Composite Wavelet Model for Stability-Oriented Speech Synthesis from Cepstral Features

Junya Koguchi† and Shigeki Sagayama†
† Meiji University, Tokyo, Japan
E-mail: {ev50552, sagayama}@meiji.ac.jp

***Abstract*—This paper discusses a stability-oriented vocoder based on Gabor wavelet approximation of the source signal for statistical speech synthesis. In conventional vocoders with recursive filters, the filter gain characteristics often cause degradations in the sound quality due to unstable behavior of recursive filters affected by sharp resonances driven by a particular overtone in the excitation signal. To cope with this problem, we have proposed Composite Wavelet Model (CWM) to avoid filter-caused problems and have made several improvements as a vocoder. Based on non-recursive filters, it enables synthesizing stable speech which is robust to changes in F_0 parameter. In this paper, we further discuss the optimal number of mixture components to improve the synthetic speech quality to determine them through subjective experimental evaluations and report them on the result of incorporating in an HMM-based speech synthesis system. Objective experimental evaluations confirmed the improved stability in the amplitude of the synthetic speech.**

I. Introduction

In the statistical approach toward text-to-speech synthesis, the waveform generation part, which is often referred to "vocoder," plays an important role as well as acoustic models. The vocoder generates speech waveforms from acoustic features obtained through statistical training. It is desirable to have vocoders synthesize high quality voices from the speech parameters and be robust to artificial changes in the parameters, so that TTS systems with vocoder can synthesize stable speech under the parameters not included in the training data, and that users can process the synthetic speech according to their preference. For these reasons, although new approaches to speech synthesis called "end-to-end models" which generates a speech waveform directly from an input text by training the interrelationship between linguistic features and speech waveforms have been proposed [1], a speech synthesis system using a vocoder is still useful.

In statistical speech synthesis, Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Log Spectrum Approximation (MLSA) filters [2] are often used as acoustic features for statistical training and waveform synthesis filter from MFCCs. MLSA filter approximates logarithmic amplitude spectra and synthesizes a speech from MFCCs and F_0 parameters using recursive filters. However, the use of the recursive filters in speech synthesis may suffer from unexpected large amplitudes of the generated waveform in the case of sharp resonances in the spectral envelope lying on one of the overtones of the excitation source and cause near-oscillation behaviors. It was shown in a previous study by Hamada *et al.* [3] that generating a signal waveform from a power spectrum without using recursive filters is effective as a means to solve this problem.

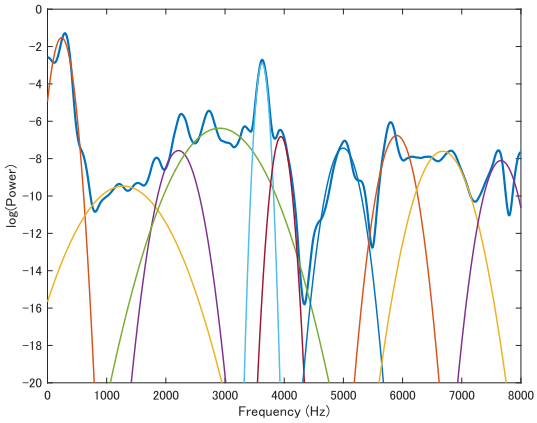

Fig. 1. GMM approximation of spectral envelope at phoneme /i/ (K=10).

On the other hand, we have proposed composite wavelet model (CWM) [4] as an alternative of the vocoder which is able to synthesize stable speeches and have utilized it in speech synthesis. The speech waveform using CWM can be regarded as convolution by non-recursive filters, its impulse responses are short and it is reported that quality degradation hardly occurs even for fluctuation of pitch. Hojo *et al.* [5] tried entirely replacing the MFCC parameters with CWM parameters both in the training and speech synthesis based on Hidden Markov Model (HMM) [6] from the same motivation. In this paper, we report the experimentally improved results by using the composite wavelet model against the problem caused by the gain characteristic of the recursive filter used in the conventional HMM speech synthesis.

II. Composite Wavelet Vocoder

A. Acoustic feature extraction

This section describes how to extract acoustic features from a speech signal and synthesize a waveform from the parameters [4, 5]. Firstly, CWM approximates a speech spectral envelope from the sum of the Gaussian distributions, interpreted as a function of frequency. This means each Gaussian distribution function roughly corresponds to a peak

in a spectral envelope [7]. CWM is thus convenient for describing both the frequency and power fluctuations of spectral peaks, because it is characterized by parameters. We use the distributions' means, variances and weights as the parameters (in the following, these are called "CWM parameters") which represent the spectrum. CWM approximates spectral envelopes by minimizing the *I*-divergence between spectral envelope and GMM sequentially using the auxiliary function approach (Fig. 1). The *I*-divergence is defined as follows

$$I[Y||F] = \sum_{\omega,t}[Y_{\omega,t}\log\frac{Y_{\omega,t}}{F_{\omega,t}} - Y_{\omega,t} + F_{\omega,t}], \quad (1)$$

where $Y_{\omega,t}, F_{\omega,t}$ denote measured and modeled spectral envelope, respectively. $F_{\omega,t}$ at time t is computed as follows.

$$F_{\omega,t} = \sum_{k=1}^{K} \frac{w_k}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(\omega-\mu_k)^2}{2\sigma_k^2}\right], \quad (2)$$

where K denotes the number of mixed Gaussians. Secondly, CWM concatenates $\mu_k, \sigma_k, w_k (k = 1, \ldots, K)$ and uses them as the time series of the spectral features. In addition to that, when GMM approximates the spectral envelope, the Gaussian functions can fit to a single harmonic structure component, which was observed in experiments. A solution is smoothing the spectrum. Saikachi *et al.* extracted spectral envelopes using a method of lag-windowed autocorrelation functions of speech waveforms and applying a Fourier transformation to them. In this paper, we obtained spectral envelopes from MFCCs by a HMM based TTS system. We initialize the values of μ_k for GMM estimation using the average of the spectrum pair obtained by $2K$ order LSP analysis. The frequency of the linear spectrum obtained from LSP is known to roughly correspond to the formant. It is expected that the convergence becomes faster compared to computation with randomly initialized values. For the variance σ we did experiments with values from 10 to 50 in increments of 10, and found the value of 10 to be optimal. This was chosen as the initial value.

B. Applying Time Transition Probability

In the extraction method mentioned in the previous subsection, extraction is performed independently for each frame. The index of each Gaussian distribution function in the CWM at one particular frame is not always consistent with that of another frame. That may cause the problem that the GMM fails to approximate some peaks of the spectrum which should be present, or to capture the smooth trajectories of the formants. Thus, we introduce the time transition probability of μ_k in the CWM parameter. When μ_k at time t is $\mu_k^{(t)}$, we assume that the time transition of $\mu_k^{(t)}$ follows the normal distribution whose mean is $\mu_k^{(t-1)}$ in the previous frame (3).

$$P(\mu_k^{(t)}|\mu_k^{(t-1)}) = \mathcal{N}(\mu_k^{(t)}; \mu_k^{(t-1)}, \nu_k^2) \quad (3)$$

The variance ν_k^2 of each index k represents the extent of allowed temporal fluctuation of μ_k. The introduction of this time transition probability smoothes the extracted time variation of

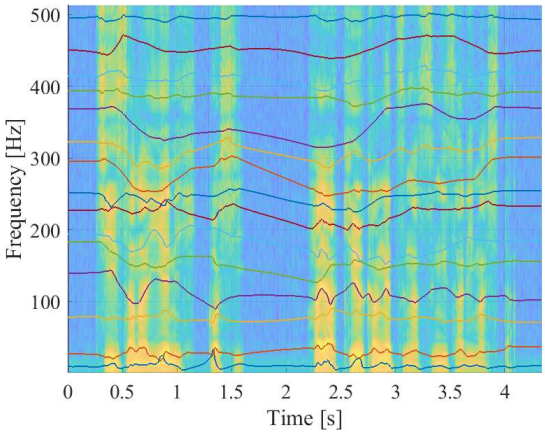

Fig. 2. The result of the extraction of μ_k without the time transition model (a Japanese sentence of A14 in the ATR-503 data set, *K*=10).

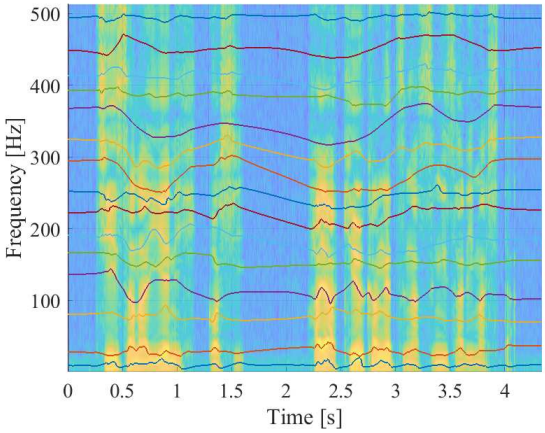

Fig. 3. The result of the extraction of μ_k with the time transition model (a Japanese sentence of A14 in the ATR-503 data set, *K*=10,).

the mean parameters. This smoothing preserves the continuities of the formant frequencies. As a result of introducing a transition probability model, the CWM extracts CWM parameters by minimizing the following objective function, which represents the difference between an observed spectrogram and a model spectrogram.

$$J[Y||F] = I[Y||F] + \sum_{k}^{K} \frac{1}{2\nu_k^2} \sum_{t=1}^{T-1} (\mu_k^{(t+1)} - \mu_k^{(t)})^2 \quad (4)$$

T is the total number of frames of the spectrum obtained from speech. In the same method as [8], we introduce an auxiliary function using Jensen's inequality. Now the parameters except for μ_k are updated with Eq. (1) in 2.1. $\boldsymbol{\mu_k} = (\mu_k^{(1)}, \mu_k^{(2)}, \cdots \mu_k^{(T)})^{\mathrm{T}}$, which is the time series vector of μ_k, is updated with the following equation considering the

time transition model.

$$\boldsymbol{\mu}_{\boldsymbol{k}}^{*} = \frac{1}{2}(D_k + E_k)^{-1} F_k, \tag{5}$$

where

$$D_k = \frac{1}{2\nu_k^2}\left\{D^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \tag{6}$$

$$D^{(i,j)} = \begin{cases} 1 & (i=j=\{1,T\}) \\ 2 & (i=j\in[2,T-1]) \\ -1 & (|i-j|=1) \\ 0 & (\text{other}), \end{cases} \tag{7}$$

$$E_k = \frac{1}{2\sigma_k^2}\left\{E_k^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \tag{8}$$

$$E_k^{(i,j)} = \begin{cases} \sum_\omega Y(\omega,i)\lambda_k(\omega,i) & (i=j), \\ 0 & (i\neq j), \end{cases} \tag{9}$$

$$F_k = \frac{1}{\sigma_k^2}(F_k^{(i)})_i, \quad (i \in [1,T]), \tag{10}$$

$$F_k^{(i)} = \sum_\omega \omega Y(\omega,i)\lambda_k(\omega,i), \tag{11}$$

where λ_k denotes an auxiliary variable in the auxiliary function. Fig. 2 shows the result of the extraction of $\boldsymbol{\mu_k}$ without the time transition model. Fig. 3 shows the result of the extraction of $\boldsymbol{\mu_k}$ with the time transition model. The speech sample used for extraction, the number of mixed Gaussians in the GMM, as well as the initial values are the same. ν_k is set so that the standard deviation of the time variation of μ_k is almost constant on the mel-frequency axis. It can be confirmed that the crossing of the index, which was often observed when the time transition model is not introduced, is continuously fluctuating due to smoothing of the time fluctuations of μ_k.

C. Waveform Synthesis from CWM Parameters

We describe a method of synthesizing speech waveforms using CWM parameters. As a method to solve the problem of the recursive filter mentioned in Chapter I, The FIR type filter obtained from inverse Fourier transform of GMM envelope approximation is used. Note that the inverse Fourier transform of the Gaussian function is the Gabor function, which is the product of the Gaussian function and the trigonometric function, as follows.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\omega-\mu)^2}{2\sigma^2}\right] \Leftrightarrow \frac{1}{\sqrt{2\pi^2}}\exp\left[-\frac{\sigma^2 t^2}{2} + j\mu t\right], \tag{12}$$

where ω, μ and σ can be regarded as frequency, peak frequency and Q-factor respectively mentioned in Chapter II.A.

By using this property, we apply inverse Fourier transformation of GMM and obtain the fundamental waveform of the Gabor wavelet. In the voiced segments, the speech waveform is obtained by arranging the fundamental waveform of the Gabor wavelet at intervals corresponding to the fundamental frequency in the time domain, which is equivalent to driving the FIR filter with an impulse train corresponding to the fundamental frequency. The synthesis of the unvoiced section, considering its aperiodicity of the waveform, is realized by arranging the fundamental waveforms at random intervals.

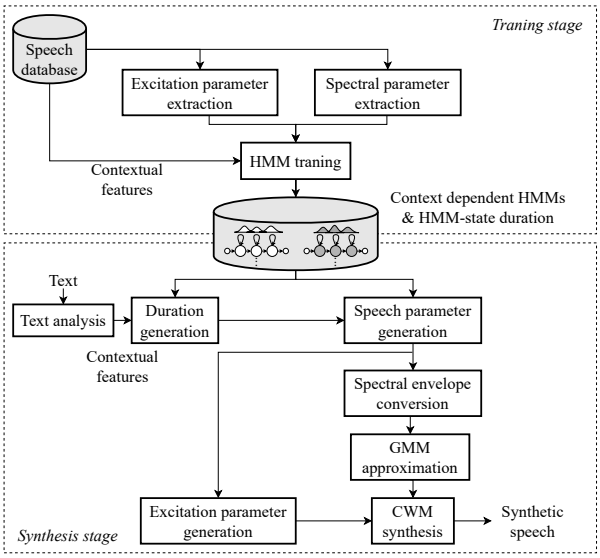

Fig. 4. An overview of TTS system based on HMM and CWM.

In CWM, since the speech waveform is synthesized directly from the spectral envelope not using recursive filters, it can be expected that the problem caused by the gain characteristic of the recursive filter does not occur. In our previous study [5], CWM was used both for generating a series of the spectral envelope and for waveform synthesis. In this paper, we focus on its aspect of waveform generation and combined it with HMM-based speech synthesis as a generator of cepstral features.

III. Speech Synthesis by HMM and CWM

A. Outline of the CWM speech synthesis

Combining the CWM vocoder with the existing HMM-based TTS system (Fig. 4), the proposed procedure is outlined as follows.

1) Give an input text to a HMM-based TTS system (e.g., HTS (HMM-based text-to-speech system) by Zen $et al.$) [9] to produce a mel-cepstrum vector sequence along with the F_0 trajectory. to produce a mel-cepstrum vector sequence along with the F0 trajectory
2) Convert the obtained mel-cepstrum coefficients into the linear spectra.
3) Generate the waveform from the linear spectra by the method of section II.

These steps are stated in more details below.

B. Generation of cepstral feature and F_0

For the input text, HTS generates a sequence of cepstral coefficient vectors along with fundamental frequencies. In the waveform synthesis part, HTS originally uses the MLSA filter. whereas CWM is used in this research instead.

C. Conversion from MFCCs to spectrum

The generated MFCCs are converted into spectrum as follows.

$$H(\omega) = s_{\gamma}^{-1}(\sum_{m=0}^{M} c_{\gamma}(\tilde{m})e^{-j\tilde{\omega}t}) \tag{13}$$

$$s_{\gamma}^{-1}(\omega) = \begin{cases} (1+\gamma\omega)^{(}1/\gamma) & (0 < |\gamma| \leq 1) \\ \exp\omega & (\gamma = 0), \end{cases} \tag{14}$$

$$\tilde{\omega} = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{15}$$

,where $H(\omega)$ denotes the spectrum, ${s_{\gamma}}^{-1}$ the inverse function of the generalized logarithmic function, c_{γ} the mel-frequency cepstrum coefficient (MFCC) [10].

D. Speech waveform synthesis by CWM

The method of synthesizing speech waveforms using the CWM features and the fundamental frequency information is similar to the method mentioned at section II.

IV. Decision of the number of mixture components

In conventional research related to CWM, it has not been specifically mentioned how many Gaussian functions approximating the spectral envelope are required to synthesize the best quality speech. Therefore, we determined the number of Gaussian functions in a subjective evaluation experiment as follows.

A. Experimental conditions

To investigate the quality of synthesized speech, listening tests were conducted. The number of mixed Gaussian function was changed from 15 to 40 in increments of 5. We selected five sentences of 3–5 seconds of duration from the ATR speech database [11]. HTS was used to generate MFCC vector sequences and F0 trajectories under the HTS conditions: $\gamma = 1.0$, $\alpha = 0.55$, and the sampling frequency 16 kHz. Ten men and women in their teens to twenties participated in the experiment using their headphones or earphones that they regularly use. The experiment was conducted in a quiet room.

B. Result

Figure 5 shows that the sound quality is highest when using 25, 30, or 35 Gaussian functions in the mixture. As the number of parameters becomes smaller, the time taken to update the GMM becomes shorter, so we conclude that 25 is the most suitable number of Gaussian functions for CWM.

V. Objective Experimental Evaluation

We investigated the gain characteristics (amplitude deviations) of synthetic speech to investigate whether the speech synthesis method by the CWM is effective or not in accordance with the experiment of [4]. In the comparison, characteristics of synthesized speech by MLSA filter were investigated.

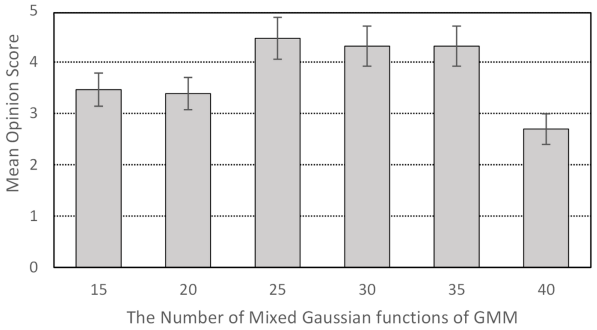

Fig. 5. An overview of TTS system based on HMM and CWM.

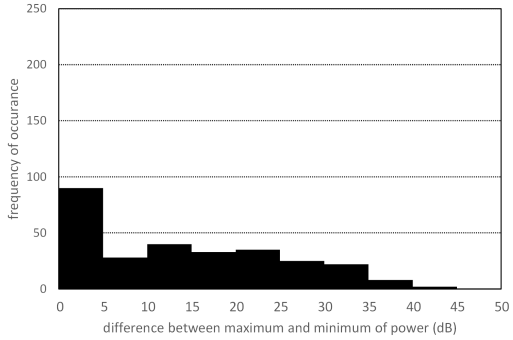

Fig. 6. Amplitude deviations in MLSA filter synthesis.

A. Experimental conditions

To investigate changes in gain with changes in F_0, the F_0 parameters were modified from 0.8 to 1.2 times of the original with the interval of 0.05 under the same conditions as Section IV. We approximated the spectra by GMM with the number of mixture components = 25.

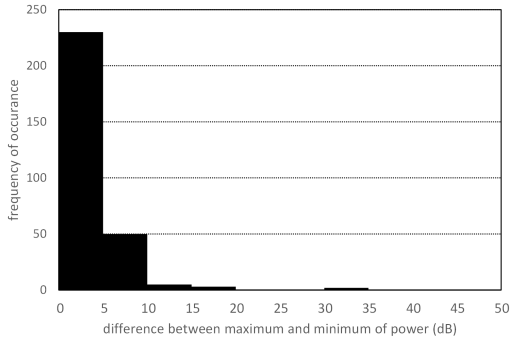

Fig. 7. Amplitude deviations in CWM synthesis.

B. Result

Since text-to-speech synthesis produces entirely new speech that lacks the original reference speech to compare with,

as in analytical synthesis [4], we explored the deviation of amplitude caused by the change in fundamental frequency of voiced sounds using the difference between the maximum value and the minimum value of the gain of the frame.

The results of comparing the amplitude characteristics are shown in the histograms in Figures 6, 7. The horizontal axis of the histograms represents the gain stability (*i.e.* amplitude stability). From the figures, it is suggested that the speech synthesis using the CWM is more stabilized in amplitude control than the method using a MLSA recursive filter.

VI. Conclusions

In this research, we proposed the use of CWM speech synthesis from cepstral features in the HMM-based text-to-speech system. We converted the cepstral features generated by HMM into spectrum and approximated it with GMM, *i.e.* sums of Gaussian functions. Based on the mixture weights, means and variances of the Gaussian functions obtained from the approximation, the composite Gabor wavelets were periodically and aperiodically concatenated to generate synthetic speech. The subjective experimental result showed that the CWM synthesized speech when the spectral envelope obtained from MFCC was approximated by 25 Gaussian functions is the most suitable. The objective experimental results showed improved gain characteristics of synthetic speech. Future work includes incorporating deep neural networks for improving the synthetic speech quality and applying this technique to singing voice synthesis.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17H00749.

References

[1] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *https://arxiv.org/abs/1703.10135*, 2017.

[2] S. Imai, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *IECE Trans. Fundamentals (Japanese Edition), A*, vol. 66, no. 2, pp. 122–129, 1983 (in Japanese).

[3] Y. Hamada *et al.*, "Non-filter waveform generation from cepstrum using spectral phase reconstruction," *Proc. SSW*, pp. 28–32, 2016.

[4] T. Saikachi *et al.*, "Speech analysis and synthesis based on composite wavelet model," *IEICE technical report. Speech*, vol. 105, no. 370, pp. 1–6, 2005 (in Japanese).

[5] N. Hojo *et al.*, "Hmm speech synthesis using speech analysis based on composite wave model," *Reports of 2012 autumn meeting the Acoustical Society of Japan (CD-ROM)*, vol. 2012, pp. 2–2–7, 2012 (in Japanese).

[6] T. Masuko *et al.*, "1996 ieee international conference on acoustics, speech, and signal processing conference proceedings," vol. 1, pp. 389–392, 1996.

[7] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 1229–1232, 1996.

[8] H. Kameoka *et al.*, "Composite function model of spectral envelope and harmonic structure for speech analysis," *Reports of the 2015 autumn meeting the Acoustical Society of Japan*, no. 2-6-7, 2005.

[9] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pp. 294–299, 2007.

[10] K. Tokuda *et al.*, "Spectral estimation of speech by mel-generalized cepstral analysis," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 76, no. 2, pp. 30–43, 1993 (in Japanese).

[11] K. Takeda *et al.*, "A japanese speech database for various kinds of research purposes," *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 44, no. 10, pp. 747–754, 1988 (in Japanese).