PAPER

DNN-Based Full-Band Speech Synthesis Using GMM Approximation of Spectral Envelope

Junya KOGUCHI[†a)], ***Nonmember***, **Shinnosuke TAKAMICHI**[††b)], **Masanori MORISE**[†c)], **Hiroshi SARUWATARI**[††d)], ***and*** **Shigeki SAGAYAMA**[†††e)], ***Members***

SUMMARY We propose a speech analysis-synthesis and deep neural network (DNN)-based text-to-speech (TTS) synthesis framework using Gaussian mixture model (GMM)-based approximation of full-band spectral envelopes. GMMs have excellent properties as acoustic features in statistic parametric speech synthesis. Each Gaussian function of a GMM fits the local resonance of the spectrum. The GMM retains the fine spectral envelope and achieve high controllability of the structure. However, since conventional speech analysis methods (i.e., GMM parameter estimation) have been formulated for a narrow-band speech, they degrade the quality of synthetic speech. Moreover, a DNN-based TTS synthesis method using GMM-based approximation has not been formulated in spite of its excellent expressive ability. Therefore, we employ peak-picking-based initialization for full-band speech analysis to provide better initialization for iterative estimation of the GMM parameters. We introduce not only prediction error of GMM parameters but also reconstruction error of the spectral envelopes as objective criteria for training DNN. Furthermore, we propose a method for multi-task learning based on minimizing these errors simultaneously. We also propose a post-filter based on variance scaling of the GMM for our framework to enhance synthetic speech. Experimental results from evaluating our framework indicated that 1) the initialization method of our framework outperformed the conventional one in the quality of analysis-synthesized speech; 2) introducing the reconstruction error in DNN training significantly improved the synthetic speech; 3) our variance-scaling-based post-filter further improved the synthetic speech.

key words: *Gaussian mixture model, spectral envelope, vocoder, deep neural network, text-to-speech synthesis*

1. Introduction

Text-to-speech synthesis (TTS) [1] is an important technology for users and computers to engage in natural spoken dialogue. In this regard, black-boxed end-to-end models can synthesize high-fidelity speech in TTS [2]. In a production scenario, however, it is important not only to achieve full-band and high-quality synthesis but also to allow users to control speech characteristics according to their preferences. Statistical parametric speech synthesis (SPSS) [3] is expected to be applied to be a full-band and highly-controllable TTS system because it uses acoustic features as a low-dimensional intermediate representation in the process of generating the speech waveform from the text.

In SPSS, acoustic features significantly affect the quality of synthetic speech and controllability. Mel-cepstrum [4] is a well-known example of representation; it approximates the spectral envelope with a superposition of trigonometric functions. However, statistical averaging of mel-cepstrum in SPSS changes the entire original structure and significantly degrades synthetic speech quality. Also, decomposition by trigonometric functions does not result in high controllability. To address this problem, approximation of spectral envelopes by using Gaussian mixture models (GMMs) has been proposed [5]. An overview of our proposed analysis-synthesis framework using GMMs is shown in Fig. 1. The mean, variance, and weight of a Gaussian function*, which are called *GMM parameters* in this current paper, respectively express frequency, sharpness, and amplitude (or power) of a peak of a spectrum such as a formant. Since formants are well-known features for visualization [6], the GMM parameters, related to formants, are more intuitive than mel-cepstrum. In addition, the GMM parameters are more stable than line spectral pair (LSP) parameters, which are other formant-related features [7]. SPSS using LSP parameters suffers from instability owing to the misordering of predicted LSP parameters [8] and typically requires an additional refinement to avoid the problem. This is because LSP models the formant structure with a pair of

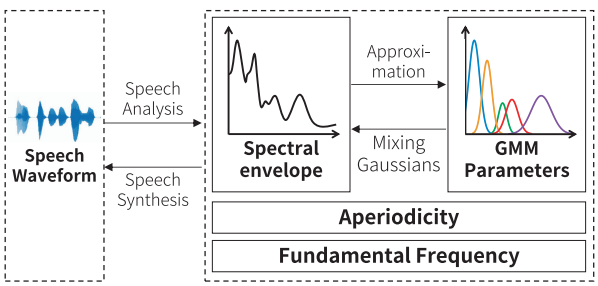


Fig. 1 Overview of proposed analysis-synthesis framework using GMM-based approximation of spectral envelope.

Manuscript received April 3, 2020.
Manuscript revised July 15, 2020.
Manuscript publicized September 3, 202.

†The authors are with Meiji University, Tokyo, 164–8525 Japan.
††The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.
†††The author is with The University of Electro-Communications, Chofu-shi, 182–8585 Japan.
a) E-mail: cs202027@meiji.ac.jp
b) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
c) E-mail: mmorise@meiji.ac.jp
d) E-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
e) E-mail: sagayama@m.ieice.org
DOI: 10.1587/transinf.2020EDP7075

*In this current paper, "Gaussian function" refers to a function that approximates a spectrum. "Normal distribution" refers to the probability distribution. Note that, the former does not satisfy the definition of probability distribution, but the latter does.

line spectral frequencies. On the other hand, our GMM represents a formant with a single Gaussian function. Therefore, our GMM never suffers from instability. Furthermore, our GMM allows us to control formants of synthetic speech more flexibly than LSP. This is because our GMM can independently control an amplitude and sharpness of formants whereas LSP cannot [9]. In GMM parameters estimation, an iterative algorithm can estimate the GMM parameters that approximate the spectral envelope and is initialized based on LSP analysis. Unlike mel-cepstrum, statistical averaging of GMM parameters is expected to alleviate the loss of fine structures because each Gaussian function fits each peak of the spectral envelope. However, LSP-based initialization creates GMM parameters that fit densely in a low-frequency band. Therefore, it is affected by quality degradation during analysis and synthesis of full-band speech. In addition, a highly controllable TTS system can be expected by formulating a DNN-based method [10] using GMM parameters, but it is never formulated.

We propose a speech analysis-synthesis and deep neural network (DNN)-based TTS framework for a full-band speech that uses GMM parameters. The framework also consists of a peak-picking-based initialization method for full-band speech analysis. The initialization provides the initial GMM parameters that accurately fit the full-band spectral envelopes. We introduce two objective criteria of DNN training by using the GMM parameters. One is a minimization of prediction errors in the GMM parameter domain and the other is a minimization of the reconstruction error of the spectral envelope. We also developed a variance-scaling-based post-filter for our framework. The post-filter efficiently uses GMM-based modeling and enhances the quality of synthetic speech by modifying the variance parameters. Experimental results from evaluating our framework indicate that 1) the initialization method of our framework outperforms the conventional one in quality of analysis-synthesized speech; 2) the DNN training criterion that introduces the reconstruction error is highly effective in improving synthetic speech; 3) our variance-scaling-based post-filter further improves the synthetic speech. The rest of this paper is organized as follows and an overview is given in Fig. 2. In Sect. 2, we briefly review the conventional analysis-synthesis framework using GMM parameters and describe an iterative algorithm for GMM-parameter estimation. In Sect. 3, we point out the problem of the conventional initialization method and introduce the peak-picking-based initialization of our proposed framework. In Sect. 4, we explain the DNN-based TTS method of our framework that uses GMM parameters and our post-filter. In Sect. 5, we discuss the evaluation of the quality of analysis-synthesized and TTS-synthesized speech generated with our framework. In Sect. 6, we conclude the paper.

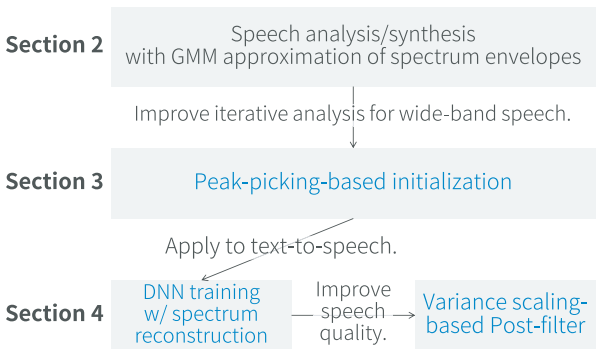


Fig. 2 The overview of this paper.

2. Speech Analysis-Synthesis Framework Using GMM Parameters

As shown in Fig. 1, our framework consists of speech analysis and synthesis. In Sect. 2.1 and Sect. 2.2, we describe the details of speech analysis and synthesis with GMM parameters, respectively.

2.1 Speech Analysis: Spectral Envelope Approximation with GMM Parameters

We first extract spectral envelope, fundamental frequency (F_0), and aperiodicity by usual speech analysis such as WORLD [11] (D4C edition [12]). We also approximate the extracted spectral envelope with a GMM [13]. The GMM parameters are estimated by minimizing a loss function of the observed spectral envelope $H(\omega)$ and the GMM-approximated one $G(\omega)$ expressed by

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

where ω denotes frequency, K is the number of mixture components, and μ_k, ${\sigma_k}^2$, w_k denote mean, variance, and weight of a Gaussian function with index k, respectively. The loss function to be minimized is a divergence between two different probability distributions, and this framework uses the I-divergence $I(H, G)$ given as

$$I(H,G) = \sum_{\omega}\left[H(\omega)\log\frac{Y(\omega)}{G(\omega)} - H(\omega) + G(\omega)\right]. \tag{2}$$

In addition, a mean transition modeling term [13] helps to estimate a temporally smooth mean trajectory. Since it is difficult to directly minimize the loss function, majorization-minimization (MM) is used as an algorithm to iteratively estimate GMM parameters. Since the MM algorithm theoretically guarantees monotonic non-increase in the objective function during iterations, the value of the objective function always converges without setting any hyperparameters. Originally, we should use the IS (Itakura–Saito) divergence that has better peak-sensitive property [14]. However, the convergence-guaranteed algorithm, i.e., the MM algorithm, cannot be applied to analysis using the IS divergence. On the other hand, the I-divergence has a peak-sensitive property, and the convergence-guaranteed algorithm can be applied to analysis using the I-divergence. Therefore, we used

the I-divergence for analysis, instead of the IS divergence.

Before executing the iterative algorithm, GMM parameters are initialized with arbitrary values. Specifically, the initial value of μ_k greatly affects not only the convergence speed of the algorithm but also to which resonance component the Gaussian function fits. The conventional method [13] uses the average of odd- and even-order pairs in the $2K$-order LSP parameters $\omega_1, \ldots, \omega_{2K}$ as follows [7]:

$$\mu_k = \frac{\omega_{2k-1} + \omega_{2k}}{2}. \tag{3}$$

In addition, w_k and σ_k are initialized with an amplitude value at the frequency of μ_k and a constant value, respectively.

2.2 Speech Synthesis: Spectral Envelope Reconstruction from GMM Parameters

The speech waveform is generated by the vocoder using the aperiodicity, F_0, and the spectral envelopes, which are reconstructed from the GMM parameters by using Eq. (1). With our framework, speech waveforms are generated by filtering based on minimum-phase response.

3. GMM-Parameter Initialization Method for Full-Band Speech Analysis-Synthesis

3.1 Problem of LSP-Based Parameter Initialization

As described in Sect. 2, the mean parameters in iterative speech analysis are initialized using the LSP parameters. The LSP parameters tends to correspond to the resonance frequencies of a vocal tract and are an appropriate initial values for a narrow-band speech. However, since not all LSP frequencies fit the resonance frequencies, the approximation accuracy decreases [15]. Moreover, since the resonances are observed in lower-frequency bands where a large amount of energy is distributed, the LSP frequencies and mean parameters fit densely in the lower frequency band. The full-band speech suffers from this LSP's problem; the spectral envelopes in the higher-frequency band are approximated inappropriately. Figure 3 shows the LSP-initialized (top) and finally obtained GMM parameters (bottom). The initial GMM parameters overfit the lower-frequency band and underfit the higher one. This tendency is also observed in finally obtained parameters. This degrades the approximation accuracy and quality of analysis-synthesized speech.

3.2 Proposed Framework: Initialization Method Based on Peak-Picking

To solve the above problem, our framework consists of an alternative initialization method using peak-picking. Peak picking generally means finding all local maxima. The method is applied to the spectral envelope and used for finding peak frequencies[†]. The peak frequencies are used as initial mean parameters. The remaining parameters of the GMM, i.e., weights and variances, are initialized in the same manner as with the conventional initialization method. Figure 4 shows the initial Gaussian functions with our peak-picking-based method (top) and finally obtained ones (bottom). Unlike the conventional method shown in Fig. 3, our initialization places the Gaussian functions over the entire frequency band. Also, they are still placed in the entire band even after the iterations. This accurate approximation provides a higher quality of analysis-synthesized speech than with the conventional method. Another simple method for widely placing the mean parameters over the entire band is a flat start; the mean parameters are placed at equal

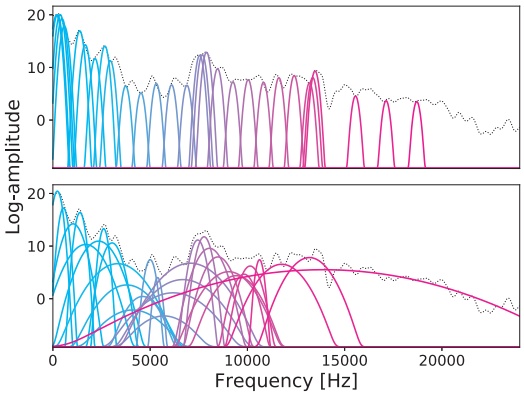


Fig. 3 Conventional LSP-based initial (top) and finally obtained (bottom) Gaussian functions ($K = 30$). Broken line is observed spectral envelope. Gaussian functions densely fit in low band, and not all fit peaks of spectrum.

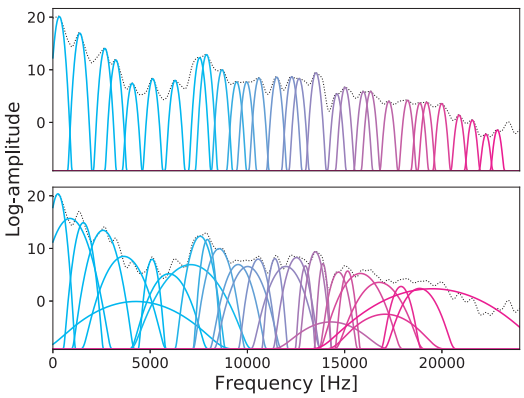


Fig. 4 Our peak-pick-based initial (top) and finally obtained (bottom) Gaussian functions with our peak-pick ($K = 30$). broken line is an observed spectral envelope. All of initial Gaussian functions fit to the peaks directly. Moreover, they still fit over the entire frequency band after the iteration.

[†]Our initialization suffers from the remaining fine structure (e.g., F_0 harmonics) or noisy spectral envelopes. Actually, when approximating a amplitude spectrum of discrete Fourier transform with a GMM, some Gaussian functions are allocated to pitch and harmonic structures [5]. In this current paper, we assume a spectrum structure is smooth, and a peak-picking algorithm we used finds local maxima of a smooth spectrum.

frequency intervals. However, there is no guarantee that the initial value corresponds to the peak frequency like the LSP-based initialization. Our preliminary experimental results indicate that the flat start degrades approximation accuracy, convergence speed, and speech quality. Therefore, we use the peak-picking-based method.

4. Proposed Framework: DNN-Based TTS Method Using GMM Parameters

In this section, we explain our DNN-based TTS method using GMM parameters. We introduce two DNN training criteria: prediction error of GMM parameters (Sect. 4.1.1) and reconstruction error of spectral envelope (Sect. 4.1.2). We also explain our a post-filter based on variance scaling to improve the quality of synthetic speech (Sect. 4.2).

4.1 DNN Training with GMM Parameters

The DNN acoustic model using the GMM parameters outputs acoustic features from an input context vector like a conventional DNN-based TTS method using mel-cepstrum [10]. In the case of the GMM parameters as acoustic features, the DNN outputs GMM parameters [16], F_0 and aperiodicity. The training criterion for F_0 and aperiodicity prediction is a minimization of mean squared error (MSE). In Sect. 4.1.1 and Sect. 4.1.2, we discuss the training criteria for the GMM parameters.

4.1.1 Prediction Error in GMM-Parameter Domain

The simple way to train a DNN is using the MSE between the predicted and target GMM parameters, as when using mel-cepstrum in the conventional DNN-based TTS. The DNNs are trained by minimizing the MSE as follows:

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2, \tag{4}$$

where $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ are the vectors of the target and predicted GMM parameters at each frame, respectively. They are concatenated vectors of μ_k, σ_k, and w_k for each frame. The middle of Fig. 5 corresponds to this criterion.

4.1.2 Reconstruction Error of Spectral Envelopes

Since the extraction of the GMM parameters is estimated independently among utterances, the resulting GMM parameters are different even between contextually similar utterances. For example, there is no guarantee that one unique Gaussian function (e.g., $k = 1$) always fits the unique formant (e.g., the first formant) of every utterance. This lack of uniformity negatively affects the MSE-based training, i.e., different formants (e.g., first and second ones) are averaged and disappear by statistical averaging. Figure 6 shows two temporal trajectories of mean parameters: the μ_k estimated from two contextually similar utterances. We can see that μ_2 of the left- and right-sides is expected to draw

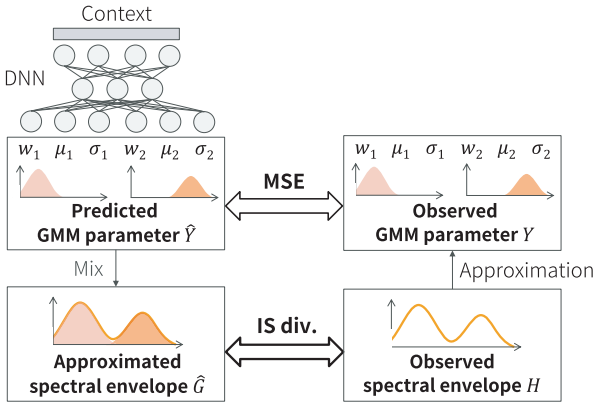


Fig. 5 The training criteria for DNN training. MSE is used for minimizing prediction errors of GMM parameters. IS div. is used for minimizing reconstruction errors of spectral envelopes. Note that, GMM parameters vary frame by frame, but frame notation (i.e., frame index) is omitted for simple illustration.

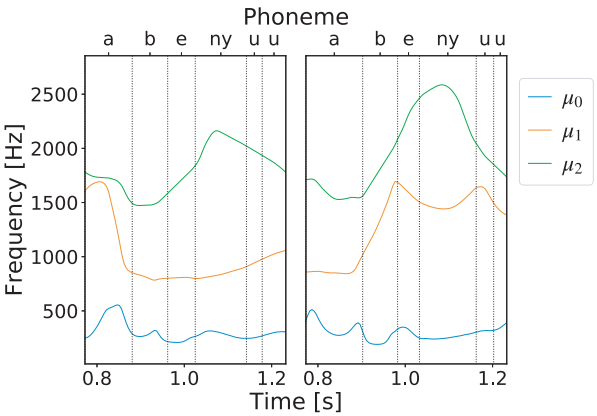


Fig. 6 Temporal trajectories of mean parameters: μ_k estimated from two contextually similar utterances. μ_2 of left- and right-sides is expected to draw similar trajectory but actually fit different respective formants.

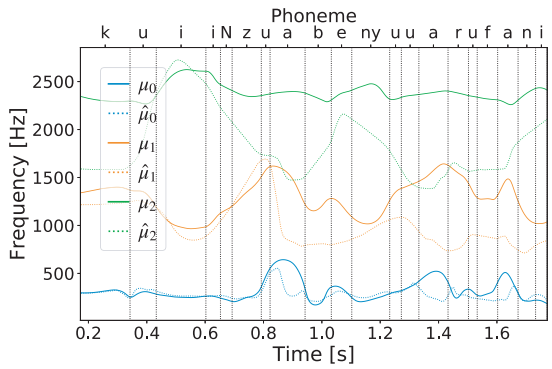


Fig. 7 Temporal trajectories of mean parameters: target μ_k and predicted one $\hat{\mu}_k$. $\hat{\mu}_2$ is between μ_2 and μ_3.

a similar trajectory but actually fit different respective formants. Also, Fig. 7 shows two temporal trajectories of mean parameters: the target one μ_k and predicted one $\hat{\mu}_k$ from the DNN trained by MSE minimization (Sect. 4.1.1). We can see that a temporal trajectory of $\hat{\mu}_2$ is between μ_2 and

μ_3. This indicates that μ_2 fits the second formant in one utterance in speech analysis but fits the third formant in another contextually similar utterance, as shown in Fig. 6. As described above, this problem is caused by context-ignored extraction of the target GMM parameters. Therefore, the target of DNN training should avoid the context-ignored processes. Since the GMM parameters work to approximate the spectral envelope, we can use the spectral envelope as the target of DNN training. Therefore, we use a reconstruction error of spectral parameters. We minimize the error between the observed spectral envelopes $H(\omega)$ and approximated ones $G(\omega)$ reconstructed from the predicted GMM parameters. The reconstruction error is the IS divergence $\mathcal{L}_{\mathrm{IS}}(\boldsymbol{H}, \hat{\boldsymbol{G}})$, which is defined as

$$\mathcal{L}_{\mathrm{IS}}(\boldsymbol{H}, \hat{\boldsymbol{G}}) = \sum_{\omega} \left[\frac{H(\omega)}{\hat{G}(\omega)} - \log \frac{H(\omega)}{\hat{G}(\omega)} - 1 \right]. \tag{5}$$

Here, $\hat{G}(\omega)$ is calculated from $\hat{\boldsymbol{Y}}$ using Eq. (1). As described in Sect. 2.1, the IS divergence has better peak-sensitive property to quantify the similarity of two spectra, not GMM parameters. The DNN is pre-trained by only prediction error and followed by training using only reconstruction error. This training can alleviate the problem of context-ignored extraction discussed above. Of course, it is possible to train a DNN without the pre-training. However, a preliminary experiment confirmed that the quality of the synthetic speech without the pre-training is significantly worse than that with the pre-training.

4.1.3 Multi-Task Learning Using Prediction and Reconstruction Errors

The above two criteria consider one error in training. Here, we propose another criterion that minimizes the prediction error and the reconstruction error simultaneously. The training criterion is the weighted sum of the two errors as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) + \alpha \mathcal{L}_{\mathrm{IS}}(\boldsymbol{H}, \hat{\boldsymbol{G}}), \tag{6}$$

where α is a weight. Equation (6) can be considered as multi-task learning of minimizing the prediction error in the GMM-parameter domain and the reconstruction error in the spectral-envelope domain.

4.2 Variance-Scaling-Based Post-Filter

In this section, we discuss our post-filter for the GMM parameters to enhance the quality of the TTS-synthesized speech. We focus on a difference in the variance parameters of natural and synthetic speech. Figure 8 shows the histograms of the variance parameters of natural and synthetic speech respectively. We can see that the variance parameters of natural speech are smaller than of synthetic speech. This appears to be because the predicted variance parameters are averaged by the statistical training of the DNN. Since a variance parameter corresponds to sharpness of the resonance component, the large variance parameters over-smoothen the spectral envelopes and degrade speech quality.

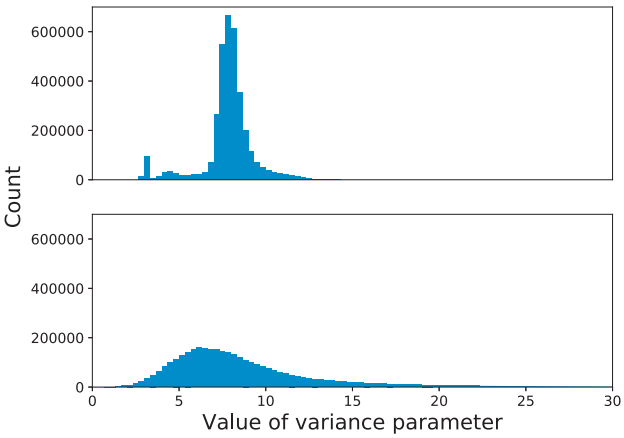


Fig. 8 Histograms of variance parameters: one estimated from natural speech (top) and one predicted from DNN (bottom). Predicted ones tend to be larger than natural ones.

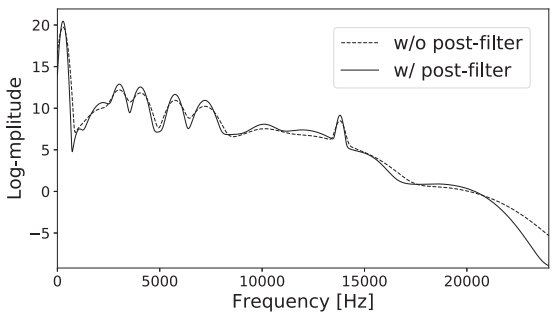


Fig. 9 Spectral envelopes with (solid line) and awithout (broken line) post-filter. Post-filtering enhances peak and dip of spectrum (the coefficient is 0.75).

To enhance the quality of synthetic speech, we use a post-filtering method to scale variance parameters[†]. The smaller the variance parameter is, the sharper the Gaussian function. Therefore, we can obtain the formant-enhancement effect by multiplying the variance parameters by a constant emphasis coefficient of less than 1. This variance scaling works to sharpen the GMM functions, i.e., alleviates over-smoothing. Figure 9 shows the spectral envelopes with and without post-filtering. We can see that the peak and dip of the spectrum are enhanced due to our post-filtering. Note that, in this paper, we consider this variance scaling as a post-filter to enhance the TTS-synthesized speech; however, the method is also effective for improving the quality of the analysis-synthesized speech. We conducted a preliminary experiment using the analysis-synthesized speech with and without variance scaling and confirmed the quality improvements.

[†]This is a GMM-specific post-filtering method whereas the cepstrum emphasis [17] is a mel-cepstrum-specific one.

5. Experimental Evaluation

5.1 Experimental Evaluation for Analysis-Synthesized Speech

5.1.1 Experimental Setup

We evaluated the quality of the analysis-synthesized speech by using our framework. We conducted a preference AB test. Twenty-five people participated in the experiment on a crowdsourcing platform [18]. We used the JVS corpus, a multi-speaker speaking voice corpus [19]. Male and female speakers were randomly selected, and 100 utterances per speaker were selected from the PARALLEL100 subset for analysis-synthesis speech. The WORLD vocoder [11] (D4C edition [12]) was used for extracting the spectral envelope, F_0, and aperiodicity. The sampling frequencies were 16, 24, and 48 kHz, and the analysis window lengths were 1024, 1024 and 2048 samples respectively. The frame shift was set to 5 ms. The number of mixed components of the GMMs was set to 30 according to the results of a previous subjective evaluation [20]. We used a peak-picking algorithm implemented in MATLAB [21], but the Python version is also available [22]. Picking peaks is equivalent to find local maxima, and the algorithm simply finds data points that are larger than the neighboring two points. After finding peaks, we sorted them in frequency order and used the first K peaks for initialization.

We compared the following acoustic features.

- **GMM-LSP**: GMM parameters estimated by the LSP-based initialization method (Sect. 3.1)
- **GMM-PEAK** (ours): GMM parameters estimated using our peak-picking-based initialization method (Sect. 3.2)
- **MCEP40**: 40-dimensional mel-cepstrum
- **MCEP60**: 60-dimensional mel-cepstrum

The comparisons with mel-cepstrum were not main aim of this study; however, it is beneficial to evaluate the performance of the GMM-based approximation.

5.1.2 Experimental Results

Table 1, Table 2, and Table 3 list the results for 16, 24, and 48 kHz-sampled speech, respectively. GMM-PEAK has no significant improvement in quality from GMM-LSP in 16 kHz-sampled speech. The result is reasonable because LSP extraction works for the narrow-band speech as described in Sect. 3.1. Otherwise, it performed better than GMM-LSP in 24 kHz- and 48 kHz-sampled speech. According to these results, our peak-picking-based initialization method works in analysis-synthesis of full-band speech.

In comparison with mel-cepstrum, the quality of GMM-PEAK is comparable with MCEP40 and MCEP60 in 24 kHz-sampled speech. On the other hand, it is significantly degraded compared to MCEP40 in 16 kHz- and 48 kHz-sampled speech and MCEP60 in 16 kHz- and 48 kHz-sampled speech. We expect that optimizing the number of mixture components can alleviate this degradation. Automatic optimization of the number of mixture components is for future work.

Table 1 Results of preference AB tests on naturalness of 16 kHz-sampled analysis-synthesized speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
GMM-PEAK	0.505 vs. 0.495	8.41×10^{-1}	GMM-LSP
GMM-PEAK	0.454 vs. **0.556**	4.56×10^{-2}	MCEP40
GMM-PEAK	0.455 vs. 0.545	7.21×10^{-2}	MCEP60

Table 2 Results of preference AB tests on naturalness of 24 kHz-sampled analysis-synthesized speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
GMM-PEAK	**0.570** vs. 0.430	5.03×10^{-3}	GMM-LSP
GMM-PEAK	0.520 vs. 0.480	4.24×10^{-1}	MCEP40
GMM-PEAK	0.470 vs. 0.530	2.31×10^{-1}	MCEP60

Table 3 Results of preference AB tests on naturalness of 48 kHz-sampled analysis-synthesized speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
GMM-PEAK	**0.690** vs. 0.310	$< 10^{-10}$	GMM-LSP
GMM-PEAK	0.365 vs. **0.635**	$< 10^{-10}$	MCEP40
GMM-PEAK	0.345 vs. **0.655**	$< 10^{-10}$	MCEP60

5.2 Experimental Evaluation for TTS-Synthesized Speech

5.2.1 Experimental Setup

We evaluated the quality of TTS-synthesized speech. Fifty people participated in this experiment. We used the JSUT corpus, a corpus of speech spoken by a single female speaker [23], as a training dataset. We selected 4,500 and 500 utterances from the subset BASIC5000 for training and validation, respectively. We also selected 100 utterances from the subset VOICEACTRESS100 for evaluation.

The sampling frequencies were 24 and 48 kHz. The conditions of acoustic feature extraction were the same as those mentioned in Sect. 5. The DNN architectures were feed-forward networks that included three hidden layers, each of which has 1024 units and rectifier linear units as activation functions [24]. The weights and biases for each layer are randomly initialized. The input linguistic features were 535-dimensional context information normalized in the range of 0.01 to 0.99. The speech features output from the DNNs were 3×30-dimensional GMM parameters, log-scaled continuous F_0, averaged parameters of aperiodicity in five bands [25], their dynamic features (Δ and $\Delta\Delta$), and the voiced/unvoiced flag. These features are standardized to mean 0 and variance 1 The weight of IS divergence α in IS+MSE is set to 10^{-3} to match the scales of MSE and IS

Table 4 Results of preference AB tests on naturalness of 24 kHz-sampled synthetic speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
IS	**0.632** vs. 0.368	$< 10^{-10}$	MSE
MSE+IS	**0.704** vs. 0.296	$< 10^{-10}$	MSE
MSE+IS	0.504 vs. 0.496	8.58×10^{-1}	IS
MSE+IS	0.468 vs. 0.532	1.53×10^{-1}	MCEP40
MSE+IS	0.448 vs. 0.552	2.00×10^{-1}	MCEP60

Table 5 Results of preference AB tests on naturalness of 48 kHz-sampled synthetic speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
IS	**0.928** vs. 0.072	$< 10^{-10}$	MSE
MSE+IS	**0.822** vs. 0.178	$< 10^{-10}$	MSE
MSE+IS	0.508 vs. 0.492	7.21×10^{-1}	IS
MSE+IS	0.484 vs. 0.516	4.75×10^{-1}	MCEP40
MSE+IS	0.428 vs. **0.572**	1.24×10^{-3}	MCEP60

divergence. We compared following methods.

- **MSE** (ours): DNN w/ GMM parameters, trained by only MSE (Sect. 4.1.1)
- **IS** (ours): DNN w/ GMM parameters, trained by only IS divergence (Sect. 4.1.2)
- **MSE+IS** (ours): DNN w/ GMM parameters, trained by MSE and IS divergence (Sect. 4.1.3)
- **MCEP40**: DNN using 40-dimentional mel-cepstrum
- **MCEP60**: DNN using 60-dimentional mel-cepstrum

DNNs of MCEP40 and MCEP60 were trained based on minimization of MSE in the mel-cepstrum domain [10].

5.2.2 Comparison of Training Criteria

Table 4 and Table 5 list the results for 24 kHz- and 48 kHz-sampled speech, respectively. IS and IS+MSE performed significantly better than MSE in 24 kHz- and 48 kHz-sampled speech to the same extent. According to these results, introducing spectral reconstruction improved speech quality.

In comparison with mel-cepstrum, the quality of MSE+IS was comparable with MCEP40 in 24 kHz- and 48 kHz-sampled speech and MCEP60 in 24 kHz-sampled speech. On the other hand, it was performed worse than MCEP60 in 48 kHz-sampled speech; however, its score was not much lower as the that of analysis-synthesis. There was no significant difference in the quality of 24 kHz-sampled TTS-synthesized speech. For 48 kHz-sampled synthetic speech, there was no difference in the quality between MSE+IS and MCEP40. According to these results, the DNN could learn speech characteristics efficiently from the GMM parameters.

5.2.3 Effect of Post-Filtering

To evaluate the quality of post-filtered synthetic speech, we conducted a preference AB test using non-filtered (i.e., MSE+IS) and filtered speech. The other condition was the same as that mentioned in Sect. 5.2. The post-filter coefficient was set to 0.75. Experimental results using other coefficients are shown in Appendix.

Table 6 Results of preference AB tests on naturalness of 24 kHz-sampled synthetic speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
MSE+IS (post-filtered)	**0.720** vs. 0.280	$< 10^{-10}$	MSE+IS

Table 7 Results of preference AB tests on naturalness of 48 kHz-samples speech. Bold indicates more preferred method with p-value < 0.05.

A	Scores	p-value	B
MSE+IS (post-filtered)	**0.609** vs. 0.391	1.08×10^{-4}	MSE+IS

The results are shown in Table 6 and Table 7. The synthetic speech post-filtered obtained a significantly higher score. This indicates that formant enhancement based on variance parameter scaling is effective for improving the quality of synthetic speech.

6. Conclusion

We proposed a speech analysis-synthesis and DNN-based TTS framework for full-band speech using GMM parameters. The GMM parameters are promising as acoustic features in SPSS. The framework also consisted of a peak-picking initialization method for the iterative estimation algorithm to apply the GMM approximation to full-band speech. We introduced criteria for not only the error between the GMM parameters but also for the reconstruction error of the spectral envelope for the DNN training. We also developed a post-filter for our framework, which is based on variance scaling to improve the quality of synthetic speech. The experimental results of evaluating our framework indicated that 1) our initialization method outperforms the conventional one in the quality of analysis-synthesized speech; 2) it is effective to introduce the reconstruction error of the spectra to train a DNN using the GMM parameters, and the GMM parameters have a representation from which the DNNs can efficiently learn speech features; and 3) our variance-scaling-based post-filter is effective in improving the quality of synthetic speech. Future work includes optimizing the number of mixture components and improving extraction accuracy.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19H01116, JST PRESTO Grant Number JPMJPR18J8, and JSPS and CAS under the Japan–People's Republic of China Research Cooperative Program.

References

[1] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," ICSLP90, Kobe, Japan, pp.1089–1092, Nov. 1990.
[2] Y. Wang, R.J.S.-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Proc. INTERSPEECH, Stockholm, Sweden, pp.4006–4010, Aug. 2017.
[3] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Communication, vol.51, no.11, pp.1039–1064, 2009.
[4] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," Proc. ICSLP, Yokohama, Japan, pp.410–415, Sept. 1994.
[5] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," Proc. ICSLP, vol.2, pp.1229–1232, 1996.
[6] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.16) [computer program]. retrieved 10 july 2020," 2020.
[7] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," The Journal of the Acoustical Society of America, vol.57, no.S1, p.S35, 1975.
[8] Y. Ohtani, M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "Histogram-based spectral equalization for HMM-based speech synthesis using mel-LSP," Proc. INTERSPEECH, Portland, U.S.A., Sept. 2012.
[9] B.P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and gaussian mixture model," Acoustical Science and Technology, vol.30, no.3, pp.170–179, 2009.
[10] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," Proc. ICASSP, Vancouver, Canada, May 2013.
[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1877–1884, July 2016.
[12] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," Speech Communication, vol.84, pp.57–65, 2016.
[13] N. Hojo, K. Minami, D. Saito, H. Kameoka, and S. Sagayama, "HMM speech synthesis using speech analysis based on composite wavelet model," Proc. ASJ Autumn Meeting, vol.2012, pp.2–2–7, 2012, in Japanese.
[14] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," Journal of the Royal Statistical Society B, vol.39, pp.185–197, 1968.
[15] S. Sagayama and F. Itakura, "Theorical study in CSM, LSP and their properties," Transaction of the Comittee on Speech Research, The Acoustical Society of Japan, vol.S82-14, pp.105–112, 1982, in Japanese.
[16] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," Proc. ICASSP, Florence, Italy, pp.3872–3876, May 2014.
[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a mixed excitation model and postfilter into hmm-based text-to-speech synthesis," Systems and Computers in Japan, vol.36, no.12, pp.43–50, 11 2005.
[18] "Lancers," https://www.lancers.jp/.
[19] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint 1908.06248, Aug. 2019.
[20] J. Koguchi and S. Sagayama, "Composite wavelet model for stability-oriented speech synthesis from cepstral features," Proc. APSIPA ASC, pp.1697–1701, Nov. 2018.
[21] MathWorks, "MATLAB findpeaks," https://jp.mathworks.com/help/signal/ref/findpeaks.html.
[22] Scipy, "Scipy find_peaks," https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html.
[23] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," abs/1711.00354, Nov. 2017.
[24] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. AISTATS, Lauderdale, U.S.A., pp.315–323, April 2011.
[25] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. INTERSPEECH, Pittsburgh, U.S.A., pp.2266–2269, Sept. 2006.

Appendix: Detailed Investigation of Post-Filter Coefficients

In Sect. 5.2.3, we used 0.75 of the post-filter coefficient. To deeply discuss the effect of the coefficient in naturalness, we evaluated synthetic speech enhanced with a variety of coefficients in both 24 and 48 kHz of sampling rates. Changing the coefficient from 0.55 (over-emphasis) to 1.00 (non-emphasis), we conducted five-scale mean opinion score (MOS) tests on naturalness. 100 listeners participated in each test. Each listener answered to 30 speech samples.

Figure A· 1 and Fig. A· 2 show the results. First, the scores gently increase as the coefficient decreases from 1.00 to 0.75. Therefore, we can say that 0.75 we used in Sect. 5.2.3 is the best setting. However, over-emphasis setting, i.e., coefficients lower than 0.60, significantly degrades

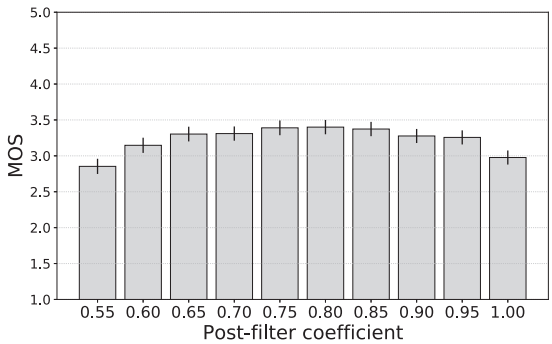


Fig. A· 1 MOS of naturalness of 24 kHz-sampled, post-filtered synthetic speech. Error bar indicates 95% confidence interval.

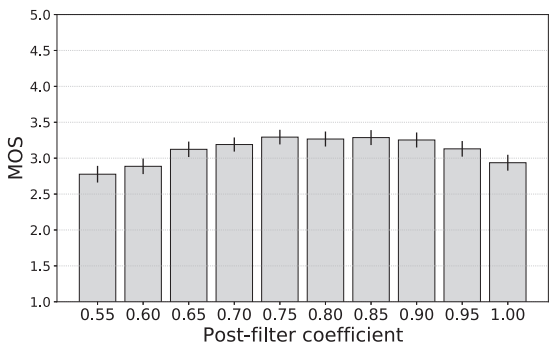
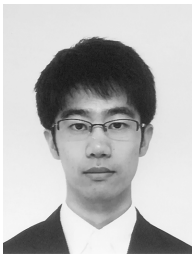


Fig. A· 2 MOS of naturalness of 48 kHz-sampled, post-filtered synthetic speech. Error bar indicates 95% confidence interval.

the naturalness. One reason of this observation is spectrum reduction at high frequency. As shown in Fig. 9, our post-filter enhances the spectrum peak but simultaneously reduces spectra at high frequency[†].



Junya Koguchi received the B.S. degree from Meiji University, Tokyo, Japan, in 2019. He was a assistant technical staff in 2019-2020. He is currently an M.E. student of Graduate School of Advanced Mathematical Sciences, Meiji University. His research interests include speech analysis/synthesis, singing voice synthesis and musical information retrival.

Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2016. He is currently an Assistant Professor at The University of Tokyo. He has received more than ten paper/achievement awards including the 3rd IEEE Signal Processing Society Japan Young Author Best Paper Award.

Masanori Morise received the Ph.D. degree in engineering from Wakayama University in 2008. He was a JSPS Research Fellow (DC1) in 2006-2008, a postdoctoral researcher at Kwansei Gakuin University in 2008-2009, an Assistant Professor at Ritsumeikan University in 2009-2013, a Project Assistant Professor at University of Yamanashi in 2013-2017, and an Associate Professor at University of Yamanashi in 2017-2019. He is currently an Associate Professor at Meiji University. His research interests include speech analysis/synthesis and speech design interface.

[†]This reduction also affects to spectra at low frequency, but the effect is less than high frequency. This is because the GMMs are placed at low frequency more dense than high frequency, as shown in Fig. 4.

Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor of The University of Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation (BSS), and speech enhancement. He has put his research into the world's first commercially available Independent-Component-Analysis based BSS microphone in 2007. He received paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE IROS2005 in 2006, and from APSIPA in 2013 and 2018. He received DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. He is an APSIPA Distinguished Lecturer from 2018.

Shigeki Sagayama was born in 1948. He received his B.E., M.S. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics. After spending 24 years with NTT Laboratories in Tokyo and Yokosuka, Japan, and ATR Interpreting Telephony Laboratories, Kyoto, Japan, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa. In 2000, he was appointed Professor at the University of Tokyo. He was a Professor of Meiji University from 2014 to 2019. His major research interests include processing, recognition and synthesis of speech, music, acoustic signals, handwriting, and images. Prof. Sagayama received the National Invention Award from the Institute of Invention of Japan in 1991, the Director General's Award from the Science and Technology Agency of Japan in 1996, and other academic awards. He is a fellow of IEICEJ, a life member of IEEE and a member of the ASJ (Acoustical Society of Japan) and IPSJ.