A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech

Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

Abstract

In statistical speech synthesis, the quality of the synthesized speech depends on the quality of training data. As the sampling rate of speech is one of the effective factors, speech data has been recently recorded at a high sampling rate. However, the sampling rates of speech data recorded in the past or collected from the Internet were often low. Therefore, to use these speech data effectively for model training, we propose a mel-cepstral analysis technique that restores missing high frequency components from low-sampling-rate speech with a statistical approach. In this technique, high-sampling-rate speech waveforms are modeled directly by integrating feature extraction and modeling processes. This framework makes it possible to optimize whole processes on the basis of an integrated objective function. Then, mel-cepstral coefficients are estimated from the low-sampling-rate speech by using the model as a prior distribution. Experimental results show that the proposed method improved the quality of synthesized speech.

Index Terms: integrative model, HMM-based speech synthesis, mel-cepstral analysis

1. Introduction

Statistical speech synthesis based on hidden Markov models (HMMs) has been proposed to enable machines to naturally speak like humans [1]–[3] and is widely used for TTS systems. In HMM-based speech synthesis, a spectral envelope, F0, and duration are modeled simultaneously on the basis of generative models. The quality of the synthesized speech strongly depends on the training data because HMM-based speech synthesis is a "corpus-based" method. The sampling rate of the training speech data is one of the factors that affect the quality of the synthesized speech. Although speech data has recently come to be recorded at a high sampling rate, e.g., 48 kHz, a lot of old speech data were recorded at a low sampling rate, e.g., 16 kHz. Furthermore, although some approaches that use speech data stored on the Internet as training data are becoming common, that kind of data is not always recorded at a high sampling rate. Low-sampling-rate speech data degrades the quality of the synthesized speech. However, recording voices and labeling them for a new speech database requires a huge cost. Thus, these low-sampling-rate speech databases should be used effectively.

Mel-cepstral coefficients are widely used as the spectral features, and low-sampling-rate speech data mainly affects the spectral features in HMM-based speech synthesis, We propose a mel-cepstral analysis technique that restores missing high frequency components from low-sampling-rate speech data by using a statistical method in the framework of the optimization integration. The idea of using the optimization integration has been seen in the construction of large scale systems, e.g., speech recognition systems [4], speech translation systems [5, 6], and spoken dialog systems [7]–[9]. For TTS systems, we proposed a technique for integrating feature extraction and acoustic modeling and optimizing them as an integrated generative model of speech waveforms for HMM-based speech synthesis [10]. Thus, optimization integration is an important trend for improving the performance of systems on the basis of statistical approaches. In this paper, speech waveforms are modeled directly as Gaussian mixture models (GMMs) by integrating feature extraction and modeling processes, and these processes are optimized on the basis of an integrated objective function. Then, mel-cepstral coefficients are estimated from the low-sampling-rate speech by using the GMMs as prior distributions.

The rest of this paper is organized as follows. Section 2 is a summary of the statical mel-cepstral analysis technique. In Section 3, the training algorithm of the integration model that represents speech waveforms directly and the restoring technique of high frequency components from a low-sampling-rate speech are derived. Difference from related work is discussed in Section 4, and experimental results are presented in Section 5. Concluding remarks and future plans are presented in the final section.

2. Mel-cepstral analysis

For spectral feature extraction, a statistical parametric mel-cepstral analysis [11, 12] has been widely used. In this method, mel-cepstral coefficients, i.e., frequency-transformed cepstral coefficients, are regarded as parameters of a generative model, and they are estimated by the maximum likelihood criterion on the basis of the likelihood of waveform domain.

A synthesis filter $H(z)$ is represented by mel-cepstral coefficients $\boldsymbol{c} = [c(0), \cdots, c(M-1)]^{\top}$ [1] defined as frequency-transformed cepstral coefficients:

$$H(z) = \exp \sum_{m=0}^{M-1} c(m)\, \tilde{z}^{-m}, \tag{1}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1, \tag{2}$$

where α is a frequency warping parameter. If $\alpha = 0$, mel-cepstral coefficients are equivalent to standard cepstral coefficients. If $\alpha > 0$, the system function defined as Eq. (1) has a high resolution at low frequencies, and if $\alpha < 0$, it has a high resolution at high frequencies.

For a given input signal, $\boldsymbol{x} = [x(0), \cdots, x(N-1)]^{\top}$, the mel-cepstral coefficients are determined by minimizing a spectral evaluation function with respect to $\boldsymbol{c}$ [13],

$$E(\boldsymbol{x}, \boldsymbol{c}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\exp R(\omega) - R(\omega) - 1\}\, d\omega, \tag{3}$$

[1] In section 2, $\boldsymbol{x}$ and $\boldsymbol{c}$ correspond to not an utterance but a frame. The frame index t is abbreviated.

where

$$R(\omega) = \log I_N(\omega) - \log\left|H\left(e^{j\omega}\right)\right|^2 \quad (4)$$

and ω denotes the angular frequency. The modified periodogram $I_N(\omega)$ of weakly stationary process $x(n)$ with a time window $w(n)$ of length N is represented as:

$$I_N(\omega) = \frac{\left|\sum_{n=0}^{N-1} w(n)\, x(n)\, e^{-j\omega n}\right|^2}{\sum_{n=0}^{N-1} w^2(n)}. \quad (5)$$

Mel-cepstral coefficients are determined easily by using an iterative algorithm, e.g., the Newton-Raphson method, because $E(\boldsymbol{x}, \boldsymbol{c})$ is convex with respect to $\boldsymbol{c}$.

There are some techniques for approximating time series signals by a zero-mean Gaussian process [14]. When $x(n)$ is assumed to be a zero-mean Gaussian process, the log likelihood can be approximated by:

$$\log P(\boldsymbol{x}|\boldsymbol{c}) \simeq -\frac{N}{2}\left[\log(2\pi) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\log\left|H\left(e^{j\omega}\right)\right|^2 + \frac{I_N(\omega)}{\left|H\left(e^{j\omega}\right)\right|^2}\right\}d\omega\right]. \quad (6)$$

Accordingly, the minimization of $E(\boldsymbol{x}, \boldsymbol{c})$ corresponds to the maximization of $P(\boldsymbol{x}|\boldsymbol{c})$. It should be noted that the spectral evaluation function of mel-cepstral analysis has the same form as that of LPC analysis [15]. Furthermore, taking the gain factor outside from $H\left(e^{j\omega}\right)$ indicates that the minimization of $E(\boldsymbol{x}, \boldsymbol{c})$ with respect to $\boldsymbol{c}$ is equivalent to both the minimization of residual energy and the maximization of the prediction gain. A mel-log spectrum approximation (MLSA) filter [16] is generally used to re-synthesize speech from mel-cepstral coefficients.

3. Mel-cepstral analysis restoring high frequency components

The goal of this paper is to estimate mel-cepstral coefficients that restores high frequency components from low-sampling-rate speech. To accomplish this goal, we employ statistical models of speech waveforms as prior distributions for mel-cepstral analysis. The proposed method consists of two parts, a modeling part and a restoring part. In the modeling part, speech waveforms are modeled directly as GMMs from high-sampling-rate speech waveforms. This modeling technique can be regarded as an application of the integration technique of acoustic modeling and mel-cepstral analysis for HMM-based speech synthesis [10], which we have already proposed as a technique for modeling speech waveforms. In the restoring part, they are used as prior distributions to estimate mel-cepstral coefficients from low-sampling-rate speech.

3.1. Technique for modeling speech waveforms

In the modeling part, speech waveforms $\boldsymbol{x}$ sampled at a high frequency are used to train the model. The model parameters $\tilde{\boldsymbol{\Lambda}}$ are estimated by maximizing the following likelihood,

$$\tilde{\boldsymbol{\Lambda}} = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{\Lambda}) = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}} \sum_{\forall \boldsymbol{h}} \int P(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{h}|\boldsymbol{\Lambda})\, d\boldsymbol{c}, \quad (7)$$

where $\boldsymbol{c}$ is a mel-cepstral coefficient sequence and $\boldsymbol{h}$ is a mixture index sequence of GMMs. To overcome the difficulty of the optimization of Eq. (7), a Q function is defined and maximized to estimate $\boldsymbol{\Lambda}$ by using the EM algorithm [17].

$$Q(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Lambda}}) = \sum_{\forall \boldsymbol{h}} \int Q(\boldsymbol{c}, \boldsymbol{h}) \log P(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{h}|\hat{\boldsymbol{\Lambda}})\, d\boldsymbol{c}, \quad (8)$$

where $Q(\boldsymbol{c}, \boldsymbol{h})$ is assumed as $Q(\boldsymbol{c}|\boldsymbol{h})Q(\boldsymbol{h})$ and the optimal posterior distributions are obtained by maximizing the objective Q function as:

$$Q(\boldsymbol{c}|\boldsymbol{h}) = \frac{1}{Z_{\boldsymbol{c}}} P(\boldsymbol{x}, \boldsymbol{c}|\boldsymbol{h}, \boldsymbol{\Lambda}), \quad (9)$$

$$Q(\boldsymbol{h}) = \frac{1}{Z_{\boldsymbol{h}}} P(\boldsymbol{h}|\boldsymbol{\Lambda}) \exp \int Q(\boldsymbol{c}|\boldsymbol{h}) \left(\log P(\boldsymbol{x}, \boldsymbol{c}|\boldsymbol{h}, \boldsymbol{\Lambda}) - \log Q(\boldsymbol{c}|\boldsymbol{h})\right) d\boldsymbol{c}, \quad (10)$$

where $Z_{\boldsymbol{c}}$ and $Z_{\boldsymbol{h}}$ are the normalization terms of $Q(\boldsymbol{c}|\boldsymbol{h})$ and $Q(\boldsymbol{h})$, respectively. These optimizations can be effectively performed by iterative calculations as the EM algorithm, which increases monotonically the value of the objective Q function at each iteration until convergence. Although the posterior distribution $Q(\boldsymbol{c}|\boldsymbol{h})$ should be ideally estimated with consideration for neighboring frames, it is estimated frame-by-frame to simplify the computation and reduce the computational complexity.

$$Q(\boldsymbol{c}|\boldsymbol{h}) = \prod_{t=1}^{T} Q(\boldsymbol{c}_t|h_t) \quad (11)$$

It is difficult to calculate the integral of $\boldsymbol{c}$ in Eq. (10) because of its high computational cost. Thus, $Q(\boldsymbol{c}_t|h_t)$ is assumed as a Gaussian probability distribution by using the Laplace approximation [18]. The posterior distribution $Q(\boldsymbol{c}_t|h_t)$ is represented by the unnormalized probability $Q^*(\boldsymbol{c}_t|h_t)$ as:

$$Q(\boldsymbol{c}_t|h_t) = \frac{1}{Z_{\boldsymbol{c}_t}} Q^*(\boldsymbol{c}_t|h_t), \quad (12)$$

where

$$Q^*(\boldsymbol{c}_t|h_t) = P(\boldsymbol{x}_t, \boldsymbol{c}_t|h_t, \boldsymbol{\Lambda}), \quad (13)$$

$$Z_{\boldsymbol{c}_t} = \int Q^*(\boldsymbol{c}'_t|h_t)\, d\boldsymbol{c}'_t. \quad (14)$$

Taking the first three terms of the Taylor series expansion around $\boldsymbol{c}_t = \tilde{\boldsymbol{c}}_t$, the logarithm of $Q^*(\boldsymbol{c}_t|h_t)$ then becomes:

$$\log Q^*(\boldsymbol{c}_t|h_t) \simeq \log Q^*(\tilde{\boldsymbol{c}}_t|h_t) + \left(\frac{\partial}{\partial \boldsymbol{c}_t} \log Q^*(\boldsymbol{c}_t|h_t)\,|_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t}\right) - \frac{1}{2}(\boldsymbol{c}_t - \tilde{\boldsymbol{c}}_t)^\top \left(\frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} \log Q^*(\boldsymbol{c}_t|h_t)\,|_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t}\right)(\boldsymbol{c}_t - \tilde{\boldsymbol{c}}_t), \quad (15)$$

where

$$\tilde{\boldsymbol{c}}_t = \underset{\boldsymbol{c}_t}{\operatorname{argmax}}\, Q(\boldsymbol{c}_t|h_t). \quad (16)$$

As the first derivation of $\log Q^*(\boldsymbol{c}_t|h_t)$ at $\tilde{\boldsymbol{c}}_t$ is equal to zero, Eq. (15) can be represented as:

$$\log Q^*(\boldsymbol{c}_t|h_t) \simeq \log Q^*(\tilde{\boldsymbol{c}}_t|h_t) - \frac{1}{2}(\boldsymbol{c}_t - \tilde{\boldsymbol{c}}_t)^\top \boldsymbol{A}_t (\boldsymbol{c}_t - \tilde{\boldsymbol{c}}_t), \quad (17)$$

$$
\begin{aligned}
\boldsymbol{A}_t &= -\frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^T} \log Q^*(\boldsymbol{c}_t|h_t) \mid_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t} \\
&= -\frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^T} \log P(\boldsymbol{x}_t|\boldsymbol{c}_t) \mid_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t} \\
&\quad -\frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^T} \log P(\boldsymbol{c}_t|h_t, \boldsymbol{\Lambda}) \mid_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t} \\
&= \frac{N}{2}\boldsymbol{H}_t \mid_{\boldsymbol{c}_t=\tilde{\boldsymbol{c}}_t} + \boldsymbol{\Sigma}_{h_t}^{-1}, \qquad (18)
\end{aligned}
$$

where $\boldsymbol{\Sigma}_{h_t}$ is the h_t-th covariance matrix of the GMMs, and $\boldsymbol{H}_t$ is the Hessian matrix of the spectral evaluation function $E(\boldsymbol{x}_t, \boldsymbol{c}_t)$ in Eq. (3) at time t:

$$
\boldsymbol{H}_t = \frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} E(\boldsymbol{x}_t, \boldsymbol{c}_t) = -\frac{2}{N}\frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} \log P(\boldsymbol{x}_t|\boldsymbol{c}_t). \quad (19)
$$

To approximate $Q(\boldsymbol{c}_t|h_t)$ by a Gaussian probability distribution, the normalization term $Z_{\boldsymbol{c}_t}$ is approximated as:

$$
Z_{\boldsymbol{c}_t} \simeq Q^*(\tilde{\boldsymbol{c}}_t|h_t)\sqrt{(2\pi)^M \left|\boldsymbol{A}_t^{-1}\right|}. \quad (20)
$$

By using the Laplace approximation, $Q(\boldsymbol{c}_t|h_t)$ is represented as:

$$
Q(\boldsymbol{c}_t|h_t) \simeq \mathcal{N}(\boldsymbol{c}_t|\tilde{\boldsymbol{c}}_t, \boldsymbol{A}_t^{-1}). \quad (21)
$$

From the above, the posterior distribution $Q(\boldsymbol{c}, \boldsymbol{h})$ can be calculated.

3.2. Technique for restoring high frequency components

In the restoring part, the mel-cepstral coefficients $\tilde{\boldsymbol{c}}$ with the high frequency components restored from the low-sampling-rate speech waveform $\boldsymbol{x}^{(L)}$ and the model parameter $\boldsymbol{\Lambda}$ are estimated by maximizing the posterior probability for the given speech waveform $\boldsymbol{x}_L$ as follows:

$$
\begin{aligned}
\tilde{\boldsymbol{c}} &= \operatorname*{argmax}_{\boldsymbol{c}} P(\boldsymbol{c}|\boldsymbol{x}^{(L)}, \boldsymbol{\Lambda}) \\
&= \operatorname*{argmax}_{\boldsymbol{c}} P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) P(\boldsymbol{c}|\boldsymbol{\Lambda}) \\
&= \operatorname*{argmax}_{\boldsymbol{c}} \Big\{ \log P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) + \log \sum_{\forall \boldsymbol{h}} P(\boldsymbol{c}, \boldsymbol{h}|\boldsymbol{\Lambda}) \Big\} \quad (22)
\end{aligned}
$$

The probability $P(\boldsymbol{c}|\boldsymbol{\Lambda})$ of mel-cepstral coefficients is expected to work as the prior distribution of mel-cepstral coefficients. When $\boldsymbol{c}$ is estimated by maximizing only $P(\boldsymbol{x}^{(L)}|\boldsymbol{c})$, the high frequency components of the spectral envelope from the estimated $\boldsymbol{c}$ are not always appropriate because high frequency components cannot be considered in $P(\boldsymbol{x}^{(L)}|\boldsymbol{c})$. However, $P(\boldsymbol{c}|\boldsymbol{\Lambda})$ leads the high frequency components of the spectral envelope to the reasonable curve. The probability $P(\boldsymbol{x}^{(L)}|\boldsymbol{c})$ of speech waveforms is calculated from the low-sampling-rate periodogram. If the log likelihood function of the partial periodgram from l_1-th to l_2-th dimension is defined as:

$$
\begin{aligned}
D(l_1, l_2) = -\frac{1}{2}\Bigg\{ & (l_2 - l_1 + 1)\log(2\pi) \\
& + \sum_{i=l_1}^{l_2} \left(\log \left| H\left(e^{j\omega_i}\right)\right|^2 + \frac{I_N(\omega_i)}{\left|H\left(e^{j\omega_i}\right)\right|^2} \right) \Bigg\}, \quad (23)
\end{aligned}
$$

the original log likelihood function is represented by

$$
\begin{aligned}
\log P(\boldsymbol{x}_t|\boldsymbol{c}_t) &= D(0, N-1) \\
&= D(0, \tilde{N}-1) + D(\tilde{N}, N-1) \\
&= \log P(\boldsymbol{x}_t^{(L)}|\boldsymbol{c}_t) + \log P(\boldsymbol{x}_t^{(H)}|\boldsymbol{c}), \quad (24)
\end{aligned}
$$

where $\boldsymbol{x}_t^{(L)}$ and $\boldsymbol{x}_t^{(H)}$ are the low and high frequency components of a speech waveform, and $\tilde{N}$ is a dimension of the boundary between them. The likelihood of the low and high frequency components can be calculated separately.

Equation (22) is converted by using Jensen's inequality:

$$
\begin{aligned}
& \log P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) + \log \sum_{\forall \boldsymbol{h}} P(\boldsymbol{c}, \boldsymbol{h}|\boldsymbol{\Lambda}) \\
& \geq \log P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) + \sum_{\forall \boldsymbol{h}} Q'(\boldsymbol{h}) \log \frac{P(\boldsymbol{c}, \boldsymbol{h}|\boldsymbol{\Lambda})}{Q'(\boldsymbol{h})}, \quad (25)
\end{aligned}
$$

where

$$
\begin{aligned}
Q'(\boldsymbol{h}) &= P(\boldsymbol{h}|\boldsymbol{c}, \boldsymbol{\Lambda}) \\
&= \frac{P(\boldsymbol{c}, \boldsymbol{h}|\boldsymbol{\Lambda})}{\sum_{\forall \boldsymbol{h}'} P(\boldsymbol{c}, \boldsymbol{h}'|\boldsymbol{\Lambda})}. \qquad (26)
\end{aligned}
$$

To maximize $P(\boldsymbol{c}|\boldsymbol{x}^{(L)}, \boldsymbol{\Lambda})$, $\tilde{\boldsymbol{c}}$ and $Q'(\boldsymbol{h})$ are updated alternately. The mel-cepstral coefficients $\tilde{\boldsymbol{c}}$ can be estimated by using an optimization algorithm such as Rprop [19].

3.3. Avoidance of local maxima problem

The estimated mel-cepstral coefficients $\boldsymbol{c}$ depend heavily on the initial value. To overcome the serious local maxima problem, an annealing technique hardly depending on the initial value is used. It is similar to the deterministic annealing EM (DAEM) algorithm [20]. Two terms related to $\boldsymbol{c}$ in Eq. (25) are shown as:

$$
\mathcal{F} = \log P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) + \log P(\boldsymbol{c}|\boldsymbol{\Lambda}). \quad (27)
$$

It is modified by using a parameter β that decides the ratio between two terms.

$$
\mathcal{F}_\beta = \beta \log P(\boldsymbol{x}^{(L)}|\boldsymbol{c}) + (2-\beta)\log P(\boldsymbol{c}|\boldsymbol{\Lambda}). \quad (28)
$$

If $\beta = 1$, $\mathcal{F}_\beta$ becomes equal to the original objective function. The parameter β is gradually changed in the estimation of $\tilde{\boldsymbol{c}}$ according to the following function.

$$
\beta = \left(\frac{s}{S}\right)^r \quad (s = 1, 2, \cdots, S), \quad (29)
$$

where s denotes the iteration number of updates.

4. Related work

As mentioned above, the proposed method restores missing high frequency components from low-sampling-rate speech. Some similar approaches have been found in previous pieces of research. One famous method converts low-sampling-rate speech into high-sampling-rate speech by using the voice conversion (VC) method [21, 22]. In the VC-based method, the feature extraction and restoration of the high frequency components are independent. Furthermore, as the trained model depends on the sampling rate of input speech, different models are required for different sampling rates of input speech. In contrast to the VC-based methods, the feature extraction and the restoration of the high frequency components are integrated and optimized on the basis of the unified criterion in the proposed method. Also, as the sampling rate of the input speech does not depend on the model, only one model is required for any sampling-rate of input speech.

5. Experiments

To evaluate the effectiveness of the proposed method, subjective comparison tests on the mean opinion score (MOS) for the analysis-synthesis and HMM-based speech synthesis were conducted. For the speech database, 503 phonetically balanced sentences from the ATR Japanese speech database (Set B) [23] uttered by a male speaker were used. The following three methods were compared in the evaluation.

- **48 kHz (original)**: Use mel-cepstrum extracted from original 48-kHz sampling-rate speech.
- **Conventional**: Use mel-cepstrum converted from a sampling rate of 16 kHz to that of 48 kHz in the mel-cepstrum domain by using the VC-based method. The joint feature vectors of the mel-cepstral coefficients of the 16 kHz and 48-kHz sampling rates were modeled as GMMs. The number of mixture components of GMMs was set to 64, and each distribution was modeled with a cross covariance matrix.
- **Proposed**: Use mel-cepstrum estimated from 16-kHz sampling-rate speech by the proposed method. To train GMMs to restore the high frequency components, speech waveforms recorded at a sampling rate of 48 kHz were used. The numbers of mixture components of GMMs were set to 64. The output probability distribution was modeled with a diagonal covariance matrix. The parameter r in Eq. (29) was varied as $r = 2^n$ and decided to $r = 2^{-3}$ which obtained the best likelihood for the test data.

5.1. Experiments of analysis-synthesis

In this experiment, mel-cepstral coefficients were estimated by using the above three methods, and 48-kHz sampling-rate speech waveforms were reconstructed from them. For the conventional and proposed methods, 200 sentences were used for training models. The speech data was windowed at a frame rate of 5 ms by using a 25-ms Hamming window. The windowed waveforms were used as the input data for training GMMs and restoring high frequency components in the proposed method, and 35 mel-cepstral coefficients including the zero coefficient, which are estimated with the standard mel-cepstral analysis technique, were used for other methods. The dimension of the hidden mel-cepstral coefficients of the proposed method was set to the same as that of the other methods. The frequency warping parameter α was set to 0.55. The evaluation data was prepared by downsampling each speech waveform from the 48-kHz sampling rate to the 16-kHz sampling rate. For the conventional method, mel-cepstral coefficients estimated from the 16-kHz sampling-rate speech were used as the input of the conversion process. Speech Signal Processing Toolkit (SPTK) [24] was used for downsampling. The other 53 sentences were used for evaluation. Ten subjects were asked to rate the naturalness of the synthesized speech on a MOS with a scale from 1 (poor) to 5 (good). Ten randomly selected sentences were presented to each subject. The experiments were carried out in a sound-proof room.

Figure 1 shows the results of MOS evaluation for analysis-synthesis speech. The proposed method obtained a significant improvement compared with the conventional method. The score of the proposed method was almost the same as that of the original 48-kHz one. Thus, the proposed method seems to be able to restore the missing high frequency components.

5.2. Experiments of HMM-based speech synthesis

Next, speech synthesized by HMM-based speech synthesis was evaluated. To train HMMs, 250 sentences not included in the training data of GMMs were used. Mel-cepstral coefficients of these sentences were prepared by using the above three methods. A five-state, left-to-right, no-skip structure was used for the HMMs. The excitation parameters were modeled with multi-space probability distribution HMMs [25]. Each state output probability distribution was modeled by using a single Gaussian distribution with a diagonal covariance matrix. The HMMs were estimated as context-dependent models [26] and applied the decision tree based context clustering technique [27]. The minimum description length (MDL) criterion was used to determine the size of the decision trees [28]. Each probability distribution was modeled with a diagonal covariance matrix. The setting of the MOS evaluation was the same as that of analysis-synthesis.

Figure 2 shows the results of MOS evaluation for speech synthesized by the HMM-based speech synthesis. The trend of the results was almost the same as that of analysis-synthesis. Thus, the effectiveness of the proposed method for HMM-based speech synthesis was shown.

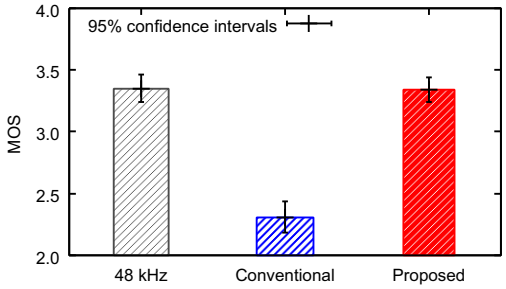

Figure 1: Mean opinion scores for analysis-synthesis speech

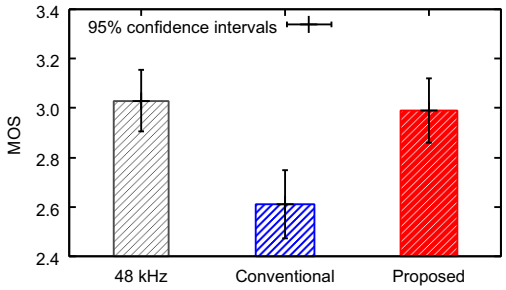

Figure 2: Mean opinion scores for speech synthesized by HMM-based speech synthesis

6. Conclusions

In this paper, a mel-cepstral analysis technique restoring missing high frequency components from low-sampling-rate speech was proposed. The feature extraction process and the modeling process of these features were integrated, and the models of speech waveforms were used as the prior models to restore the high frequency components. In subjective experiments, the naturalness of synthesized speech was significantly improved by using the proposed method. Future work includes objective evaluations and experiments with speaker-independent models.

7. Acknowledgements

The research leading to these results was partly funded by the Core Research for Evolutionary Science and Technology (CREST) program from the Japan Science and Technology Agency (JST).

8. References

[1] T. Masuko, K. Tokuda, T. Kobayashi and, S. Imai, "Speech synthesis from HMMs using dynamic features," Proceedings of ICASSP, pp. 389–392, 1996.

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

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proceedings of ICASSP, pp. 1315–1318, 2000.

[4] J. T. Chien and C. H. Chueh, "Joint acoustic and language modeling for speech recognition," Speech Communication, vol. 52, Issue 3, pp. 223–235, 2010.

[5] A. Parlikar, A. Black, and S. Vogel, "Improving speech synthesis of machine translation output," Proceedings of Interspeech, pp. 194–197, 2010.

[6] K. Hashimoto, J. Yamagishi, W. Byrne, S. King, and K. Tokuda, "Impacts of machine translation and speech synthesis on speech-to-speech translation," Speech Communication, vol. 54, Issue 7, pp. 854–866, 2012.

[7] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple target using weighted finite state transducers," Computer Speech and Language, vol. 16, pp. 533–550, 2002.

[8] C. Nakatsu and M. White, "Reranking realizations by predicted synthesis quality," Proceedings of ACL, pp. 1113–1120, 2006.

[9] C. Boidin, V. Rieser, L. Plas, O. Lemon, and J. Chevelu, "Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems," Proceedings of Interspeech, pp. 2487–2490, 2009.

[10] K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Integration of acoustic modeling and mel-cepstral analysis for HMM-based speech synthesis," Proceedings of ICASSP, pp. 7883–7887, 2013.

[11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proceedings of ICASSP, vol. 1, pp. 137–140, 1992.

[12] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generated cepstral analysis - a unified approach to speech spectral estimation," Proceedings of ICSLP, pp. 1043–1045, 1994.

[13] S. Imai and C. Furuichi, "Unbiased estimator of log spectrum and its application to speech signal processing," Proceedings of EURASIP, pp. 203–206, 1988.

[14] K. Dzhaparidze, "Parameter estimation and hypothesis testing in spectral analysis of stationary time series," New York: Springer-Verlag, 1986.

[15] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," IECE Transactions on Fundamentals (Japanese Edition), vol.J53-A, no.1, pp35–42, Jan. 1970. Translation: R.W. Schafer and J.D. Markel, eds., Speech Analysis, pp.295–302, IEEE Press, New York, 1979.

[16] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectral approximation filter for speech synthesis," IECE Translations on Fundamentals (Japanese Edition), vol. J66-A, pp. 122–129, Feb. 1983.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," J. Royal Statist. Soc., Ser. B, 39, pp. 1-38, 1977.

[18] P. S. Laplace, "Memoir on the probability of the causes of events," Statistical Science, pp. 364–378, 1986.

[19] M. Riedmiller, "Rprop - Description and implementation details," Technical Report, University of Karlsruhe, 1994.

[20] N. Ueda, R. Nakano, "Deterministic annealing EM algorithm," Neural Networks, vol.11, pp.271–282, Mar. 1998.

[21] K.-H. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," Proceedings of ICASSP, vol. 3, pp. 1843–1846, 2000.

[22] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," Proceedings of NCSP, pp. 283–286, 2008.

[23] A. Kuramatsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kawabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[24] "Speech Signal Processing Toolkit (SPTK)," http://sp-tk.sourceforge.net/

[25] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proceedings of ICASSP, pp. 229–232, 1999.

[26] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 4, pp. 599–609, 1990.

[27] S. Young, J. J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," Proceedings of ARPA Workshop on Human Language Technology, pp. 307–312, 1994.

[28] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proceedings of Eurospeech, pp. 99–102, 1997.