PAPER *Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application*

Integration of Spectral Feature Extraction and Modeling for HMM-Based Speech Synthesis

Kazuhiro NAKAMURA[†a)], **Kei HASHIMOTO**[†], *Nonmembers*, **Yoshihiko NANKAKU**[†], *and* **Keiichi TOKUDA**[†], *Members*

SUMMARY This paper proposes a novel approach for integrating spectral feature extraction and acoustic modeling in hidden Markov model (HMM) based speech synthesis. The statistical modeling process of speech waveforms is typically divided into two component modules: the frame-by-frame feature extraction module and the acoustic modeling module. In the feature extraction module, the statistical mel-cepstral analysis technique has been used and the objective function is the likelihood of mel-cepstral coefficients for given speech waveforms. In the acoustic modeling module, the objective function is the likelihood of model parameters for given mel-cepstral coefficients. It is important to improve the performance of each component module for achieving higher quality synthesized speech. However, the final objective of speech synthesis systems is to generate natural speech waveforms from given texts, and the improvement of each component module does not always lead to the improvement of the quality of synthesized speech. Therefore, ideally all objective functions should be optimized based on an integrated criterion which well represents subjective speech quality of human perception. In this paper, we propose an approach to model speech waveforms directly and optimize the final objective function. Experimental results show that the proposed method outperformed the conventional methods in objective and subjective measures.

key words: *integrative model, HMM-based speech synthesis, acoustic modeling, mel-cepstral analysis, trajectory HMM*

1. Introduction

Statistical speech synthesis based on HMMs has been proposed to enable machines to naturally speak like humans [1]–[3] and widely used for TTS systems. In this technique, spectral and F0 features are extracted from speech waveforms and modeled by statistical techniques [2]. In general, a TTS system consists of several component modules, e.g., text analysis, spectral estimation, F0 estimation and acoustic modeling, that are usually optimized independently each other. It is important to improve the performance of each component module for achieving higher quality synthesized speech. However, the final objective of TTS systems is to generate natural speech waveforms from given texts, and the improvement of each component module does not always lead to the improvement of the quality of synthesized speech. Therefore, ideally all component modules should be optimized based on an integrated criterion which well represents subjective speech quality of human perception. A similar idea using the optimization integration has been seen in the construction of large scale systems, e.g., acoustic and language models of speech recognition systems [4], speech translation systems [5], [6] and spoken dialog systems [7]–[9]. In TTS systems, an approach integrating text analysis and acoustic modeling modules has been proposed [10]. By integrating linguistic and acoustic models, it became robust against text analysis errors and improved the quality of synthesized speech. Thus, the optimization integration is an important trend for improving the performance of systems based on statistical approaches.

In this paper, we integrate the feature extraction and the acoustic modeling of HMM-based TTS systems. These modules are typically connected in series and optimized independently. We optimize them as an integrated generative model of speech waveforms. As the component modules of feature extraction and acoustic modeling, statistical generative model-based approaches that are suitable for the integration have already been proposed and employed in HMM-based speech synthesis. For 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 based on the likelihood of waveform domain. For the acoustic modeling, "trajectory HMM" [15]–[17] has been proposed as a generative model of static features considering the temporal continuity of feature sequences. It is well known that an acoustic modeling technique considering the temporal continuity of each feature sequence improves the quality of synthesized speech [1]. In the standard HMM, dynamic features calculated from extracted static features are typically modeled with static features. However, as the proposed method requires a generative model of only static features, the trajectory HMM should be used. We integrate the statistical mel-cepstral analysis and the trajectory HMM and redefine as a generative model.

The rest of this paper is organized as follows. Section 2 summarizes HMM-based speech synthesis, including the mel-cepstral analysis and the trajectory HMMs. In Sect. 3, the integration method of the mel-cepstral analysis and the acoustic modeling is derived. Experimental results are presented in Sect. 4. Concluding remarks and future plans are presented in the final section.

Manuscript received September 27, 2013.

Manuscript revised January 18, 2014.

[†]The authors are with the Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

a) E-mail: nkazu@sp.nitech.ac.jp

DOI: 10.1587/transinf.E97.D.1438

2. HMM-Based Speech Synthesis

In HMM-based TTS systems, spectral envelope, F0, and duration are modeled simultaneously based on generative models, i.e., MSD-HSMM (Multi-Space Probability Distribution Hidden Semi-Markov Models). However, this paper focuses only on the spectral modeling based on the standard HMMs (or trajectory HMMs). When a target text is given to the TTS system, the spectral parameter sequence is generated from HMMs, and a speech waveform is finally synthesized from them via the source-filter based production model. In the training process, the spectral feature extraction followed by the training HMMs is firstly performed. The statistical mel-cepstral analysis [11], [12] which regards mel-cepstral coefficients as the model parameters is widely used in the standard HMM-based TTS systems, and the mel-cepstral coefficients are estimated from a given input signal $\boldsymbol{x}$ in the maximum likelihood (ML) sense:

$$\hat{\boldsymbol{c}}_t = \underset{\boldsymbol{c}_t}{\operatorname{argmax}}\, P\left(\boldsymbol{x}_t | \boldsymbol{c}_t\right) \tag{1}$$

The training of HMMs using extracted mel-cepstrum sequences $\boldsymbol{c} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T)$ is also performed based on the ML criterion

$$\hat{\boldsymbol{\Lambda}} = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, P\left(\boldsymbol{c} | \boldsymbol{w}, \boldsymbol{\Lambda}\right) \tag{2}$$

where $\hat{\boldsymbol{\Lambda}}$ is a set of the model parameters of HMMs and $\boldsymbol{w}$ is a text corresponding to the training data ($\boldsymbol{w}$ is omitted in the following formulas for simplicity). In this paper, trajectory HMMs are used for acoustic modeling instead of standard HMMs, because the standard HMMs generate step-wise parameter sequences with discontinuity at state boundaries due to the shortcoming of model structures while training HMMs. To overcome this problem, the consistency between static and dynamic features that causes the smooth trajectory is considered in the spectral parameter generation. In the rest of this section, the mel-cepstral analysis and trajectory HMMs will be briefly reviewed.

2.1 Mel-Cepstral Analysis

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

$$H(z) = \exp \sum_{m=0}^{M-1} c(m)\, \tilde{z}^{-m} \tag{3}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{4}$$

where α is a frequency warping parameter. If $\alpha = 0$, mel-cepstral coefficients are equivalent to standard cepstral coefficients. Figure 1 shows the frequency warping function with varying α. The vertical axis gives the warped frequencies. If $\alpha > 0$, the system function defined as Eq. (3) has a high resolution at low frequencies, and if $\alpha < 0$, it has a high resolution at high frequencies.

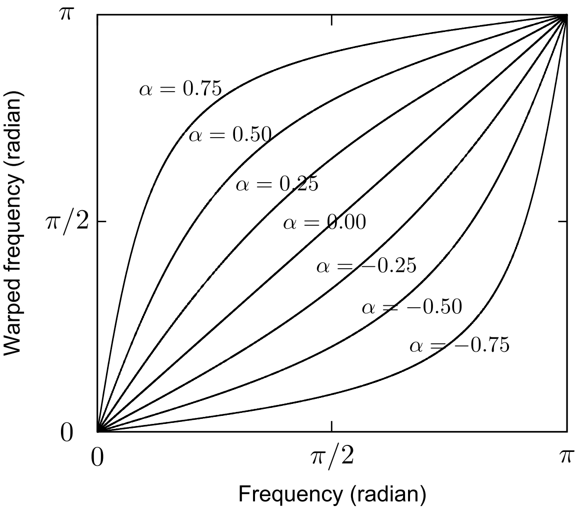


Fig. 1 Frequency warping function.

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 $\boldsymbol{c}$ [24],

$$E(\boldsymbol{x}, \boldsymbol{c}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{\exp R(\omega) - R(\omega) - 1\right\} d\omega \tag{5}$$

where

$$R(\omega) = \log I_N(\omega) - \log \left| H\left(e^{j\omega}\right) \right|^2 \tag{6}$$

and $I_N(\omega)$ is the modified periodogram of weakly stationary process $x(n)$ with a time window $w(n)$ of length N:

$$I_N(\omega) = \frac{\left| \sum_{n=0}^{N-1} w(n)\, x(n)\, e^{-j\omega n} \right|^2}{\sum_{n=0}^{N-1} w^2(n)} \tag{7}$$

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 $\boldsymbol{c}$.

When $x(n)$ is assumed to be a zero-mean Gaussian process, the log likelihood can be approximated by

$$\begin{aligned} \log P(\boldsymbol{x}|\boldsymbol{c}) \simeq -\frac{N}{2}\Bigg[& \log(2\pi) \\ & + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \log \left| H\left(e^{j\omega}\right) \right|^2 + \frac{I_N(\omega)}{\left| H\left(e^{j\omega}\right) \right|^2} \right\} d\omega \Bigg] \end{aligned} \tag{8}$$

There are some techniques to approximate time series signals by a zero-mean Gaussian process [25]. The approximation used in this paper is shown in Appendix. 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 [26]. Furthermore, taking the gain

[†] In Sect. 2.1, $\boldsymbol{x}$ and $\boldsymbol{c}$ correspond to not an utterance but a frame. The frame index t is abbreviated.

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 minimization of residual energy and maximization of the prediction gain. Mel-log spectrum approximation (MLSA) filter [27] is generally used to re-synthesize speech from the mel-cepstral coefficients.

2.2 Trajectory HMM

In HMM-based speech synthesis systems, observation vector sequences are quasi-stationary and each stationary part is represented by a state of the HMMs. The statistics of each state do not change dynamically, and intra-state time-dependency cannot be represented. Therefore, a technique that augments the dimensionality of an acoustic static feature vector by appending its dynamic feature vectors is widely used. The standard HMMs with static and dynamic features are improper in the sense of statistical modeling because they model the static and dynamic features independently. By imposing the explicit relationship between them, the standard HMMs are naturally translated into trajectory HMMs. The trajectory HMMs can overcome the impropriety in the standard HMM framework without any additional parameters, and be a consistent generative model of the static feature sequences.

Let a spectral feature vector sequence be $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$, where $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top\right]^\top$ includes not only static but also dynamic features. Mel-cepstral coefficients $\boldsymbol{c}_t$ are a M dimensional vector, and T is the number of frames. In the standard model, the probability density of $\boldsymbol{o}$ is shown as $P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\Lambda})$ and assumed as a Gaussian distribution, where $\boldsymbol{q} = (q_1, q_2, \cdots, q_T)$ is a state sequence of HMMs. By imposing an explicit relationship between static and dynamic features, which is given by $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{W}$ is a $3MT \times MT$ window matrix as shown in Fig. 2, the standard HMM is reformed as the trajectory HMM as:

$$P(\boldsymbol{c}|\boldsymbol{\Lambda}) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\Lambda})\, P(\boldsymbol{q}|\boldsymbol{\Lambda}) \tag{9}$$

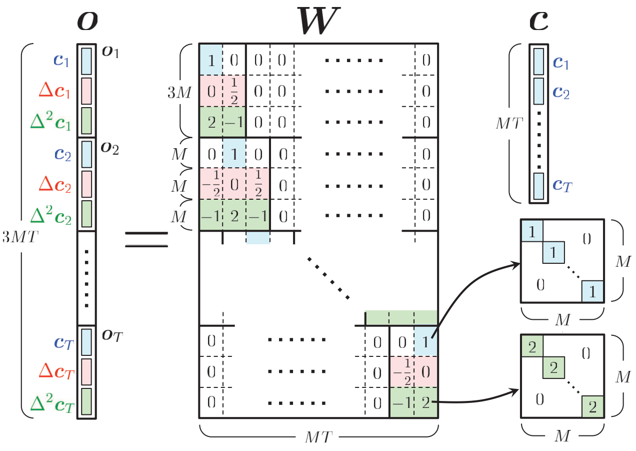


Fig. 2 Example of the relationship between the static feature vector sequence $\boldsymbol{c}$ and the speech parameter vector sequence $\boldsymbol{o}$ in a matrix form.

$$P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{c}|\bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right) = \frac{1}{Z} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\Lambda}) \tag{10}$$

$$P(\boldsymbol{q}|\boldsymbol{\Lambda}) = P(q_1|\boldsymbol{\Lambda}) \prod_{t=2}^{t} P(q_t|q_{t-1}, \boldsymbol{\Lambda}) \tag{11}$$

where Z is a normalization term. In Eq. (10), $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ and $\boldsymbol{P}_{\boldsymbol{q}}$ are the $MT \times 1$ mean vector and the $MT \times MT$ covariance matrix given by $\boldsymbol{q}$, respectively. They are represented as:

$$Z = \frac{\sqrt{(2\pi)^{MT}\left|\boldsymbol{P}_{\boldsymbol{q}}\right|}}{\sqrt{(2\pi)^{3MT}\left|\boldsymbol{\Sigma}_{\boldsymbol{q}}\right|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{\boldsymbol{q}}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}}\right)\right\} \tag{12}$$

$$\boldsymbol{R}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}} \tag{13}$$

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1} \tag{14}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} \tag{15}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \cdots, \boldsymbol{\mu}_{q_T}^\top\right]^\top \tag{16}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{q}} = \operatorname{diag}\left[\boldsymbol{\Sigma}_{q_1}^\top, \cdots, \boldsymbol{\Sigma}_{q_T}^\top\right]^\top \tag{17}$$

where $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$ are the $3M \times 1$ mean vector and the $3M \times 3M$ covariance matrix associated with the state q_t, respectively. The elements of $\boldsymbol{W}$ are given as regression window coefficients to calculate delta and delta-delta features as follows:

$$\Delta^d \boldsymbol{c}_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)\, \boldsymbol{c}_{t+\tau}, \quad d = 1, 2 \tag{18}$$

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{M\times M} \tag{19}$$

$$\boldsymbol{W}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right] \tag{20}$$

$$\begin{aligned} \boldsymbol{w}_t^{(d)} = \Big[&\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(0), \\ &\ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-\left(t+L_+^{(d)}\right)}\Big]^\top, d = 0, 1, 2 \end{aligned} \tag{21}$$

where $L_-^{(0)} = L_+^{(0)} = 0$, $\boldsymbol{w}^{(0)} = 1$, and $\otimes$ denotes the Kronecker product for matrices.

Note that $\boldsymbol{c}$ is modeled by a Gaussian distribution whose dimensionality is MT, and the covariance matrices $\boldsymbol{P}_{\boldsymbol{q}}$ are generally full. As a result, the trajectory HMMs can overcome the drawback of the HMMs. It is also noted that the parameterization of the trajectory HMMs is completely the same as that of the HMMs with the same model topology.

3. Integration of Acoustic Modeling and Mel-Cepstral Analysis

In the conventional method, the statistical modeling processes for feature extraction and acoustic modeling are connected in series. However, the essential problem of constructing TTS systems is to comprehensively estimate models that can generate speech waveforms from texts. In this paper, we propose a technique to directly model speech

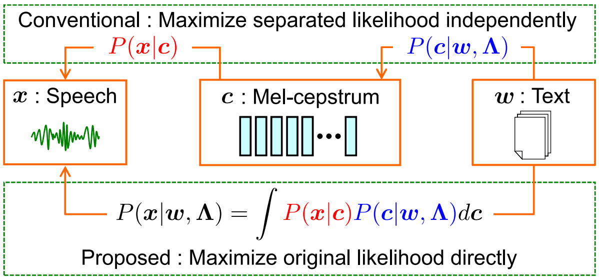

Conventional : Maximize separated likelihood independently
$P(\boldsymbol{x}|\boldsymbol{c})$
$P(\boldsymbol{c}|\boldsymbol{w}, \boldsymbol{\Lambda})$
$\boldsymbol{x}$: Speech
$\boldsymbol{c}$: Mel-cepstrum
$\boldsymbol{w}$: Text
$P(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{\Lambda}) = \int P(\boldsymbol{x}|\boldsymbol{c})P(\boldsymbol{c}|\boldsymbol{w}, \boldsymbol{\Lambda})d\boldsymbol{c}$
Proposed : Maximize original likelihood directly

Fig. 3 Basic idea of the proposed approach.

waveforms as a statistical model. The statistical mel-cepstral model $P(\boldsymbol{x}|\boldsymbol{c})$ and the statistical acoustic model $P(\boldsymbol{c}|\boldsymbol{\Lambda})$ are integrated as:

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{\Lambda}) &= \int P(\boldsymbol{x}, \boldsymbol{c}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &= \int P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \qquad (22) \end{aligned}$$

The original point of this model structure is that two statistical modeling processes are connected with the marginalization of mel-cepstral coefficients, and the proposed model is a generative model of speech waveforms. Figure 3 shows the generative process. In the conventional model structure, there is the strong constraint that only one mel-cepstral sequence is used to convey useful information from the feature extraction module to the acoustic modeling module. As the proposed method can avoid this constraint, we expect that the proposed method improve the quality of synthesized speech.

In the standard mel-cepstral analysis technique, mel-cepstral coefficients are estimated frame-by-frame. However, it is well known that considering the temporal continuity of mel-cepstral coefficients improves the quality of synthesized speech. Thus, we use the trajectory HMM to consider the temporal continuity as a statistical model of mel-cepstral coefficients.

To train the proposed model, a lower bound of log marginal likelihood $\mathcal{F}$ is maximized instead of the true likelihood. The lower bound $\mathcal{F}$ is defined by using Jensen's inequality:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}|\boldsymbol{\Lambda}) &= \log P(\boldsymbol{x}|\boldsymbol{\Lambda}) \\ &= \log \sum_{\forall \boldsymbol{q}} \int P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}, \boldsymbol{q}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &= \log \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c}, \boldsymbol{q}) \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}, \boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c}, \boldsymbol{q})} d\boldsymbol{c} \\ &= \log \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c})\, Q(\boldsymbol{q}) \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}, \boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c})\, Q(\boldsymbol{q})} d\boldsymbol{c} \\ &\geq \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c})\, Q(\boldsymbol{q}) \log \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}, \boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c})\, Q(\boldsymbol{q})} d\boldsymbol{c} \\ &= \mathcal{F} \qquad (23) \end{aligned}$$

To overcome the difficulty of optimization, it is assumed that $\boldsymbol{c}$ and $\boldsymbol{q}$ are conditionally independent. The optimal posterior distributions can be obtained by maximizing the objective function $\mathcal{F}$ with the variational method [13] as:

$$Q(\boldsymbol{c}) = \frac{1}{Z_{\boldsymbol{c}}} P(\boldsymbol{x}|\boldsymbol{c}) \exp \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \log P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\Lambda}) \qquad (24)$$

$$Q(\boldsymbol{q}) = \frac{1}{Z_{\boldsymbol{q}}} P(\boldsymbol{q}|\boldsymbol{\Lambda}) \exp \int Q(\boldsymbol{c}) \log P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\Lambda})\, d\boldsymbol{c} \qquad (25)$$

where $Z_{\boldsymbol{c}}$ and $Z_{\boldsymbol{q}}$ are the normalization terms of $Q(\boldsymbol{c})$ and $Q(\boldsymbol{q})$, respectively.

$$Z_{\boldsymbol{c}} = \int P(\boldsymbol{x}|\boldsymbol{c}') \exp \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \log P(\boldsymbol{c}'|\boldsymbol{q}, \boldsymbol{\Lambda})\, d\boldsymbol{c}' \qquad (26)$$

$$Z_{\boldsymbol{q}} = \sum_{\forall \boldsymbol{q}'} P(\boldsymbol{q}'|\boldsymbol{\Lambda}) \exp \int Q(\boldsymbol{c}) \log P(\boldsymbol{c}|\boldsymbol{q}', \boldsymbol{\Lambda})\, d\boldsymbol{c} \qquad (27)$$

These optimizations can be effectively performed by iterative calculations as the Expectation and Maximization (EM) algorithm, which increases monotonically the value of objective function $\mathcal{F}$ at each iteration until convergence.

3.1 Posterior Probabilities of Mel-Cepstral Coefficients

It is difficult to calculate the integral of $\boldsymbol{c}$ in Eq. (25) because of its high computational cost. Therefore, $Q(\boldsymbol{c})$ is assumed as a Gaussian probability distribution by using the Laplace approximation [14]. The unnormalized probability in $Q(\boldsymbol{c})$ is defined by $Q^*(\boldsymbol{c})$ as:

$$Q^*(\boldsymbol{c}) = P(\boldsymbol{x}|\boldsymbol{c}) \exp \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \log P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\Lambda}) \qquad (28)$$

Taking the first three terms of the Taylor series expansion around $\boldsymbol{c} = \tilde{\boldsymbol{c}}$ then the logarithm of Eq. (28) becomes:

$$\begin{aligned} \log Q^*(\boldsymbol{c}) \simeq \log Q^*(\tilde{\boldsymbol{c}}) &+ \left(\frac{\partial}{\partial \boldsymbol{c}} \log Q^*(\boldsymbol{c}) \,|_{\boldsymbol{c}=\tilde{\boldsymbol{c}}} \right) (\boldsymbol{c} - \tilde{\boldsymbol{c}}) \\ &+ \frac{1}{2} (\boldsymbol{c} - \tilde{\boldsymbol{c}})^\top \left(\frac{\partial^2}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top} \log Q^*(\boldsymbol{c}) \,|_{\boldsymbol{c}=\tilde{\boldsymbol{c}}} \right) (\boldsymbol{c} - \tilde{\boldsymbol{c}}) \qquad (29) \end{aligned}$$

where

$$\tilde{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} Q(\boldsymbol{c}) \qquad (30)$$

As the first derivation of $\log Q^*(\boldsymbol{c})$ at $\tilde{\boldsymbol{c}}$ is equal to 0, Eq. (29) can be represented as:

$$\log Q^*(\boldsymbol{c}) \simeq \log Q^*(\tilde{\boldsymbol{c}}) - \frac{1}{2} (\boldsymbol{c} - \tilde{\boldsymbol{c}})^\top \boldsymbol{A} (\boldsymbol{c} - \tilde{\boldsymbol{c}}) \qquad (31)$$

$$\begin{aligned} \boldsymbol{A} &= -\frac{\partial^2}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top} \log Q^*(\boldsymbol{c}) \,|_{\boldsymbol{c}=\tilde{\boldsymbol{c}}} \\ &= \frac{N}{2} \boldsymbol{H} \,|_{\boldsymbol{c}=\tilde{\boldsymbol{c}}} + \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{P}_{\boldsymbol{q}}^{-1} \qquad (32) \end{aligned}$$

The Hessian matrix $\boldsymbol{H}$ is represented as follows:

$$\boldsymbol{H} = -\frac{2}{N} \frac{\partial^2}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top} \log P(\boldsymbol{x}|\boldsymbol{c})$$

$$= \mathrm{diag}\left(\left[\boldsymbol{H}_1^\top, \boldsymbol{H}_2^\top, \cdots, \boldsymbol{H}_T^\top\right]^\top\right) \tag{33}$$

where $\boldsymbol{H}_t$ is the Hessian matrix of the spectral evaluation function $E(\boldsymbol{x}_t, \boldsymbol{c}_t)$ in Eq. (5) at time t:

$$\boldsymbol{H}_t = \frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} E(\boldsymbol{x}_t, \boldsymbol{c}_t) = -\frac{2}{N} \frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} \log P(\boldsymbol{x}_t | \boldsymbol{c}_t) \tag{34}$$

In order to approximate $Q(\boldsymbol{c})$ by a Gaussian probability distribution, the normalization term $Z_{\boldsymbol{c}}$ is approximated as:

$$Z_{\boldsymbol{c}} \simeq Q^*(\tilde{\boldsymbol{c}}) \sqrt{(2\pi)^{MT} \left|\boldsymbol{A}^{-1}\right|} \tag{35}$$

By using a Laplace approximation, $Q(\boldsymbol{c})$ is represented as:

$$Q(\boldsymbol{c}) \simeq \mathcal{N}\left(\boldsymbol{c} | \tilde{\boldsymbol{c}}, \boldsymbol{A}^{-1}\right) \tag{36}$$

As the matrix $\boldsymbol{A}$ is a $(4LM + 1)$-diagonal band symmetric matrix where L is the window length, the inverse matrix $\boldsymbol{A}^{-1}$ can be calculated in realistic time.

3.2 Posterior Probabilities of State Sequences

The Forward-Backward algorithm is generally applied to the standard HMM in E-step. However, it cannot be applied to the trajectory HMM, and the delayed decision Viterbi algorithm [17], [18] is applied instead. Thus, we derive a delayed decision Viterbi algorithm for the proposed model similarly.

By using Eq. (36), the expectation with respect to $\boldsymbol{c}$ in Eq. (25) is given by

$$\begin{aligned} &\int Q(\boldsymbol{c}) \log P(\boldsymbol{c} | \boldsymbol{q}, \boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &\simeq \int \mathcal{N}\left(\boldsymbol{c} | \tilde{\boldsymbol{c}}, \boldsymbol{A}^{-1}\right) \log \mathcal{N}\left(\boldsymbol{c} | \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right) d\boldsymbol{c} \\ &= \log \mathcal{N}\left(\tilde{\boldsymbol{c}} | \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right) - \frac{1}{2} \mathrm{tr}\left(\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right) \\ &= \log P(\tilde{\boldsymbol{c}} | \boldsymbol{q}, \boldsymbol{\Lambda}) - \frac{1}{2} \mathrm{tr}\left(\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right) \end{aligned} \tag{37}$$

In Eq. (12), although $\left|\boldsymbol{\Sigma}_{\boldsymbol{q}}\right|$ and $\boldsymbol{\mu}_{\boldsymbol{q}}^T \boldsymbol{\Sigma}_{\boldsymbol{q}} \boldsymbol{\mu}_{\boldsymbol{q}}$ can be computed time-recursively, it is difficult to recursively compute $\left|\boldsymbol{P}_{\boldsymbol{q}}\right|$ and $\boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}}$ because of the temporal full-covariance matrix $\boldsymbol{P}_{\boldsymbol{q}}$. However, by using the special structure of $\boldsymbol{P}_{\boldsymbol{q}}$, "trajectory likelihood" (Eq. (9)) can be computed in a time-recursive manner. When $\Delta \tilde{\boldsymbol{c}}_t$ and $\Delta^2 \tilde{\boldsymbol{c}}_t$ are computed as regression coefficients from $(\tilde{\boldsymbol{c}}_{t-L}, \cdots, \tilde{\boldsymbol{c}}_{t+L})$, $\boldsymbol{R}_{\boldsymbol{q}}$ becomes a $(4LM + 1)$-diagonal band symmetric positive definite matrix. Accordingly, $\boldsymbol{R}_{\boldsymbol{q}}$ can be decomposed by Cholesky decomposition:

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}} \tag{38}$$

where $\boldsymbol{U}_{\boldsymbol{q}}$ is an upper $(2LM + 1)$-band triangular matrix. From Eq. (38), $\left|\boldsymbol{P}_{\boldsymbol{q}}\right|$ can be rewritten as:

$$\left|\boldsymbol{P}_{\boldsymbol{q}}\right| = \left|\boldsymbol{R}_{\boldsymbol{q}}\right|^{-1} = \left|\boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}\right|^{-1} = \left|\boldsymbol{U}_{\boldsymbol{q}}\right|^{-2} = \prod_{t=1}^{T} \left|U_{\boldsymbol{q}_{t+L}}^{(t,t)}\right|^{-2} \tag{39}$$

where $\boldsymbol{q}_{t+L} = (q_1, \cdots, q_{t+L})$. Since $\boldsymbol{U}_{\boldsymbol{q}_{t+L}}^{(t,t)}$ depends only on the state sequence from time 1 to $t+L$, $\left|\boldsymbol{P}_{\boldsymbol{q}}\right|$ can be computed time-recursively. From Eqs. (13), (14), and (38), $\boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}}$ can be rewritten by

$$\begin{aligned} \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}} &= \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}}^\top \boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}} = \bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}} \\ &= \boldsymbol{g}_{\boldsymbol{q}}^\top \boldsymbol{g}_{\boldsymbol{q}} \qquad \left(\boldsymbol{g} = \boldsymbol{U}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{r}_{\boldsymbol{q}}\right) \\ &= \sum_{t=1}^{T} \left(\boldsymbol{g}_{\boldsymbol{q}_{t+L}}^{(t)}\right)^\top \boldsymbol{g}_{\boldsymbol{q}_{t+L}}^{(t)} \end{aligned} \tag{40}$$

where $\boldsymbol{g}_{\boldsymbol{q}}$ is a vector computed from $\boldsymbol{U}_{\boldsymbol{q}}$ and $\boldsymbol{r}_{\boldsymbol{q}}$ by forward substitutions. Since $\boldsymbol{g}_{\boldsymbol{q}_{t+L}}^{(t)}$ depends only on the state sequence from time 1 to $t + L$, $\boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}}$ can be also computed time-recursively. As a result, "trajectory likelihood" can be computed time-recursively as follows:

$$P(\tilde{\boldsymbol{c}} | \boldsymbol{q}, \boldsymbol{\Lambda}) = \prod_{t=1}^{T} \frac{1}{Z_{\boldsymbol{q}_{t+L}}^{(t)}} P(\tilde{\boldsymbol{o}}_t | q_t, \boldsymbol{\Lambda}) \tag{41}$$

where

$$\begin{aligned} Z_{\boldsymbol{q}_{t+L}}^{(t)} &= \frac{\sqrt{(2\pi)^M \left|\boldsymbol{U}_{\boldsymbol{q}_{t+L}}^{(t,t)}\right|^{-2}}}{\sqrt{(2\pi)^{3M} \left|\boldsymbol{\Sigma}_{q_t}\right|}} \\ &\quad \times \exp\left[-\frac{1}{2}\left\{\boldsymbol{\mu}_{q_t}^\top \boldsymbol{\Sigma}_{q_t}^{-1} \boldsymbol{\mu}_{q_t} - \left(\boldsymbol{g}_{\boldsymbol{q}_{t+L}}^{(t)}\right)^\top \boldsymbol{g}_{\boldsymbol{q}_{t+L}}^{(t)}\right\}\right] \end{aligned} \tag{42}$$

From Eq. (38), submatrices of $\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}$ in Eq. (37) can be rewritten as:

$$\begin{aligned} \left(\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right)^{(t,t)} &= \left(\boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right)^{(t,t)} = \left(\boldsymbol{U}_{\boldsymbol{q}} \boldsymbol{A}^{-1} \boldsymbol{U}_{\boldsymbol{q}}^\top\right)^{(t,t)} \\ &= \sum_{i=t}^{t+2L} \sum_{j=t}^{t+2L} \boldsymbol{U}_{\boldsymbol{q}_{t+2L}}^{(t,i)} \left(\boldsymbol{A}^{-1}\right)^{(i,j)} \boldsymbol{U}_{\boldsymbol{q}_{t+2L}}^{(t,j)} \end{aligned} \tag{43}$$

Since $\boldsymbol{U}_{\boldsymbol{q}_{t+2L}}^{(t,j)}$ depends only on the state sequence from time 1 to $t + 2L$, $\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}$ can be computed time-recursively. Therefore, Eq. (37) is represented as:

$$\begin{aligned} &\int Q(\boldsymbol{c}) \log P(\boldsymbol{c} | \boldsymbol{q}, \boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &\simeq \sum_{t=1}^{T} \left[\log \frac{1}{Z_{\boldsymbol{q}_{t+L}}^{(t)}} \mathcal{N}\left(\boldsymbol{W} \tilde{\boldsymbol{c}}_t | \boldsymbol{\mu}_{q_t}, \boldsymbol{\Sigma}_{q_t}\right) \right. \\ &\quad \left. - \frac{1}{2} \sum_{i=t}^{t+2L} \sum_{j=t}^{t+2L} \mathrm{tr}\left\{\boldsymbol{U}_{\boldsymbol{q}_{t+2L}}^{(t,i)} \left(\boldsymbol{A}^{-1}\right)^{(i,j)} \boldsymbol{U}_{\boldsymbol{q}_{t+2L}}^{(t,j)}\right\}\right] \end{aligned} \tag{44}$$

Thus, the proposed method can use the delayed decision Viterbi algorithm.

3.3 Update Model Parameters

Model parameters $\boldsymbol{m}$ and $\boldsymbol{\phi}$ are defined by concatenating the mean vectors and covariance matrices of all unique Gaussian components in the model set as:

$$\boldsymbol{m} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \cdots, \boldsymbol{\mu}_K^\top\right]^\top \tag{45}$$

$$\boldsymbol{\phi} = \left[\boldsymbol{\Sigma}_1^\top, \boldsymbol{\Sigma}_2^\top, \cdots, \boldsymbol{\Sigma}_K^\top\right]^\top \tag{46}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th unique Gaussian component in the model set, and K is the total number of Gaussian components in the model set, respectively.

By setting the partial derivative of $\mathcal{F}$ with respect to $\boldsymbol{m}$ to 0, a set of linear equations for determining $\boldsymbol{m}$ maximizing $\mathcal{F}$ are obtained as:

$$\sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{W}^\top \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\Phi}^{-1} \boldsymbol{m} = \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W} \tilde{\boldsymbol{c}} \tag{47}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{m} \tag{48}$$

$$\boldsymbol{\Phi}^{-1} = \mathrm{diag}(\boldsymbol{\phi}) \tag{49}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \mathrm{diag}(\boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\phi}) \tag{50}$$

$$\boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{S}_{\boldsymbol{q}} \tag{51}$$

In the above equations, $\boldsymbol{S}_{\boldsymbol{q}}$ is a $3MT \times 3MT$ matrix whose elements are 0 or 1 determined by the Gaussian component sequence $\boldsymbol{q}$.

For maximizing $\mathcal{F}$ with respect to $\boldsymbol{\phi}$, a gradient method is applied by using its partial derivative

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \boldsymbol{\phi}} \simeq \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \Bigg[\frac{1}{2} \boldsymbol{S}_{\boldsymbol{q}}^\top \,\mathrm{diag}^{-1}\Big\{ & \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{W}^\top - \boldsymbol{W} \boldsymbol{A}^{-1} \boldsymbol{W}^\top \\ & - \boldsymbol{W} \tilde{\boldsymbol{c}} \tilde{\boldsymbol{c}}^\top \boldsymbol{W}^\top + 2 \boldsymbol{\mu}_{\boldsymbol{q}} \tilde{\boldsymbol{c}}^\top \boldsymbol{W}^\top \\ & + \boldsymbol{W} \bar{\boldsymbol{c}}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top - 2 \boldsymbol{\mu}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top \Big\}\Bigg]\end{aligned} \tag{52}$$

because Eq. (52) is not a quadratic function of $\boldsymbol{\phi}$. As explained above, the parameterization of the proposed model is completely the same as that of the standard HMM and trajectory HMM.

3.4 Related Work

As mentioned above, the proposed method integrates the spectral estimation process and the spectral modeling process and the generative model is defined on the waveform domain. Some similar approaches have been found in previous researches. The vocal tract transfer function (VTTF) estimation of a speech signal based on a factor analyzed (FA) trajectory HMMs [19] is closely related to the proposed method in terms of the direct modeling of speech observation. In this method, mel-cepstral coefficients are regarded as factors and the harmonic components are represented by using linear transformation with the time-varying factor loading matrix. The likelihood function is defined in the log spectral domain and measured only on voiced frames of speech while the likelihood function of the proposed method is defined in the waveform domain. Furthermore, as the proposed method is based on the conventional acoustic model structure, the proposed method has an advantage that reasonable initial model parameters can be given by the conventional method and many techniques are regarded for the conventional models, e.g. speaker adaptation, can be applied.

In another related approach, the mel-cepstral analysis was integrated into the estimation of Gaussian mixture model (GMM) for modeling a quasi-stationary Gaussian process [20]. It can represent mel-cepstral coefficients stochastically with mixture weights of GMM. However, mel-cepstral coefficients are constant because each mixture has no variance parameters, and the temporal continuity of mel-cepstral coefficients is also not considered. Contrary to this, the proposed method assumes mel-cepstral coefficients as latent variables with variances and marginalizes out to form a single generative model. Additionally, the temporal continuity is represented by using the trajectory HMMs.

The joint estimation of the acoustic and excitation model parameters [21] is similar to the proposed method. The distance between natural and synthesized speech waveforms is minimized in the time domain by updating the cepstral sequences, the trajectory HMMs, and the excitation models iteratively. Although the proposed method treats the cepstral coefficients as probabilistic variables and estimate their distributions, the method in [21] uses only single cepstral coefficient vectors as an approximation. Furthermore, the state sequence is fixed through the entire training process in [21]. On the other hand, in the proposed method, the modified delayed decision Viterbi algorithm are derived and the state sequence can be optimized for the integrated objective function.

The proposed algorithm and a small experiment have already shown in [22]. The reason why we conducted only the small experiment in [22] was large computational cost over 1000 hours for the training of the proposed models. This is mainly caused by following processes, (1) Searching the best state sequences with the delayed decision Viterbi algorithm, (2) Iterative updates for estimating the covariance matrices, and (3) Estimating $Q(\boldsymbol{c})$ in Eq. (24). Although the process (1) and (2) are required for both the trajectory HMM and the proposed method, (3) is necessary only for the proposed method, because all mel-cepstral coefficients in each utterance have to be estimated simultaneously. For a large scale experiment, we reduced the computational cost in (3) by changing the optimization method from the Newton-Raphson method to the RPROP [23] method and using the distributed processing in the estimation of $Q(\boldsymbol{c})$.

4. Experiments

To evaluate the effectiveness of the proposed method, objective comparison tests on the likelihood measure and subjective comparison tests on the mean opinion score (MOS) were conducted. For training, two data sets which contain different number of sentences from the phonetically balanced 503 sentences of the ATR Japanese speech database (Set B) [28] recorded in NITech were used.

- Small data set: 50 sentences
- Large data set: 450 sentences

Fifty other sentences were used for evaluation. The speech data was recorded at 48 kHz and windowed at a frame rate of 5-ms by using a 25-ms Hamming window. The windowed waveforms were used as the input data in the proposed method, and 35 mel-cepstral coefficients, which include the zero coefficient estimated with the mel-cepstral analysis technique [11], and their delta and delta-delta coefficients were used in the conventional method. The dimension of the hidden mel-cepstral coefficients of the proposed method was set to the same as that of the conventional method. The excitation parameter vectors consisted of log F_0 and its delta and delta-delta. The frequency warping parameter α was set to 0.55. A five-state, left-to-right, no-skip structure was used for the HMMs. The excitation parameters were modeled with multi-space probability distributions HMMs [29] in both the proposed and conventional methods. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix.

The standard HMMs were estimated as context-dependent models [30] and applied the decision tree based context clustering technique [31]. The minimum description length (MDL) criterion was used to determine the size of the decision trees [32]. After estimating the standard HMMs, the trajectory HMMs and proposed models were re-estimated by using the standard HMMs as their initial models in accordance with the training procedure described in Sect. 3. The number of delayed frames in the delayed decision Viterbi algorithm was set to seven. In the subjective test, ten subjects were asked to rate the naturalness of the synthesized speech on a MOS with a scale from 1 (poor) to 5 (good). Fifteen randomly selected sentences were presented to each subject. The experiments were carried out in a sound-proof room.

4.1 Experiments of Small Data Set

In the experiments on the small data set, an iteration of the proposed embedded training was decided as follows: (Step A) Estimating $Q(\boldsymbol{c})$, and (Step B) estimating $Q(\boldsymbol{q})$ by delayed decision Viterbi algorithm were repeated three times, and then (Step C) the model parameters were updated. The embedded training process was repeated 5 times.

Figure 4 shows the difference of likelihood $P(\boldsymbol{x}|\boldsymbol{\Lambda})$ for the training data set (close) and the test data set (open). The vertical axis shows the average log likelihood per frame. All likelihoods were measured with the proposed model likelihood $P(\boldsymbol{x}|\boldsymbol{\Lambda})$ in the waveform domain (Eq. (22)). The proposed model outperformed the others for both data sets. This means that speech waveforms rather than mel-cepstrum were modeled appropriately in the proposed method. Although the trajectory HMMs was expected to obtain a higher likelihood than HMMs, similar likelihoods were actually obtained. This result indicates that improvement of each component does not always achieve better modeling in terms of the final objective measure. Figure 5 shows the subjective listening test results. In Fig. 5, the MOS of the proposed method was better than that of the standard HMMs and similar to or better than that of the trajectory HMMs.

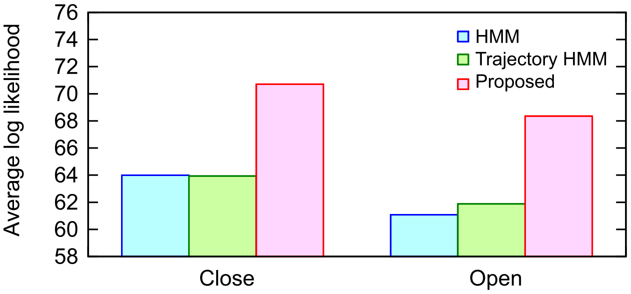


Fig. 4 Log likelihood per frame for close and open data sets (Small data set).

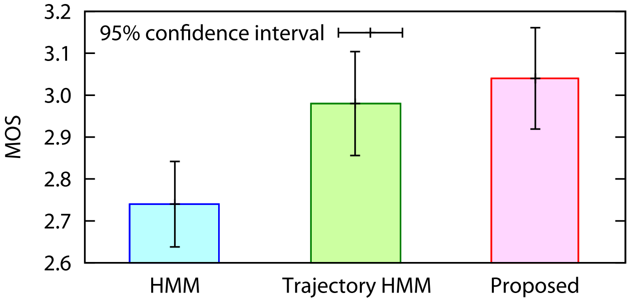


Fig. 5 Mean opinion scores for synthesized speech obtained by standard HMMs, trajectory HMMs and proposed model (Small data set).

4.2 Experiments on Large Data Set

In the experiments on the large data set, the state sequences were previously determined by using the delayed decision Viterbi algorithm, and the state sequences and the duration models were fixed to reduce the computational cost while the trajectory HMMs and the proposed models were trained. The training process of the proposed models, (Step A) estimating $Q(\boldsymbol{c})$ and (Step C) updating the model parameters, was repeated 5 times. As a result, the total computational time was about 1000 hours. Actually, the computational time was reduced by parallel processing of Step A using multiple computers.

Figure 6 shows the subjective listening test results. The MOS of the proposed method was significantly better than the others. The reason why the trajectory HMMs obtained a slightly worse MOS than the standard HMMs might be that the state sequences were fixed through the embedded training of the trajectory HMMs to reduce the computational cost. Figure 7 shows examples of spectrum sequences generated by these models. The state duration for all models was aligned to the natural spectrum sequence so as to compare these spectra easily. It can be observed that the proposed model generated sharper spectra than the other mod-

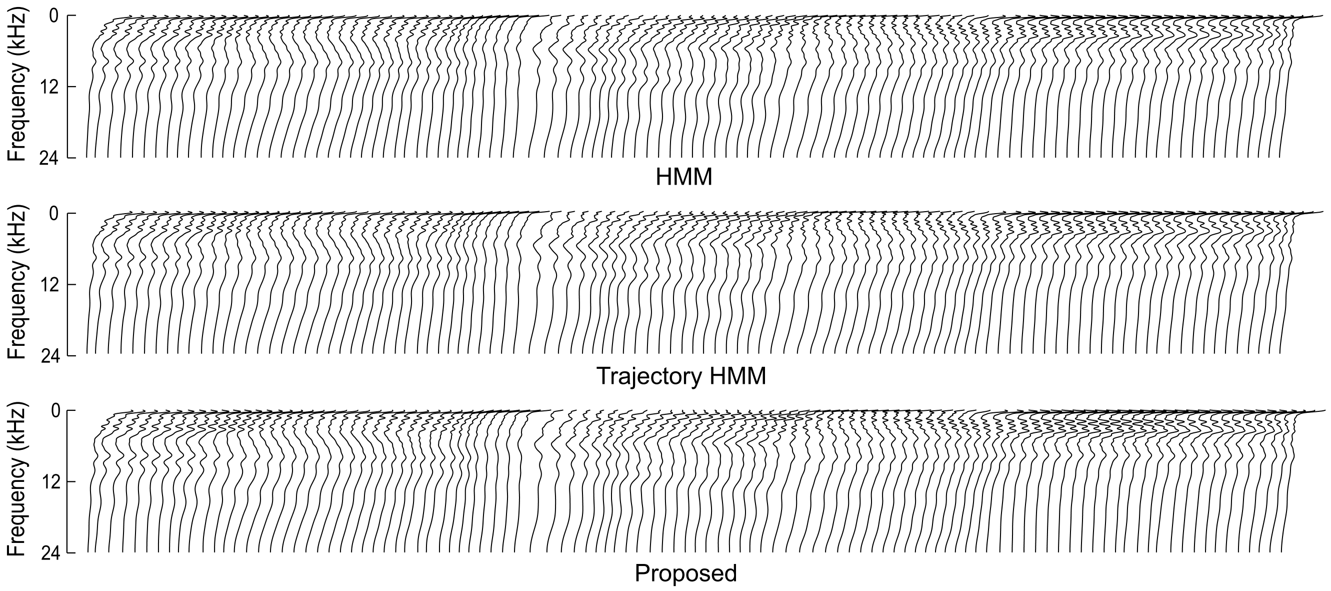


Fig. 7 Example of logarithm spectrum sequences generated using standard HMMs, trajectory HMMs and proposed model. (Large data set).

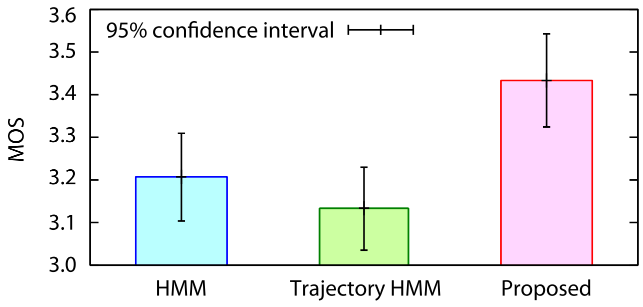


Fig. 6 Mean opinion scores for synthesized speech obtained by standard HMMs, trajectory HMMs and proposed model (Large data set).

els, especially in the low frequency band. It might contribute to naturalness of the generated voices in the proposed method.

These results suggested that the proposed method appropriately modeled speech waveforms directly, even though the proposed model have exactly the same number of parameters as the baseline system. Further improvement is expected by applying the integrated optimization not only to parameter estimation but also to the model structure selection, e.g., context clustering in future work.

5. Conclusions

In this paper, we proposed a novel technique for modeling speech waveforms directly by integrating the mel-cepstral analysis and the acoustic modeling. A generative model representing the TTS problem was constructed and optimized, in which mel-cepstrum coefficients were treated as latent variables and the statistical mel-cepstral analysis and the statistical acoustic model were integrated with marginalizing over mel-cepstral sequences. In the objective experiment, the proposed method outperformed the conventional methods. In addition, the subjective evaluation score of the proposed method was better than that of the conventional methods. These results suggested that the proposed method improves the quality of synthesized speech. Future work includes experiments and evaluation on larger data set with searching the best state sequences by the delayed decision Viterbi algorithm, and constructing a parameter tying structure based on the objective function of the proposed method. Furthermore, the use of other features rather than mel-cepstral coefficients in the proposed framework will also be future work.

Acknowledgements

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

References

[1] T. Masuko, K. Tokuda, T. Kobayashi, and, S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. 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," Proc. 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," Proc. ICASSP, pp.1315–1318, 2000.

[4] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," Speech Commun., vol.52, no.3, pp.223–235, 2010.

[5] A. Parlikar, A. Black, and S. Vogel, "Improving speech synthesis of machine translation output," Proc. 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 Commun., vol.54, no.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," Proc. ACL 2006, 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," Proc. Interspeech 2009, pp.2487–2490, 2009.

[10] K. Oura, Y. Nankaku, T. Toda, K. Tokuda, R. Maia, S. Sakai, and S. Nakamura, "Simultaneous acoustic, prosodic, and phrasing model training for TTS conversion systems," Proc. ISCSLP 2008, pp.1–4, 2008.

[11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. 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," Proc. ICSLP, pp.1043–1045, 1994.

[13] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. UAI 15, pp.21–30, 1999.

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

[15] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features," Proc. Eurospeech, pp.865–868, 2003.

[16] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," Proc. 5th ISCA Speech Synthesis Workshop, 2004.

[17] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," Computer Speech and Language, vol.21, pp.153–173, 2007.

[18] H. Zen, K. Tokuda, and T. Kitamura, "A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features," Proc. ICASSP, pp.837–840, 2004.

[19] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," Proc. ICASSP, pp.3925–3928, 2008.

[20] T. Takahashi, K. Tokuda, T. Kobayashi, and T. Kitamura, "Mixture density models based on mel-cepstral representation of Gaussian process," IEICE Trans. Fundamentals, vol.E86-A, no.8, pp.1971–1978, Aug. 2003.

[21] R. Maia, H. Zen, and M.J.F. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," Proc. Speech Synthesis Workshop 7, pp.88–93, 2010.

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

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

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

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

[26] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," IEICE Trans. Fundamentals (Japanese Edition), vol.J53-A, no.1, pp.35–42, Jan. 1970. Translation: R.W. Schafer and J.D. Markel, eds., Speech Analysis, pp.295–302, IEEE Press, New York, 1979.

[27] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectral approximation filter for speech synthesis," IEICE Trans. Fundamentals (Japanese Edition), vol.J66-A, no.2, pp.122–129, Feb. 1983.

[28] 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 Commun., vol.9, pp.357–363, 1990.

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

[30] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.38, no.4, pp.599–609, 1990.

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

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

Appendix: Likelihood Function of Mel-Cepstrum

It has been shown in literatures (e.g., [25]) that the following equation approximates the log likelihood function of a zero-mean Gaussian process when $N \to \infty$:

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

As a result, it can be seen that the minimization of Eq. (5) is equivalent to maximizing $P(\boldsymbol{x}|\boldsymbol{c})$.

This appendix shows that Eq. (8) approximates the log likelihood function with an assumption that windowed signal

$$\boldsymbol{x}' = \left[x'(0), x'(1), \cdots, x'(N-1)\right]^\top \qquad \text{(A$\cdot$2)}$$

where

$$x'(n) = \sqrt{\frac{N}{\sum_{n=0}^{N-1} w^2(n)}}\, w(n)x(n) \qquad \text{(A$\cdot$3)}$$

is generated by circular convolution of white Gaussian process

$$\boldsymbol{e} = [e(0), e(1), \cdots, e(N-1)]^\top \qquad \text{(A$\cdot$4)}$$

whose variance is unity and

$$\tilde{\boldsymbol{h}} = \left[\tilde{h}(0), \tilde{h}(1), \cdots, \tilde{h}(N-1)\right]^\top \qquad \text{(A$\cdot$5)}$$

where

$$\tilde{h}(n) = \frac{1}{N}\sum_{i=0}^{N-1} H\left(e^{jw_i}\right)e^{jw_i n}, \qquad w_i = \frac{2\pi i}{N} \qquad \text{(A$\cdot$6)}$$

that is, $\boldsymbol{e}$ is obtained by circular convolution of $\boldsymbol{x}'$ and

$$\boldsymbol{g} = [g(0), g(1), \cdots, g(N-1)]^\top \qquad \text{(A$\cdot$7)}$$

where

$$g(n) = \frac{1}{N}\sum_{i=0}^{N-1} H^{-1}\left(e^{jw_i}\right)e^{jw_i n} \qquad \text{(A· 8)}$$

It is noted that $x'(n)$ is normalized so that the energy of $x(n)$ is preserved, and windowing can reduce the effect of replacing convolution by circular convolution.

From the assumption, the likelihood is written as

$$P\left(\boldsymbol{x}'|\boldsymbol{c}\right) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{U}|}}\exp\left(-\frac{1}{2}\boldsymbol{x}'^{\top}\boldsymbol{U}^{-1}\boldsymbol{x}'\right) \qquad \text{(A· 9)}$$

where

$$\boldsymbol{U} = \begin{bmatrix} u(0) & u(1) & \cdots & u(N-1) \\ u(1) & u(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & u(1) \\ u(N-1) & \cdots & u(1) & u(0) \end{bmatrix} \qquad \text{(A· 10)}$$

and

$$u(k) = \frac{1}{N}\sum_{i=0}^{N-1}\left|H\left(e^{j\omega_i}\right)\right|^2 e^{j\omega_i k} \qquad \text{(A· 11)}$$

We can show

$$\boldsymbol{x}'^{\top}\boldsymbol{U}^{-1}\boldsymbol{x}' = \sum_{i=0}^{N-1}\frac{I_N(\omega_i)}{\left|H\left(e^{j\omega_i}\right)\right|^2} \qquad \text{(A· 12)}$$

and

$$|\boldsymbol{U}| = \prod_{i=0}^{N-1}\left|H\left(e^{j\omega_i}\right)\right|^2 \qquad \text{(A· 13)}$$

Consequently, it can be shown

$$\log P\left(\boldsymbol{x}'|\boldsymbol{c}\right) = -\frac{N}{2}\left[\log(2\pi) + \frac{1}{N}\sum_{i=0}^{N-1}\left\{\log\left|H\left(e^{j\omega_i}\right)\right|^2 + \frac{I_N(\omega_i)}{\left|H\left(e^{j\omega_i}\right)\right|^2}\right\}\right] \qquad \text{(A· 14)}$$

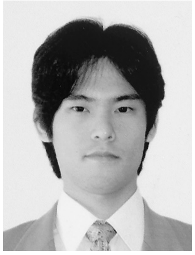
where $I_N(\omega)$ is given by Eq. (7). By replacing the summation by an integration, we obtain

$$\log P\left(\boldsymbol{x}'|\boldsymbol{c}\right) \simeq -\frac{N}{2}\left[\log(2\pi) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\log\left|H\left(e^{j\omega}\right)\right|^2 + \frac{I_N(\omega)}{\left|H\left(e^{j\omega}\right)\right|^2}\right\}d\omega\right] \qquad \text{(A· 15)}$$

Thus, maximizing $P\left(\boldsymbol{x}'|\boldsymbol{c}\right)$, i.e., maximizing Eq. (A· 15) with respect to $\boldsymbol{c}$ is equivalent to the minimization of Eq. (5) with respect to $\boldsymbol{c}$.

Kazuhiro Nakamura received the B.E., and M.E. degrees in intelligence and computer science, and computer science and engineering from Nagoya Institute of technology, Nagoya, Japan in 2005, and 2007, respectively. He is a engineer at Brother Industries, LTD from April 2007. He took a leave of absence, and he is currently a Doctor's candidate at Nagoya Institute of Technology. His research interests include machine learning, statistical speech recognition and synthesis. He is a member of the Acoustical Society of Japan.

Kei Hashimoto receieved the B.E., M.E., and Ph.D. degrees in computer science, computer science and engineering, and scientific and engineering simulation from Nagoya Institute of Technology, Nagoya, Japan in 2006, 2008, and 2011, respectively. From October 2008 to January 2009, he was an intern researcher at National Institute of Information and Communications Technology (NICT), Kyoto, Japan. From April 2010, he was a Research Fellow of the Japan Society for the Promotion of Science (JSPS) at Nagoya Institute of Technology, Nagoya, Japan. From May 2010 to September 2010, he was a visiting researcher at University of Edinburgh and Cambridge University. From April 2012, he is now an Assistant Professor at Nagoya Institute of Technology, Nagoya, Japan. His research interests include statistical speech recognition, speech synthesis and machine translation. He is a member of the Acoustical Society of Japan.



Yoshihiko Nankaku received the B.E. degree in Computer Science, and the M.E. and Ph.D. degrees in the Department of Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004, respectively. After a year as a postdoctoral fellow at the Nagoya Institute of Technology, he is currently an Assistant Professor at the same Institute. His research interests include statistical machine learning, speech recognition, speech synthesis, image recognition, and multi-modal interface. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He published over 60 journal papers and over 150 conference papers, and received 5 paper awards. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 2000 to 2003. Currently he is a member of ISCA Advisory Council and an associate editor of IEEE Transactions on Audio, Speech & Language Processing, and acts as organizer and reviewer for many major speech conferences, workshops and journals. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.