Integration of Acoustic Modeling and Mel-cepstral analysis for HMM-based Speech Synthesis

著者（英）	Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda
journal or publication title	2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)
page range	7883-7887
year	2013-05
URL	http://id.nii.ac.jp/1476/00004649/

doi: 10.1109/ICASSP.2013.6639199(http://dx.doi.org/10.1109/ICASSP.2013.6639199)

INTEGRATION OF ACOUSTIC MODELING AND MEL-CEPSTRAL ANALYSIS FOR HMM-BASED SPEECH SYNTHESIS

Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

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

ABSTRACT

In this paper, a novel approach for integrating acoustic modeling and mel-cepstral analysis is proposed. The aim of HMM-based speech synthesis is to model speech waveforms with a statistical model. However, the conventional techniques divide the modeling process into two steps: the frame by frame feature extraction step and the acoustic modeling step. Although it is reasonably effective, the deterioration of speech quality is caused by the divide of the objective function. In this paper, we propose an approach to modeling them as an integrative model and show the possibility of improving synthesized speech.

***Index Terms*—** integrative model, HMM-based speech synthesis, acoustic modeling, mel-cepstal analysis, trajectory HMM

1. INTRODUCTION

HMM-based speech synthesis was proposed to enable machines to speak naturally like humans [1]. In this method, spectral and F0 features are extracted and modeled with a statistical technique. Recent large systems are often constructed with combinations of some modules that use statistical models. A famous example is language and acoustic models for speech recognition systems. Recently, the integration of these statistical models is an important research subject. In text-to-speech (TTS) systems, an approach integrating text analysis and speech synthesis modules was proposed [2]. It can optimize linguistic and acoustic models simultaneously. As the essential aim of TTS is to synthesize speech from given texts, the integration of these statistical models is a desirable future of TTS systems. In HMM-based speech synthesis, some heuristic methods have been used to extract spectral features from speech waveforms previously. A statistical method that consists of mel-cepstral analysis was proposed and is widely used [3]. In this method, mel-cepstral coefficients, i.e., frequency transformed cepstral coefficients, are regarded as statistical model parameters and estimated with a parametric method.

In the standard HMM, observation vector sequences are quasi-stationary, and each stationary part can be represented by a state of the HMM. 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. A trajectory model, named a "trajectory HMM" [4], was derived by reformulating the HMM. The standard HMM with static and dynamic features allows inconsistent statistics between the model parameters for static and dynamic features. By imposing the explicit relationship between them, the standard HMM is naturally translated into a trajectory model. The trajectory HMM can overcome the limitations in the standard HMM framework without any additional parameters.

In the mel-cepstral analysis and the acoustic modeling, two statistical models are estimated independently. Since the aim of these two techniques is to model speech with a statistical model, in this paper, we propose integrating the feature extraction and the acoustic modeling by using a probabilistic representation of extracted features. This approach can model speech waveforms directly without having any intermediate representation, and it can be regarded as a generative model of speech waveforms.

In this model, the spectral extraction process and the spectral modeling process are simultaneously optimized for the given speech waveforms. This framework is similar to the vocal tract transfer function (VTTF) estimation of a speech signal based on a factor analyzed (FA) trajectory HMM [5]. Mel-cepstral coefficients were regarded as factors, and by using the time-varying factor loading matrix, the harmonic components were represented with the FA method. In another approach, the mel-cepstral analysis was integrated into the Gaussian mixture model (GMM) for modeling a quasi-stationary Gaussian process [6]. It can represent mel-cepstral coefficients stochastically with mixture weights of GMM.

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

2. HMM-BASED SPEECH SYNTHESIS

In HMM-based TTS training, spectral envelope, fundamental frequency, and duration are modeled simultaneously by using the corresponding HMMs. Mel-cepstral coefficients $\boldsymbol{c}$ are widely used as spectral features. They are regarded as statistical model parameters and estimated from a given input signal, $\boldsymbol{x}$, in the maximum likelihood (ML) sense:

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} P\left(\boldsymbol{x}|\boldsymbol{c}\right) \tag{1}$$

Extracted mel-cepstral coefficients are used for training HMMs with dynamic ("delta" and "delta-delta") feature constraints. A speech waveform is finally synthesized from the generated spectral and excitation parameters via the source-filter based production model. Recently, the trajectory HMM was derived by reformulating the HMM. The model parameters $\boldsymbol{\Lambda}$ are trained in the ML sense by using the static features:

$$\hat{\boldsymbol{\Lambda}} = \operatorname*{argmax}_{\boldsymbol{\Lambda}} P\left(\boldsymbol{c}|\boldsymbol{\Lambda}\right) \tag{2}$$

2.1. MEL-CEPSTRAL ANALYSIS

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

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

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

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

where α is a frequency warping parameter. If $\alpha = 0$, mel-cepstral coefficients are equivalent to cepstral coefficients.

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 $\boldsymbol{c}$ [7],

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

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 likelihood can be approximated by

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

and, accordingly, minimization of $E(\boldsymbol{x}, \boldsymbol{c})$ corresponds to the maximization of $P(\boldsymbol{x}|\boldsymbol{c})$.

2.2. TRAJECTORY HMM

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 $M+1$ dimensional vector, and T is the number of frames. In the conventional model, the probability density of $\boldsymbol{o}$ is shown as $P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\Lambda})$, where $\boldsymbol{q} = (q_1, q_2, \cdots, q_T)$ is a state sequence. However, the conventional model is mathematically improper in the sense of statistical modeling. In this model, the static and dynamic features are modeled as independent statistical variables. When it is used as a generative model, it allows inconsistent static and dynamic features. 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 $3(M+1)T \times (M+1)T$ window matrix as shown in Fig. 1, the conventional HMM is reformed as the trajectory HMM as:

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

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

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

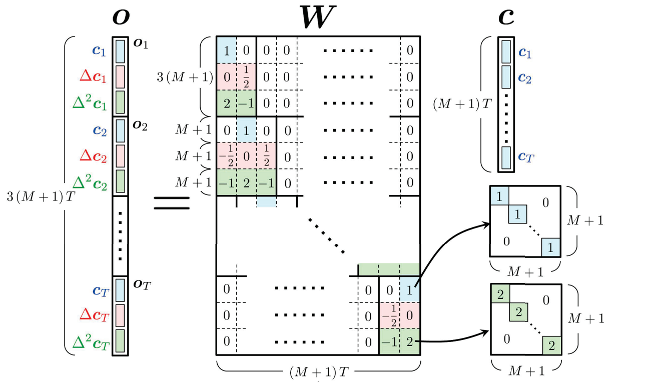

Fig. 1. Example of the relationship between the static feature vector sequence $\boldsymbol{c}$ and the speech parameter vector sequence $\boldsymbol{o}$ in a matrix form

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

$$\boldsymbol{R}_{\boldsymbol{q}}\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}} \tag{12}$$

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1} \tag{13}$$

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

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \cdots, \boldsymbol{\mu}_{q_T}^\top\right]^\top \tag{15}$$

$$\boldsymbol{\mu}_i = \left[\boldsymbol{\mu}_i^\top, \Delta\boldsymbol{\mu}_i^\top, \Delta^2\boldsymbol{\mu}_i^\top\right]^\top, \; i = 1, \cdots, N \tag{16}$$

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

$$\boldsymbol{\Sigma}_i = \mathrm{diag}\left[\boldsymbol{\Sigma}_i^\top, \Delta\boldsymbol{\Sigma}_i^\top, \Delta^2\boldsymbol{\Sigma}_i^\top\right]^\top, \; i = 1, \cdots, N \tag{18}$$

where N is the total number of state output PDFs, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the $3(M+1)T \times 1$ mean vector and the $3(M+1)T \times 3(M+1)T$ covariance matrix associated with the i-th state, respectively. The elements of $\boldsymbol{W}$ are given as regression window coefficients to calculate delta and delta-delta features as:

$$\Delta^d \boldsymbol{c}_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)\, \boldsymbol{c}_{t+\tau}, \quad d = 1, 2 \tag{19}$$

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{(M+1)\times(M+1)} \tag{20}$$

$$\boldsymbol{W}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \ldots, \boldsymbol{w}_t^{(D-1)}\right] \tag{21}$$

$$\boldsymbol{w}_t^{(d)} = \Big[\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(0), \ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-\left(t+L_+^{(d)}\right)}\Big]^\top, d = 0, 1, 2 \tag{22}$$

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 Gaussian distributions whose dimensionality is $(M+1)T$, and the covariance matrices $\boldsymbol{P}_{\boldsymbol{q}}$ are generally full. As a result, the trajectory HMM can overcome the deficiencies of the HMM. It is also noted that the parameterization of

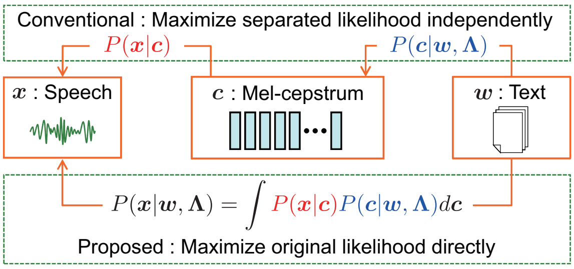

Fig. 2. Basic idea of the proposed approach

the trajectory HMM is completely the same as that of the HMM with the same model topology.

3. INTEGRATION ALGORITHM OF ACOUSTIC MODELING AND MEL-CEPSTRAL ANALYSIS

Figure 2 shows the difference between the conventional and the proposed approaches. In the proposed approach, we integrate the statistical mel-cepstral model $P(\boldsymbol{x}|\boldsymbol{c})$ and the statistical acoustic model $P(\boldsymbol{c}|\boldsymbol{\Lambda})$ as:

$$\begin{aligned}\hat{\boldsymbol{\Lambda}} &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} P(\boldsymbol{x}|\boldsymbol{\Lambda}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \int P(\boldsymbol{x},\boldsymbol{c}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &= \operatorname*{argmax}_{\boldsymbol{\Lambda}} \int P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \end{aligned} \tag{23}$$

In the proposed method, the mel-cepstral coefficients are represented as a random variable, and the dynamic feature vectors cannot be calculated from static features. Therefore, a modeling technique including the extraction of dynamic features, like trajectory HMM, should be introduced. As far as the conventional method, the standard mel-cepstral analysis can be assumed as extracting 1-best 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:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}|\boldsymbol{\Lambda}) &= \log P(\boldsymbol{x}|\boldsymbol{\Lambda}) \\ &= \log \sum_{\forall \boldsymbol{q}} \int P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c},\boldsymbol{q}|\boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &= \log \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c},\boldsymbol{q}) \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c},\boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c},\boldsymbol{q})} d\boldsymbol{c} \\ &= \log \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c})\, Q(\boldsymbol{q}) \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c},\boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c})\, Q(\boldsymbol{q})} d\boldsymbol{c} \\ &\geq \sum_{\forall \boldsymbol{q}} \int Q(\boldsymbol{c})\, Q(\boldsymbol{q}) \log \frac{P(\boldsymbol{x}|\boldsymbol{c})\, P(\boldsymbol{c},\boldsymbol{q}|\boldsymbol{\Lambda})}{Q(\boldsymbol{c})\, Q(\boldsymbol{q})} d\boldsymbol{c} \\ &= \mathcal{F} \end{aligned} \tag{24}$$

To overcome the difficulty of optimization, we assume that $\boldsymbol{c}$ and $\boldsymbol{q}$ are independent and that posterior distribution $Q(\boldsymbol{c})$ and $Q(\boldsymbol{q})$ approximate the true posterior distributions. The optimal posterior distributions can be obtained by maximizing the original objective function $\mathcal{F}$ with the variational method as:

$$Q(\boldsymbol{c}) = \frac{1}{Z_{\boldsymbol{c}}} P(\boldsymbol{x}|\boldsymbol{c}) \exp \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \log P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\Lambda}) \tag{25}$$

$$Q(\boldsymbol{q}) = \frac{1}{Z_{\boldsymbol{q}}} P(\boldsymbol{q}|\boldsymbol{\Lambda}) \exp \int Q(\boldsymbol{c}) \log P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\Lambda})\, d\boldsymbol{c} \tag{26}$$

where $Z_{\boldsymbol{c}}$ and $Z_{\boldsymbol{q}}$ are the normalization terms of $Q(\boldsymbol{c})$ and $Q(\boldsymbol{q})$, respectively. These optimizations can be effectively performed by iterative calculations as the Expectation and Maximization (EM) algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence.

3.1. Posterior Probabilities of Mel-cepstral coefficients

It is difficult to integrate $Q(\boldsymbol{c})$ with respect to $\boldsymbol{c}$, so we approximate $Q(\boldsymbol{c})$ as a Gaussian probability distribution. By using a Laplace approximation, $Q(\boldsymbol{c})$ is represented as:

$$Q(\boldsymbol{c}) \simeq \mathcal{N}\left(\boldsymbol{c}|\tilde{\boldsymbol{c}}, \boldsymbol{A}^{-1}\right) \tag{27}$$

$$\tilde{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} Q(\boldsymbol{c}) \tag{28}$$

$$\boldsymbol{A} = \frac{N}{2} \boldsymbol{H}\, |_{\boldsymbol{c}=\tilde{\boldsymbol{c}}} + \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{P}_{\boldsymbol{q}}^{-1} \tag{29}$$

where

$$\boldsymbol{H} = -\frac{2}{N} \frac{\partial^2}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top} \log P(\boldsymbol{x}|\boldsymbol{c}) \tag{30}$$

$$= \operatorname{diag}\left(\left[\boldsymbol{H}_1^\top, \boldsymbol{H}_2^\top, \cdots, \boldsymbol{H}_T^\top\right]^\top\right) \tag{31}$$

$\boldsymbol{H}_t$ is the Hessian matrix of the spectral evaluation function at time t:

$$\boldsymbol{H}_t = \frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} E(\boldsymbol{x}_t, \boldsymbol{c}_t) = -\frac{2}{N} \frac{\partial^2}{\partial \boldsymbol{c}_t \partial \boldsymbol{c}_t^\top} \log P(\boldsymbol{x}_t|\boldsymbol{c}_t) \tag{32}$$

In the standard mel-cepstral analysis, mel-cepstral coefficients for each frame can be estimated frame by frame independently. However, in the proposed method, mel-cepstral coefficients for all utterances should be estimated simultaneously. Thus, a large computation cost is required for this Newton-Raphson method.

3.2. Posterior Probabilities of State Sequences

The expectation with respect to $\boldsymbol{c}$ in Eq. (26) is given by

$$\begin{aligned}&\int Q(\boldsymbol{c}) \log P(\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\Lambda})\, d\boldsymbol{c} \\ &= \log \mathcal{N}(\tilde{\boldsymbol{c}}|\bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}) - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right)\end{aligned} \tag{33}$$

where the matrix $\boldsymbol{R}_{\boldsymbol{q}}$ is a positive definite $(4L(M+1)+1)$-diagonal band symmetric matrix. Thus, $\boldsymbol{R}_{\boldsymbol{q}}$ can be decomposed into its Cholesky factorization:

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}} \tag{34}$$

where $\boldsymbol{U}_{\boldsymbol{q}}$ is an upper-triangular $(2L(M+1)+1)$-diagonal matrix. Elements of $\boldsymbol{U}_{\boldsymbol{q}}$ are calculated in a recursive manner and depend only on substates from time 1 to $t+2L$.

$$\begin{aligned}\left(\boldsymbol{R}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right)^{(t,t)} &= \left(\boldsymbol{U}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}} \boldsymbol{A}^{-1}\right)^{(t,t)} = \left(\boldsymbol{U}_{\boldsymbol{q}} \boldsymbol{A}^{-1} \boldsymbol{U}_{\boldsymbol{q}}^\top\right)^{(t,t)} \\ &= \sum_{i=t}^{t+2L} \sum_{j=t}^{t+2L} \boldsymbol{U}_{\boldsymbol{q}}^{(t,i)} \left(\boldsymbol{A}^{-1}\right)^{(i,j)} \boldsymbol{U}_{\boldsymbol{q}}^{(t,j)}\end{aligned} \tag{35}$$

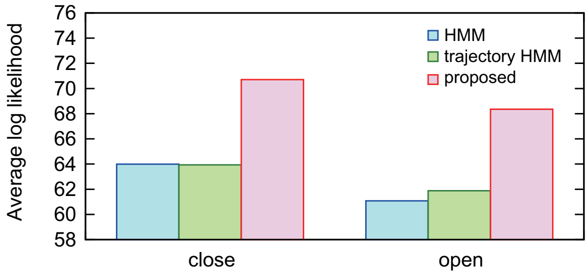

Fig. 3. Log likelihood per frame for closed and opened data sets

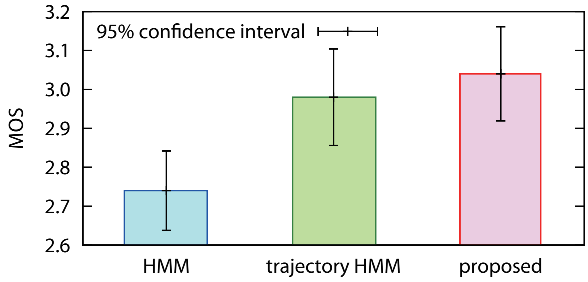

Fig. 4. Mean opinion scores for synthesized speech obtained by standard HMM, trajectory HMM and proposed model.

Thus, the delayed decision Viterbi algorithm [4] can be applied to the proposed method.

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:

$$\boldsymbol{m} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \cdots, \boldsymbol{\mu}_N^\top\right]^\top \tag{36}$$

$$\boldsymbol{\phi} = \left[\boldsymbol{\Sigma}_1^\top, \boldsymbol{\Sigma}_2^\top, \cdots, \boldsymbol{\Sigma}_N^\top\right]^\top \tag{37}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are the mean vector and covariance matrix of the n-th unique Gaussian component in the model set, and N 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:

$$\sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{W}^\top \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\Phi}^{-1} \boldsymbol{m} = \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q})\, \boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W} \tilde{\boldsymbol{c}} \tag{38}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{q}} &= \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{m} && (39)\\ \boldsymbol{\Phi}^{-1} &= \mathrm{diag}(\boldsymbol{\phi}) && (40)\\ \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} &= \mathrm{diag}(\boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\phi}) && (41)\\ \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{\Phi}^{-1} &= \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{S}_{\boldsymbol{q}} && (42) \end{aligned}$$

In the above equations, $\boldsymbol{S}_{\boldsymbol{q}}$ is a $3(M+1)T \times 3(M+1)T$ 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

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \boldsymbol{\phi}} = \sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) \Bigg[& \frac{1}{2} \boldsymbol{S}_{\boldsymbol{q}}^\top \,\mathrm{diag}^{-1} \Big\{ \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{W}^\top - \boldsymbol{W} \boldsymbol{A}^{-1} \boldsymbol{W}^\top \\ & - \boldsymbol{W} \tilde{\boldsymbol{c}} \tilde{\boldsymbol{c}}^\top \boldsymbol{W}^\top + 2 \boldsymbol{\mu}_{\boldsymbol{q}} \tilde{\boldsymbol{c}}^\top \boldsymbol{W}^\top \\ & + \boldsymbol{W} \bar{\boldsymbol{c}}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top - 2 \boldsymbol{\mu}_{\boldsymbol{q}} \bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top \boldsymbol{W}^\top \Big\} \Bigg] \end{aligned} \tag{43}$$

because Eq. (43) is not a quadratic function of $\boldsymbol{\phi}$.

4. EXPERIMENTS

4.1. Experimental Conditions

To evaluate the effectiveness of the proposed method, objective comparison tests of likelihood and subjective comparison test of Mean opinion Score (MOS) were conducted. For training, 50 sentences of the phonetically balanced 503 sentences from the ATR Japanese speech database (Set B) uttered by a male speaker M001 in Nitech, were used. Fifty other sentences were used for evaluating. 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 standard mel-cepstral analysis technique 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 frequency warping parameter α was set to 0.55. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix. In the subjective test, 10 subjects were asked to rate the naturalness of the synthesized speech voices 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.

4.2. Experimental Results

After estimating the standard HMMs, the proposed models were re-estimated by using the standard HMMs as their initial models in accordance with the training procedure described in Section 3. Figure 3 shows the average log likelihood per frame for a training data set (close) and test data set (open) to compare the proposed models with the standard HMMs. The proposed models outperformed the baseline standard HMMs for both data sets. This result for the test data set especially indicates the possibility of improving synthesized speech. Figure 4 shows the subjective listening results. In Figure 4, the MOS of the proposed technique was higher than that of the standard HMMs and equivalent to or higher than that of the trajectory HMMs. These results mean that the proposed technique can represent speech waveforms better by modeling them directly, even though baseline and proposed models have the same number of parameters.

5. CONCLUSION

In this paper, we defined a novel kind of acoustic model for modeling speech waveforms directly by integrating the mel-cepstral analysis and the acoustic modeling. In experiments, the objective and subjective evaluation scores of proposed models were equivalent to or higher than the baseline models. These results indicate the possibility of improving the quality of synthesized speech. Experiments on larger data sets will be future work.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

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

[2] K. Oura, Y. Nankaku, T. Toda, and K. Tokuda, "Simultaneous acoustic, prosodic, and phrasing model training for TTS conversion systems," *Proceedings of ISCSLP2008*, pp. 1–4, 2008.

[3] 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.

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

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

[6] T. Takahashi, K. Tokuda, T. Kobayashi, and T. Kitamura, "Mixture density models based on mel-cepstral representation of gaussian process," *Proceedings of IEICE*, pp. 1971–1978, 2003.

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