Text-to-speech synthesizer based on combination of composite wavelet and hidden Markov models

Nobukatsu Hojo[1], *Kota Yoshizato*[1], *Hirokazu Kameoka*[1,2], *Daisuke Saito*[1], *Shigeki Sagayama*[1,*]

[1]Graduate School of Information Science and Technology, the University of Tokyo, Japan
[2] Communication Science Laboratories, NTT Corporation, Japan
{hojo, yoshizato, kameoka, dsaito, sagayama}@hil.t.u-tokyo.ac.jp

Abstract

This paper proposes a text-to-speech synthesis (TTS) system based on a combined model consisting of the Composite Wavelet Model (CWM) and the Hidden Markov Model (HMM). Conventional HMM-based TTS systems using cepstral features tend to produce over-smoothed spectra, which often result in muffled and buzzy synthesized speech. This is simply caused by the averaging of spectra associated with each phoneme during the learning process. To avoid the over-smoothing of generated spectra, we consider it important to focus on a different representation of the generative process of speech spectra. In particular, we choose to characterize speech spectra using the CWM, whose parameters correspond to the frequency, gain and peakiness of each underlying formant. This idea is motivated by our expectation that the averaging of these parameters would not lead directly to the over-smoothing of spectra, as opposed to the cepstral representations. To describe the entire generative process of a sequence of speech spectra, we combine the generative process of a formant trajectory using an HMM and the generative process of a speech spectrum using the CWM. A parameter learning algorithm for this combined model is derived based on an auxiliary function approach. We confirmed through experiments that our speech synthesis system was able to generate speech spectra with clear peaks and dips, which resulted in natural-sounding synthetic speech.
Index Terms: text-to-speech synthesis, hidden Markov model, composite wavelet model, formant, Gaussian mixture model, auxiliary function

1. Introduction

This paper proposes a new model for text-to-speech synthesis (TTS). One promising approach for TTS involves methods based on statistical models. In this approach, the first step is to formulate a generative model of a sequence of speech features. The second step is to train the parameters of the assumed generative model given a set of training data in a speech corpus. The third step is to produce the most likely sequence of speech features given a text input and transform it into a speech signal. With this approach, one key to success is that the assumed generative model reflects the nature of real speech well. To model the entire temporal evolution of speech features, a hidden Markov model (HMM) and its variants including the "trajectory HMM" have been introduced with notable success [1, 2, 3]. HMMs are roughly characterized by the structure of the state transition network (i.e., a transition matrix) and an output distribution assumption. In conventional HMM-based TTS systems, a Gaussian mixture emission distribution of a cepstral feature [1, 2] or a line spectrum pair (LSP) feature [4, 3] is typically used as the output distribution of each state. The aim of this paper is to seek an alternative to the conventional speech feature and the state output distribution. Namely, this paper is concerned with formulating a new model for the generative process of speech spectra and combining it with an HMM.

The Gaussian distribution assumption of a cepstral feature describes probabilistic fluctuations of spectra in the power direction, while that of an LSP feature describes probabilistic fluctuations of spectral peaks in the frequency direction. Since the frequencies and powers of peaks in a spectral envelope correspond to the resonance frequencies and gains of the vocal tract, they both vary continuously over time according to the physical movement of the vocal tract. In particular, resonance frequencies and gains both vary significantly in the boundary between one phoneme and another. To achieve higher quality speech synthesis, we consider it important to describe a generative model that takes account of the fluctuations of spectral peaks in both the frequency and power directions rather than a fluctuation in just one direction. Conventional HMM-based TTS systems using cepstral features tend to produce over-smoothed spectra, which often result in muffled and buzzy synthesized speech. This is simply caused by the averaging of observed log-spectra assigned to each state during the training process (Fig. 2 (a)). Although some attempts were made to emphasize the peaks and dips of generated spectra via post-processing [5], it is generally difficult to restore original peaks and dips once spectra are over-smoothed. By contrast, this paper proposes tackling this problem by introducing a different spectral representation called the Composite Wavelet Model (CWM) [6, 7, 8] as an alternative to cepstral and LSP representations, on which basis we formulate a new generative model for TTS.

2. Generative model of spectral sequence

2.1. Motivation

Taking the mean of cepstral features associated with the same phoneme amounts to taking the mean of the corresponding log spectra. Since resonance frequencies and gains of the vocal tract both vary continuously during the transition from one phoneme to another, if we simply take the mean of the log spectra associated with a particular phoneme, the spectral peaks and dips will be be blurred and indistinct (Fig. 2 (a)). By contrast, if we can appropriately treat the frequencies and powers of individual spectral peaks as speech features, we expect that the spectral blurring will not occur when taking the mean (Fig. 2 (b)). CWM approximates a speech spectral envelope from the

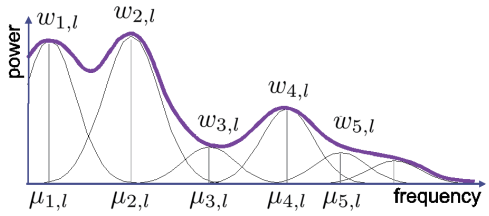
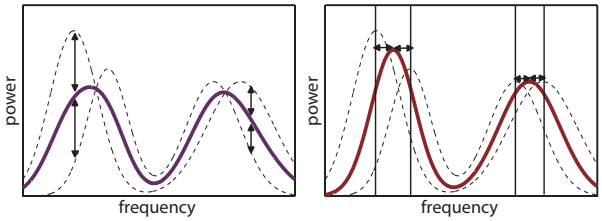


Figure 1: Spectral representation based on CWM



(a) Mean in cepstral domain (b) Mean in CWM domain

Figure 2: Expectations of spectra taken in different domains. The dashed lines indicate the training samples and the solid lines indicate the mean spectra.

sum of the Gaussian distributions, interpreted as a function of frequency (see Fig. 1 for a graphical illustration). This means each Gaussian distribution function roughly corresponds to a peak in a spectral envelope. CWM is thus convenient for describing both the frequency and power fluctuations of spectral peaks because it is characterized by parameters corresponding to the frequency and power of each spectral peak.

Another important feature of CWM is that the CWM parameters can be easily transformed into a signal. Since a Gaussian function in the frequency domain corresponds to a Gabor function in the time domain, CWM parameters can be directly converted into a signal by superimposing the corresponding Gabor functions. CWM can thus be regarded as a speech synthesizer with an FIR filter and its superiority to conventional systems with IIR filters has already been shown in [6].

One straightforward way of incorporating CWM into an HMM-based TTS system would be to first perform CWM parameter extraction on a frame-by-frame basis, and then train the parameters of an HMM by treating the extracted CWM parameters as a sequence of feature vectors. However, our preliminary investigation revealed that this simple method did not work well [9]. One reason for this is the common difficulty of formant tracking. Although formants and their time courses (formant trajectories) are clearly visible to the human eye in the spectrum and in the spectrogram, automatic formant extraction and tracking are far from trivial. Many formant extraction algorithms miss a formant that is present, insert a formant when there is none, or mislabel them (such as label F_1 as F_2, or F_3 as F_2), mostly as a result of incorrectly pruning the correct formant at the frame level. This is also the case for frame-by-frame CWM parameter extraction since it can be thought of as a sort of formant extraction. Fig. 3 shows an example of results obtained with frame-by-frame CWM analysis. As this example shows, the index of each Gaussian distribution function in the CWM at one particular frame is not always consistent with that at another frame. This was problematic when training the HMM parameters since the mean of the emission distribution of each state is given as the average of the feature vectors assigned to the same state. To train the HMM parameters appropriately, the indices of the Gaussian functions of the CWM assigned to the same state must always be consistent. This implies the need for the joint modeling of the CWM and HMM. Motivated by the above discussion, we here describe the entire generative process of a sequence of speech spectra by combining the generative process of a speech spectrum based on the CWM with the generative process of a CWM parameter trajectory based on an HMM.

2.2. Formulation

The CWM consists of parameters (roughly) corresponding to the frequency and power of spectral peaks. Specifically, the CWM approximates a spectral envelope by employing a Gaussian mixture model (GMM) interpreted as a function of frequency. Namely, the CWM $f_{\omega,l}$ is defined as

$$f_{\omega,l} = \sum_{k=1}^{K} \frac{w_{k,l}}{\sqrt{2\pi}\sigma_{k,l}} \exp\left(- \frac{(\omega - \mu_{k,l})^2}{2\sigma_{k,l}^2} \right), \tag{1}$$

where ω, l denotes the indices of frequency and time, respectively, and K is the number of mixture components of the GMM. See Fig. 1 for a graphical illustration. $\mu_{k,l}, \sigma_{k,l}$ and $w_{k,l}$ are the mean, variance and weight of each Gaussian (when interpreted as a probability density function), and thus correspond to the frequency, peakiness and power of the peaks in a spectral envelope, respectively.

Using the CWM representation given above as a basis, here we describe the generative process of an entire sequence of observed spectra. We consider an HMM that generates a set consisting of $\mu_{k,l}$, $\rho_{k,l} := 1/\sigma_{k,l}^2$, and $w_{k,l}$ at time l (see Fig. 4). Each state of the HMM represents a label indicating linguistic information, which can be simply either a phoneme label as shown in Fig. 4 or a context label (as used in [10]). Given state s_l at time l, we assume that the CWM parameters are generated according to

$$P(\mu_{k,l}|s_l) = \mathcal{N}(\mu_{k,l}; m_{k,s_l}, \eta_{k,s_l}^2), \tag{2}$$

$$P(\rho_{k,l}|s_l) = \mathrm{Gamma}(\rho_{k,l}; a_{k,s_l}^{(\rho)}, b_{k,s_l}^{(\rho)}), \tag{3}$$

$$P(w_{k,l}|s_l) = \mathrm{Gamma}(w_{k,l}; a_{k,s_l}^{(w)}, b_{k,s_l}^{(w)}), \tag{4}$$

where $\mathcal{N}(x; m, \eta^2)$ denotes the normal distribution and $\mathrm{Gamma}(x; a, b)$ the gamma distribution:

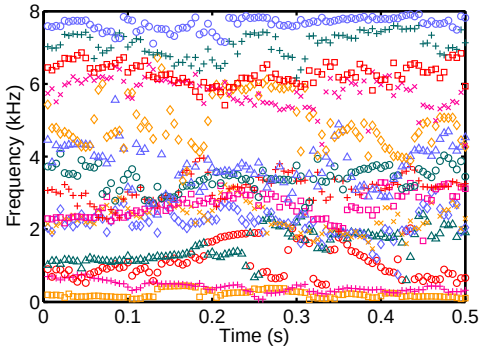
$$\mathrm{Gamma}(x; a, b) = x^{a-1} \frac{\exp(-x/b)}{\Gamma(a)\, b^a}. \tag{5}$$

These distribution families are chosen for simplifying the derivation of the parameter estimation algorithm described in the next section. Given the following sequence of CWM parameters, $\boldsymbol{\mu} = \{\mu_{k,l}\}_{k,l}, \boldsymbol{\rho} = \{\rho_{k,l}\}_{k,l}, \boldsymbol{w} = \{w_{k,l}\}_{k,l}$, we further assume that a spectrum $\{y_{\omega,l}\}_\omega$ observed at time l is generated according to

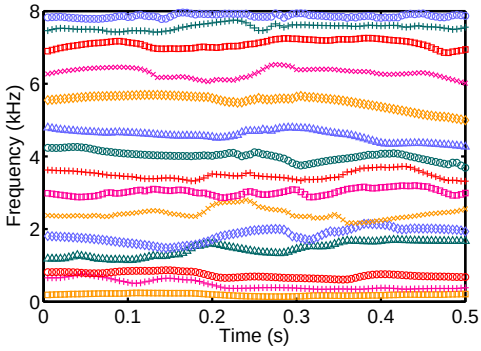
$$P(y_{\omega,l}|\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w}) = \mathrm{Poisson}(y_{\omega,l}; f_{\omega,l}), \tag{6}$$

where $\mathrm{Poisson}(x; \lambda)$ denotes the Poisson distribution:

$$\mathrm{Poisson}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{7}$$



(a) The frame-independent CWM parameter extraction method [9]



(b) The proposed method

Figure 3: An example of the formant frequencies extracted using (a) the frame-independent CWM parameter extraction method [9] and (b) the proposed method. The extracted formant frequencies are plotted with different colors according to the indices of the Gaussian distribution functions in CWM.

This assumption is also made for simplifying the derivation of the parameter estimation algorithm described in the next section. $f_{\omega,l}$ denotes the sequence of spectrum models defined by Eq. 1. It should be noted that the maximization of the Poisson likelihood with respect to $f_{\omega,l}$ amounts to optimally fitting $f_{\omega,l}$ to $y_{\omega,l}$ by using the I-divergence as the fitting criterion [8]. The next section describes the parameter estimation algorithm of the generative model presented above.

3. Parameter estimation algorithm

Here we describe the parameter learning algorithm given a set of training data. The spectral sequence of the training data is considered to be an observed sequence and a set of unknown parameters of the present generative model is denoted by Θ. Θ consists of the state sequence $\boldsymbol{s} = \{s_l\}_l$, the parameters of the state emission distributions $\boldsymbol{\theta} = \{m_{k,i}, \eta_{k,i}, a^{(\rho)}_{k,i}, b^{(\rho)}_{k,i}, a^{(w)}_{k,i}, b^{(w)}_{k,i}\}_{k,i}$, and the CWM parameters $\{\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w}\}$.

Given a set of observed spectra $Y = \{y_{\omega,l}\}_{\omega,l}$, we would like to determine the estimate of Θ that maximizes the posterior density $P(\Theta|Y) \propto P(Y|\Theta)P(\Theta)$, or equivalently, $\log P(Y|\Theta) + \log P(\Theta)$. Here, $\log P(\Theta|Y)$ is written as

$$\log P(\Theta|Y) \stackrel{c}{=} \log P(Y|\Theta) + \alpha \log P(\Theta), \tag{8}$$

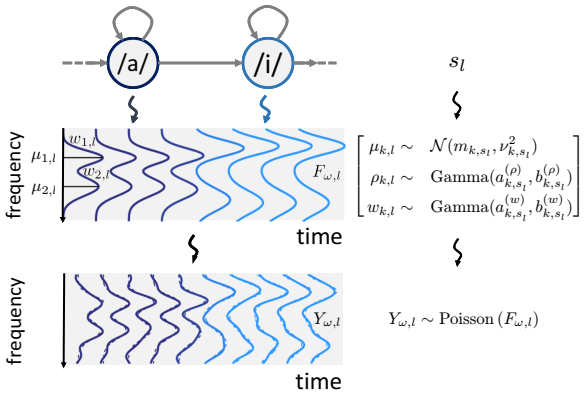


Figure 4: Illustration of the present HMM

$$\log P(\Theta) \stackrel{c}{=} \log P(\boldsymbol{s}) + \log P(\boldsymbol{\mu}|\boldsymbol{s},\boldsymbol{\theta}) + \log P(\boldsymbol{\rho}|\boldsymbol{s},\boldsymbol{\theta}) + \log P(\boldsymbol{w}|\boldsymbol{s},\boldsymbol{\theta}), \tag{9}$$

where α is a regularization parameter that weighs the importance of the log prior density relative to the log-likelihood. $\stackrel{c}{=}$ denotes equality up to constant terms. Unfortunately, this optimization problem is non-convex, and finding the global optimum is an intractable problem. However, we can employ an auxiliary function approach to find a local optimum [8].

As mentioned above, we notice from Eqs. (6) and (7) that $-\log P(Y|\Theta)$ is equal up to constant terms to the sum of the I-divergence between $y_{\omega,l}$ and $f_{\omega,l}$

$$\mathcal{I}(\Theta) := \sum_{\omega,l} \Big(y_{\omega,l} \log \frac{y_{\omega,l}}{f_{\omega,l}} - y_{\omega,l} + f_{\omega,l} \Big). \tag{10}$$

Hence, maximizing $P(\Theta|Y)$ amounts to minimizing $I(\Theta) - \alpha \log P(\Theta)$ with respect to Θ. By invoking Jensen's inequality based on the concavity of the logarithm function, we obtain an inequality

$$\mathcal{I}(\Theta) = \sum_{\omega,l} \Big(y_{\omega,l} \log \frac{y_{\omega,l}}{f_{\omega,l}} - y_{\omega,l} + f_{\omega,l} \Big) \tag{11}$$

$$\leq \sum_{\omega,l} \Big\{ y_{\omega,l} \log y_{\omega,l} - y_{\omega,l} - \sum_k \Big(\lambda_{k,\omega,l} \log \frac{g_{k,\omega,l}}{\lambda_{k,\omega,l}} + g_{k,\omega,l} \Big) \Big\}, \tag{12}$$

where

$$g_{k,\omega,l} = \sqrt{\frac{\rho_{k,l}}{2\pi}} \exp\Big(-\frac{\rho_{k,l}}{2} (\omega - \mu_{k,l})^2 \Big). \tag{13}$$

By using $\mathcal{J}(\Theta, \lambda)$ to denote the upper bound of $\mathcal{I}(\Theta)$, i.e., the right-hand side of (12), equality holds if and only if

$$\lambda_{k,\omega,l} = \frac{w_{k,l} g_{k,\omega,l}}{\sum_{j=1}^K w_{j,l} g_{j,\omega,l}}. \tag{14}$$

Now, when $\lambda_{k,\omega,l}$ is given by Eq. (14) with an arbitrary Θ, the auxiliary function $\mathcal{J}(\Theta, \lambda) - \alpha \log P(\Theta)$ becomes equal to the objective function $\mathcal{I}(\Theta) - \alpha \log P(\Theta)$. Then, the parameter

that decreases $\mathcal{J}(\Theta,\lambda)-\alpha\log P(\Theta)$ while keeping λ fixed will necessarily decrease $\mathcal{I}(\Theta)-\alpha\log P(\Theta)$, since inequation (12) guarantees that the original objective function is always even smaller than the decreased $\mathcal{J}(\Theta,\lambda)-\alpha\log P(\Theta)$. Therefore, by repeating the update of λ by Eq. (14) and the update of Θ that decreases $\mathcal{J}(\Theta,\lambda)-\alpha\log P(\Theta)$, the objective function decreases monotonically and converges to a stationary point.

With fixed $\boldsymbol{s}$ and $\boldsymbol{\theta}$, the auxiliary function $\mathcal{J}(\Theta,\lambda)-\alpha\log P(\Theta)$ is minimized with respect to the CWM parameters, $\boldsymbol{\mu}$, $\boldsymbol{\rho}$, and $\boldsymbol{w}$, under the following updates

$$\mu_{k,l}=\frac{D_{k,l}\eta_{k,i}^2\rho_{k,l}+\alpha m_{k,i}}{C_{k,l}\eta_{k,i}^2\rho_{k,l}+\alpha}, \tag{15}$$

$$\rho_{k,l}=\frac{C_{k,l}+2\alpha(a_{k,l}^{(\rho)}-1)}{C_{k,l}\mu_{k,l}^2-2D_{k,l}\mu_{k,l}+E_{k,l}-2\alpha/b_{k,l}^{(\rho)}}, \tag{16}$$

$$w_{k,l}=\frac{C_{k,l}+\alpha(a_{k,l}^{(w)}-1)}{1+\alpha/b_{k,l}^{(w)}}, \tag{17}$$

where

$$C_{k,l}=\sum_{\omega}y_{\omega,l}\lambda_{k,\omega,l}, \tag{18}$$

$$D_{k,l}=\sum_{\omega}\omega y_{\omega,l}\lambda_{k,\omega,l}, \tag{19}$$

$$E_{k,l}=\sum_{\omega}\omega^2 y_{\omega,l}\lambda_{k,\omega,l}. \tag{20}$$

With fixed $\boldsymbol{\mu}$, $\boldsymbol{\rho}$, $\boldsymbol{w}$ and $\boldsymbol{\theta}$, the state sequence minimizing $\mathcal{J}(\Theta,\lambda)-\alpha\log P(\Theta)$ can be obtained by using the Viterbi algorithm. With fixed $\boldsymbol{\mu}$, $\boldsymbol{\rho}$, $\boldsymbol{w}$ and $\boldsymbol{s}$, the auxiliary function is minimized under the following updates.

As for $\{m_{k,i},\nu_{k,i}^2\}_{k,i}$,

$$m_{k,i}=\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\mu_{k,l} \tag{21}$$

$$\nu_{k,i}^2=\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\mu_{k,l}^2-(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\mu_{k,l})^2, \tag{22}$$

where $\mathcal{T}_i=\{l|s_l=i\}$. As for $\{a_{k,i}^{(\rho)},b_{k,i}^{(\rho)},a_{k,i}^{(w)}\}_{k,l},b_{k,i}^{(w)}\}_{k,l}$, although the updates are not derived in a closed form, they are updated as the root of the following equations

$$\log a_{k,i}^{(\rho)}-\psi(a_{k,i}^{(\rho)})=\log(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{k,l})-\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{k,l} \tag{23}$$

$$\log a_{k,i}^{(w)}-\psi(a_{k,i}^{(w)})=\log(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}w_{k,l})-\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}w_{k,l} \tag{24}$$

$$b_{k,i}^{(\rho)}=\frac{\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{k,l}}{a_{k,i}^{(\rho)}} \tag{25}$$

$$b_{k,i}^{(w)}=\frac{\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}w_{k,l}}{a_{k,i}^{(w)}}, \tag{26}$$

where $\psi(a)$ denotes the digamma function

$$\psi(a)=\frac{\partial\Gamma(a)}{\partial a}/\Gamma(a). \tag{27}$$

A spectrogram of sentence A01 from the ATR503 data set 'Arayurugenjitsuwo subete jibunnohohe nejimagetanoda' and an example of the trajectory of the mean parameters of the CWM extracted from it are shown in Fig. 5.

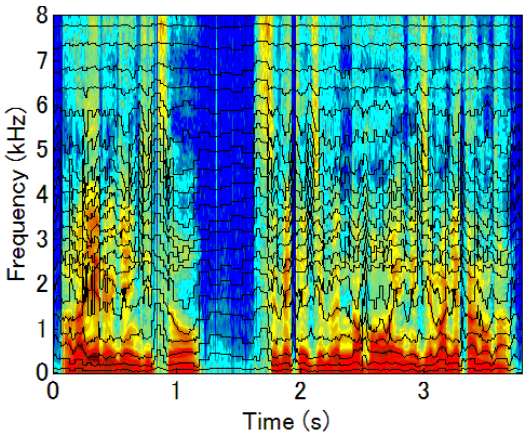


Figure 5: Spectrogram of a natural voice and the trajectories of extracted frequency parameters (a Japanese sentence of A01 in the ATR-503 data set).

4. Speech Synthesis Experiment and Evaluation

4.1. Evaluation criterion

To evaluate the spectral distortion caused by the training process, we measured the spectral distortion between the synthesized speech and real speech.

Bark spectral distortion [11] is known to be an objective measure for incorporating psychoacoustic responses. We used Bark spectral distortion and log spectral distortion to measure the spectral divergence between the synthesized and real speech. Using this criterion, we compared the present method with HTS-2.1 [10], a conventional HMM speech synthesis system with mel-cepstrum features and the global variance model [**?**].

The time alignment between the two spectral sequences was computed by using the dynamic time warping (DTW) algorithm. The average distortions on 53 synthesized speech sentences were compared and a statistical test was conducted.

4.2. Experimental Conditions

All the phoneme boundaries of the training data were given by the phoneme labels of HTS [10]. This means that the state sequences of the HMM are assumed to be known in the training stage. The number of the Gaussian functions in the CWM was set at 24. The initial CWM parameters were determined by performing the method reported by [9] to estimate the CWM parameters from the entire spectral sequence corresponding to each phoneme. We used an acoustic model of 1 state and a left-to-right HMM with no skip, as with HTS-2.1 [10]. We used "STRAIGHT" [12] to compute the spectral envelopes of the training data, with an analysis interval of 5 ms. We used 450 uttered sentences for the training and 53 uttered sentences for the synthesis and evaluation, respectively. All the speech samples were uttered by a Japanese male speaker and recorded with a 16kHz sampling rate and a 16 bit quantization, which were taken from the demo site of HTS [13].

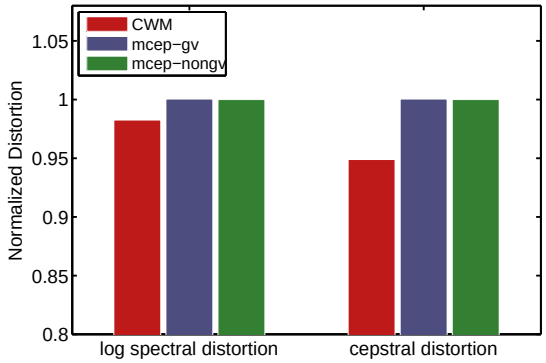


Figure 6: Normalized squared distances between the log spectra and cepstra of real and synthesized speech by the proposed method (red) , the ordinary mel-cepstral method with gv (blue) and without gv (green).

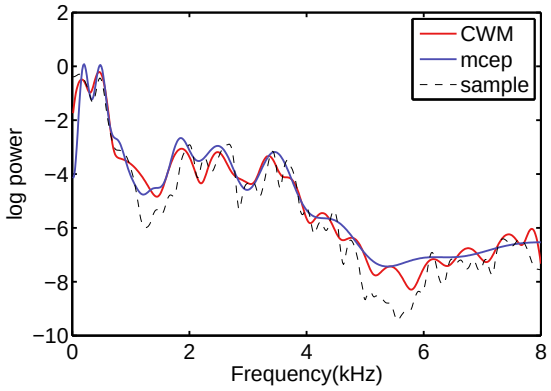


Figure 7: An example of the synthesized spectral envelope corresponding to the phoneme '/e/'

4.3. Experimental Results

The measured spectral distortions are shown in Fig. 6. The distances were normalized according to the values obtained with the proposed and conventional methods. We confirmed through a statistical test that the Bark spectral distortion obtained with the proposed method was significantly smaller than that obtained with the conventional method, while the two methods were comparable in terms of the log spectrum distortion. This result can be interpreted as showing that the proposed model was able to synthesize spectra properly especially in a low frequency region. Since the changes in the peak frequencies of spectra are mainly seen in the low frequency region, this result indicates that the proposed model is superior to the conventional model as regards modeling fluctuations in the frequency direction. Fig. 7 shows an example of the synthesized spectral envelope along with the spectral envelope of real speech corresponding to the phoneme '/e/'. In addition to the superiority of the proposed model in terms of the Bark spectral distortion, the proposed method was able to restore spectral peaks and dips more clearly than the conventional model, especially in the 1.5 to 6 kHz frequency range. In future we plan to conduct a subjective evaluation test to confirm whether the perceptual quality of the synthesized speech has actually been improved.

5. Conclusion

This paper proposed a text-to-speech synthesis (TTS) system based on a combined model consisting of the Composite Wavelet Model (CWM) and the Hidden Markov Model (HMM). To avoid the over-smoothing of spectra generated by the conventional HMM-based TTS systems using cepstral features, we considered it important to focus on a different representation of the generative process of speech spectra. In particular, we chose to characterize the speech spectra with the CWM, whose parameters correspond to the frequency, gain and peakiness of each underlying formant. This idea was motivated by our expectation that averaging these parameters would not directly cause the over-smoothing of the spectra, unlike the cepstral representations. To describe the entire generative process of a sequence of speech spectra, we combined the generative process of a formant trajectory using an HMM and the generative process of a speech spectrum using the CWM. We developed a parameter learning algorithm for this combined model based on an auxiliary function approach. We confirmed through experiments that our speech synthesis system was able to generate speech spectra with clear peaks and dips, which resulted in natural-sounding synthetic speech.

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Eurospeech 1999*, 1999, pp. 2347–2350.

[2] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 191–196.

[3] Y. Qian, Z.-J. Yan, Y.-J. Wu, F. K. Soong, G. Zhang, and L. Wang, "An HMM trajectory tiling (HTT) approach to high quality TTS," in *Proc. of Interspeech 2010*, 2010, pp. 422–425.

[4] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, no. 1, p. S35, 1975.

[5] T. Toda and T. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[6] T. Saikachi, K. Matsumoto, S. Sako, and S. Sagayama, "Discussion about speech analysis and synthesis by composite wavelet model," in *Proc. ASJ Spring Meeting*, vol. 2, 2006, pp. 89–94, in Japanese.

[7] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," in *Proc. of ICSLP '96*, vol. 2, 1996, pp. 1229–1232.

[8] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *IEEE Trans. Audio, Speech & Language Process.*, vol. 18, no. 6, pp. 1507–1516, 2010.

[9] N. Hojo, K. Minami, D. Saito, H. Kameoka, and S. Sagayama, "HMM speech synthesis using speech analysis based on composite wavelet model." in *Proc. ASJ Autumn Meeting*, no. 2-7-7, 2013, in Japanese.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. Proc. of 6th ISCA Speech Synthesis Workshop*, vol. 6, 2007, pp. 294–299.

[11] S. Wan, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-freqency smoothing and an instantaneous-frequency-based f0 extraction," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[13] http://hts.sp.nitech.ac.jp/.