figure 1: Diagram of the PUSF method

The remainder of this paper is organized as follows. The PUSF method is described in Section 2, the Per/Aper PUSF method is described in Section 3, a method that introduces the FWF model to the Per/Aper PUSF method is presented in Section 4, experimental result is reported and discussed in Section 5, and a conclusion is presented in Section 6.

2. The PUSF method

Figure 1 shows the diagram of the PUSF method. It consists of a speech unit selection process, a speech unit fusion process and a speech unit concatenation process [3]. First, a phoneme sequence and prosody (duration, f_0 contour) information are input to the speech unit selection process. Next, in the speech unit selection process, plural speech units are selected. Speech units are selected that are suited to the phoneme sequence and the prosody information by using a cost function for each processing speech unit. Then, in the speech unit fusion process, a fused speech unit is generated by fusing selected plural speech units. Finally, in the speech unit concatenation process, synthesized speech is generated by concatenating fused speech units. In this paper, the word "fusion" means generating waveform which represents the selected plural speech units.

The cost function, in the speech unit selection process, consists of a target cost and a concatenation cost. The target cost is defined by the weighted sum of f_0 target cost, duration target cost, and phonetic context cost. The concatenation cost is defined by the weighted sum of f_0 concatenation cost, spectrum concatenation cost, power concatenation cost, and adjacency cost (set to 0 when two consecutive units are adjoining in the speech unit database, otherwise 1).

In the speech unit fusion process, synthesized speech of the unvoiced portion is not generated by fusing speech units, but synthesized speech of the voiced portion is generated by fusing speech units. Synthesized speech units in the unvoiced portion are used for the speech unit that has the minimum cost calculated in the speech unit selection process. Selected plural speech units of the voiced portion are divided into pitch-cycle waveforms and fused by the respective pitch-cycle waveforms. The pitch-cycle waveform is the waveform whose spectrum shape represents the spectrum envelope of speech signals. The fused

pitch-cycle waveform is generated by decomposing into several bands, averaging in each band, and adding the bands averaged waveforms.

3. The Per/Aper PUSF method

In the FWF model, periodic components are represented properly, but aperiodic components are not represented properly. Therefore, a speech unit waveform is decomposed into the two components in order to introduce the FWF method to the PUSF-based method. Firstly, we refer to a method that decomposes pitch-cycle waveform of the speech unit into periodic and aperiodic components, and properly fuses them. In this paper, we call this method the Per/Aper PUSF method. This method can improve the quality of synthesized speech because synthesized speech is generated based on an acoustic source.

3.1. Decomposing periodic and aperiodic components

For decomposing the pitch-cycle waveform, we use the PSHF (Pitch-Scaled Harmonic Filter) method [7]. The PSHF method can decompose the pitch-cycle waveform, which is composed of mixed periodic and aperiodic components in full band, into the two kinds of components by utilizing the harmonious structure of the spectrum of a periodic component. The decomposing performance of the PSHF method is satisfactory except in the case of rapid variation of pitch.

3.2. Fusion of aperiodic components

In the Per/Aper PUSF method, fused aperiodic components are generated from two features to avoid attenuation of aperiodic components. One feature concerns the acoustic source, the other feature concerns the vocal tract filter. These features are extracted from aperiodic components of pitch-cycle waveforms for selected plural speech units as follows:

Vocal tract filter feature extraction step:

1. Concatenate temporally aperiodic components.
2. Extract fused LPC coefficients from concatenated aperiodic components by LPC analysis.

Acoustic source feature extraction step:

1. Extract LPC coefficients from each aperiodic component by LPC analysis.
2. Extract linear prediction residual from LPC coefficients.
3. Extract power envelope from each linear prediction residual.
4. Generate fused power envelope by averaging extracted power envelopes with phase alignment.

Fused aperiodic component is generated from the above two features as follows:

1. Excite white noise for each pitch-cycle waveform.
2. Generate waveform of acoustic source by modulating amplitude according to the fused power envelope.
3. Generate fused aperiodic component by filtering with the fused LPC coefficients.

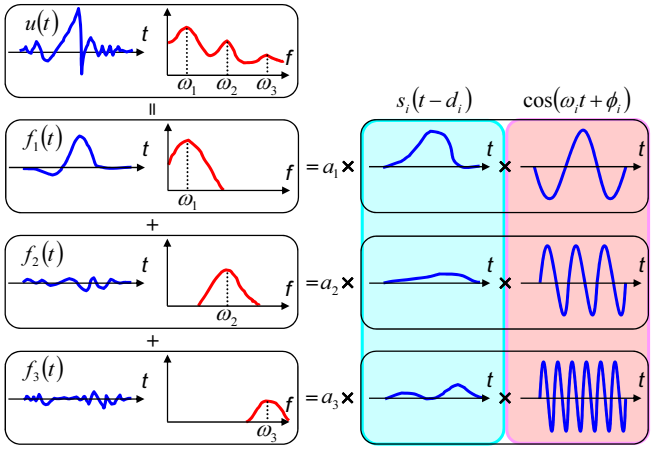

Figure 2: Diagram of the FWF model

3.3. Fusion of periodic components

In the Per/Aper PUSF method, fused periodic components are generated in the same way as with the speech unit fusion process in the PUSF method. The fused pitch-cycle waveform is generated by adding the fused periodic component and the fused aperiodic component.

4. The FWF model parameter fusion method

In the Per/Aper PUSF method, the spectral features can no longer be modified easily because of fusion of periodic components in the time domain. To overcome this problem, we propose a method that introduces the FWF model.

4.1. FWF model

Assume that a voiced portion of synthesized speech is generated by the pitch synchronous overlap addition process of a sequence of pitch-cycle waveforms composing a speech unit. The pitch-cycle waveform is represented by the FWF model defined below. A pitch-cycle waveform $u(t)$ is composed of sum of formant waveforms $f_i(t)$ as follows:

$$u(t) = \sum_{i=1}^{N_f} f_i(t) \tag{1}$$

where N_f is the number of formants. Each of the formant waveforms $f_i(t)$ has one formant. It is defined by windowed cosine waveform as follows:

$$f_i(t) = a_i s_i(t - d_i) \cos(\omega_i t + \phi_i) \tag{2}$$

where a_i, $s_i(t)$, d_i, ω_i and ϕ_i are a formant amplitude, a window function, a position of the window function, a formant frequency and a formant phase, respectively. By this modeling, frequency, amplitude, phase and spectral shape of each formant can be controlled independently. The spectral shape is represented by the spectrum of the window function. Figure 2 shows the diagram of the FWF model in the case of $N_f = 3$. In the figure, blue line shows waveform in the time domain and red line shows logarithmic power spectrum in the frequency domain. In the above formulation, the window function is generalized by arbitrary function $s_i(t)$. The freedom of this model is too great

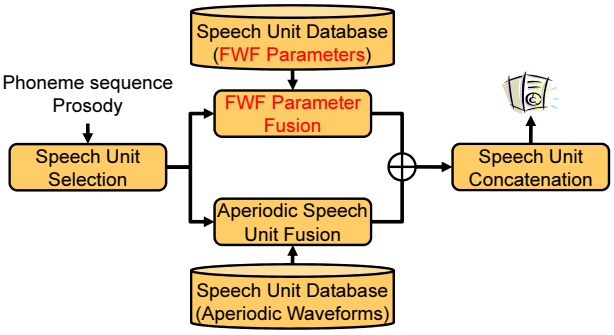

Figure 3: Diagram of the proposed method

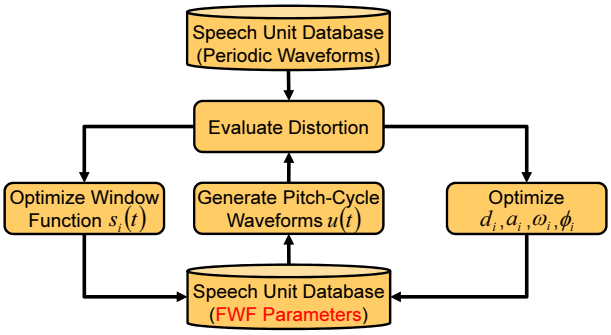

Figure 4: Diagram of the extracting FWF model parameter

for robust estimation of the parameters. It is reasonable to reduce the freedom by limiting bandwidth of the window function because its spectral shape represents a formant. Therefore, we introduced the constraint that the window function is represented by the weighted sum of the bases as follows:

$$s_i(t) = \sum_{j=1}^{N_b} w_{ij} b_j(t) \tag{3}$$

where $b_j(t)$, w_{ij} and N_b are the bases of window functions, weight of each basis and the number of bases, respectively. The basis function is defined by product of Hanning window and DCT basis as follows:

$$b_j(t) = \frac{1}{2}\left(1 - \cos\frac{2\pi t}{L_w}\right)\cos\frac{(j-1)\pi t}{L_w}, \tag{4}$$

where a constant L_w indicates the length of the window function. The aim of this formulation is that the spectrum of each basis has a single narrow peak and small sidelobes.

4.2. Basic architecture

The FWF model cannot represent aperiodic components appropriately, but can represent periodic components of pitch-cycle waveforms. Therefore, we propose a method that generates fused periodic components from fused FWF parameters. Aperiodic components are generated in the same way as with the Per/Aper PUSF method. Figure 3 shows a diagram of the proposed method.

4.3. Extracting FWF model parameters

The proposed method requires the FWF model parameters for the periodic component in the voiced portion. The FWF model parameters are extracted from periodic component of the pitch-cycle waveform. When extracting the FWF model parameters, it is assumed that the window function $s_i(t)$, which represents the detailed spectral shape of each formant, is common to all training data for the the same formant of the same phoneme, while the formant frequency ω_i varies among training data samples. The window function $s_i(t)$ is culculated under this assumption, and therefore a single function is extracted for each formant of each phoneme. The other parameters of the FWF model are extracted for every training data sample. For each pitch-cycle waveform, the FWF model parameters are extracted as follows:

1. Optimize window function $s_i(t)$ and reconstruct pitch-cycle waveform $u(t)$ using equation (1).
2. Evaluate distortion between reconstructed pitch-cycle waveform $u(t)$ and pitch-cycle waveforms $r(t)$ of training data.
3. Optimize the other FWF parameters, a_i, d_i, ω_i, and ϕ_i, and reconstruct pitch-cycle waveform $u(t)$ using equation (1).
4. Evaluate distortion between reconstructed pitch-cycle waveform $u(t)$ and pitch-cycle waveforms $r(t)$ of training data.

The above procedure is repeated until the distortion is converged. Square error of waveforms is used as the distortion criterion in the above procedure [8].

4.4. Fusion of FWF parameters

Fused FWF parameters are generated from FWF parameters corresponding with pitch-cycle waveforms of selected plural speech units. Assuming the number of selected speech units is N, each component of fused FWF parameters in i-th formant is generated as follows:

$$a_i^{\mathcal{F}} = \sum_{n=1}^{N} \frac{a_i^n}{N} \qquad d_i^{\mathcal{F}} = \sum_{n=1}^{N} \frac{d_i^n}{N}$$
$$\omega_i^{\mathcal{F}} = \sum_{n=1}^{N} \frac{\omega_i^n}{N} \qquad \phi_i^{\mathcal{F}} = \tan^{-1}\frac{\sum_{n=1}^{N} a_i^n \sin\phi_i^n}{\sum_{n=1}^{N} a_i^n \cos\phi_i^n} \tag{5}$$

where $a_i^{\mathcal{F}}$, $d_i^{\mathcal{F}}$, $\omega_i^{\mathcal{F}}$ and $\phi_i^{\mathcal{F}}$ are a fused formant amplitude, a fused portion of the window function, a fused formant frequency and a fused formant phase, respectively. And, a_i^n, d_i^n, ω_i^n and ϕ_i^n are a formant amplitude, a portion of the window function, a formant frequency and a formant phase of the n-th FWF parameter, respectively. Fused window function $s_i^{\mathcal{F}}(t)$ is generated by averaging with phase alignment between window functions. Figure 5 shows the time-series variation of $\omega_i^{\mathcal{F}}$ ($i = 1, 2, 3$) in the input Japanese sentence "/H-o-o-s-o-o-ch-u-u-n-o/", meaning "on-air" in English. And it shows ω_i^1, ω_i^2, ω_i^3 when the number of plural units N is 3. A vertical line in Figure 5 indicates boundaries between phonemes.

5. Experiment

To evaluate the proposed method, we implemented a prototype system and conducted MOS (Mean Opinion Score) evaluation.

5.1. Experimental conditions

We used half phones as speech units, and the half phone balanced speech database for a female speaker and a male speaker

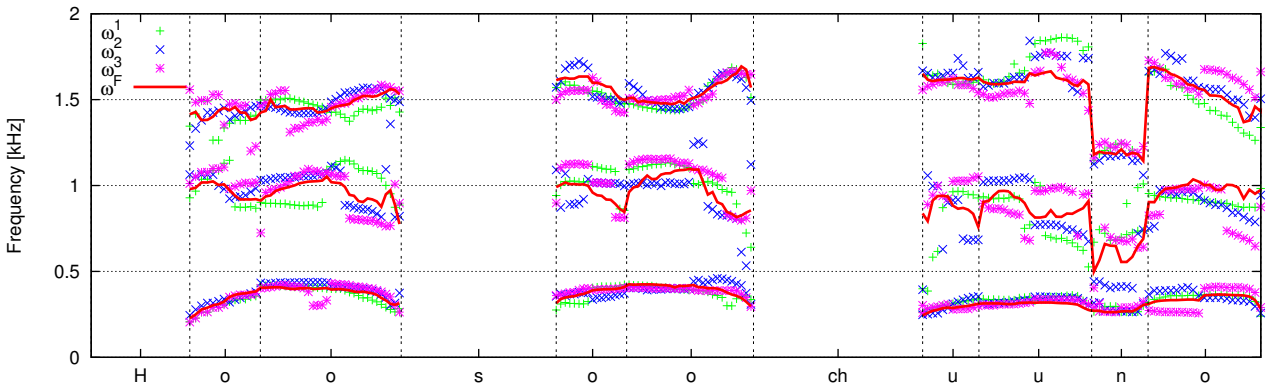

Figure 5: Time-series variation of the $\omega_i^{\mathcal{F}}$

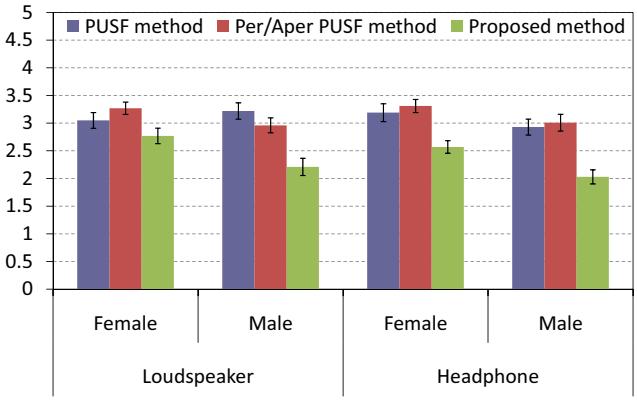

Figure 6: Result of MOS evaluation

is used. The MOS evaluation was conducted through a loudspeaker and a headphone. For each MOS evaluation, 10 subjects evaluated 10 Japanese utterances. 5 utterances were from car-navigation sentences. The rest were from news articles. For each subject, the utterances were shuffled and played in a different order so that each subject eventually evaluated 10 utterances, and was given a 5-point MOS value. Synthesized speech in the case of the PUSF method, the Per/Aper PUSF method and the proposed method was generated from the same prosody information. The number of plural speech units N is 10 and the sampling frequency f_s is 22.05 kHz.

5.2. Experimental result

Figure 6 shows the MOS evaluation result. The bars in the figure show two-sided 95% confidence intervals. The result for a loudspeaker is almost the same the result for a headphone. From Figure 6, the performance of the Per/Aper PUSF method was higher than that of the PUSF method in most conditions except in the case of the male speaker. The unsatisfactory decomposing performance for the male speaker is mainly attributable to the leaking of periodic components to aperiodic components, and consequently, aperiodic components are not fused properly. On the contrary, the performance of the proposed method was worse than that of the PUSF method and the Per/Aper PUSF method in all conditions, owing to the low accuracy in the extracting of the FWF model parameters. Window function $s_i(t)$ of the FWF model parameters is extracted by setting it to be common for all training data in same phoneme. Therefore, the spectrum of $u(t)$, which was reconstructed from the extracted FWF model parameters using equation (1), cannot accurately represent the spectrum of the pitch-cycle waveform of training data. Therefore, the result for the proposed method was inferior to those for the other methods.

6. Conclusions

We proposed a method introducing the FWF model to represent the pitch-cycle waveform for enhancing diversity of a speech synthesizer. The experimental result does not show the effectiveness of the proposed method in terms of the quality of the synthesized speech. However, we achieved enhanced diversity in the case of the PUSF-based method by introducing the FWF model, because formant parameters can be controlled independently in the case of the FWF model. One of the subjects for future work is the realization of diversity-enhancing techniques such as speaker interpolation and voice conversion in a method based on the proposed method. Since the FWF model is highly flexible for diversity-enhancing techniques, we can easily apply them to the proposed method. Other subjects for future work are improvement of the accuracy of extracting the FWF parameters and of decomposing the pitch-cycle waveform.

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," Acoustical Society of Japan, Vol.21, No.4, pp.199-206, 2000.

[2] M. Tamura and T. Kagoshima, "Voice conversion for plural unit selection and fusion method," 1-4-13, Proc. 2006 Spring Meeting of ASJ, pp.237-238, 2006 (in Japanese).

[3] T. Mizutani and T. Kagoshima, "Concatenative speech synthesis based on the plural unit selection and fusion method," IEICE, Vol.E88-D, No.11, pp.2565-2572, 2005.

[4] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS)," Proc. ICSLP'98, pp.1927-1930, Dec. 1998.

[5] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," Proc. ICASSP 2002, pp.465-468, 2002.

[6] T. Kagoshima and M. Akamine, "Modeling of pitch-cycle waveforms for controlling spectrum envelope," 1-6-7, Proc. 2003 Spring Meeting of ASJ, pp.235-236, 2003 (in Japanese).

[7] P. Jackson and C. H. Shadle, "Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech," *IEEE* Trans. Speech and Audio Processing, vol.9 pp.713-726, Oct. 2001.

[8] T. Kagoshima and M. Akamine, "Closed loop training of pitch-cycle waveforms using FWF model," 1-8-12, Proc. 2003 Autumn Meeting of ASJ, pp.205-206, 2003 (in Japanese).