FACTOR ANALYZED VOICE MODELS FOR HMM-BASED SPEECH SYNTHESIS

Kyosuke Kazumi, Yoshihiko Nankaku, Keiichi Tokuda

Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper describes factor analyzed voice models for realizing various voice characteristics in the HMM-based speech synthesis. The eigenvoice method can synthesize speech with arbitrary voice characteristics by interpolating representative HMM sets. However, the objective of PCA is to accurately reconstruct each speaker-dependent HMM set, and this is not equivalent to estimating models which represent training data accurately. To overcome this problem, we propose a general speech model which generates speech utterances with various voice characteristics directly. In the proposed method, the HMM states, factors representing voice characteristics and contextual decision trees are simultaneously optimized within a unified framework.

Index Terms— HMM-based speech synthesis, eigenvoice, factor analysis, expectation maximization algorithm, deterministic annealing EM algorithm

1. INTRODUCTION

To let machines speak naturally like a human, an HMM-based speech synthesis system has been proposed [1, 2]. The system models spectrum, pitch and state duration simultaneously in a unified framework of HMMs and synthesizes speech using parameters generated from HMM sets. One of the advantages of the system is that voice characteristics of synthesized speech can be easily changed by transforming HMM parameters.

The eigenvoice method for HMM based synthesis [3] can synthesize speech with arbitrary voice characteristics by interpolating representative HMM sets. In this method, a set of speaker-dependent HMMs is represented by a single large dimensional vector called a supervector and representative supervectors called eigenvoice vectors are constructed by applying Principal Component Analysis (PCA). Finally, speech with a desired voice characteristic is generated by a target HMM set which is reconstructed by interpolating the eigenvoice vectors using given weights representing factors of voice characteristics. However, the objective of PCA in the eigenvoice method is to accurately reconstruct each speaker-dependent HMM set, and this is not equivalent to estimating models which represent training data accurately. Furthermore, speaker-dependent HMMs for all speakers are required with the same decision trees (parameter tying structures). However, it is sometimes difficult to collect enough speech data for all speakers.

To overcome these problems, we propose a general speech model which generates speech utterances with various voice characteristics directly. In this paper, we name the proposed model factor analyzed voice model (FA-voice model), because the structure of the proposed method is based on the factor analysis (FA) model for representing various speaker characteristics. In the proposed method, the likelihood is directly calculated from speech utterances of training data, therefore speaker-dependent HMMs are not required. In the training of the proposed model, the HMM states, factors representing voice characteristics and contextual decision trees are simultaneously optimized within a unified maximum likelihood (ML) framework based on a single statistical model. This simultaneous optimization is expected to achieve a better quality of synthesized speech for a desired voice characteristics.

The parameters of the proposed model can be estimated via the expectation maximization (EM) algorithm for approximating the Maximum Likelihood (ML) estimate. However, the exact expectation step (E-step) is computationally intractable due to the combination of hidden variables. To derive a feasible algorithm, we applied the variational EM algorithm [4] to the proposed model. The variational method approximates the posterior distribution over the hidden variables by a tractable distribution. However, the EM algorithm has the problem that the solution converges to a local optimum and the convergence point depends on the initial model parameters. This problem causes that estimated model parameters are not appropriate for training data and the quality of synthesized speech may be degraded. To overcome this problem, we apply the deterministic annealing EM (DAEM) algorithm [5] to the training of the FA voice model.

The rest of the paper is organized as follows. Section 2 describes the eigenvoice method based on PCA. In Section 3, we propose the voice generation model based on factor analysis and estimation method for model parameters using the EM algorithm, and Section 4 describes the DAEM algorithm for estimating the proposed model parameters. Experimental results are presented in Section 5, and concluding remarks and future work are presented in the Section 6.

2. EIGENVOICE METHOD BASED ON PCA

In the eigenvoice method for HMM-based speech synthesis, speaker dependent HMM sets represent voice characteristics of each speaker. Therefore, mean vectors of these HMM sets are generated from eingenvoice vectors and these weights. For each of S speaker-dependent HMM sets, we extract parameters representing all HMMs. Then, we concatenate all parameters for each speaker and create a vector of a large dimensionality which is called "supervector." By applying PCA, S eigenvoice vectors are calculated from S supervectors. Eigenvoice vectors are called "eigenvoices." Using the first L eigenvoices $\boldsymbol{e}_l$, L arbitrary weight coefficients λ_l for eigenvoices $\boldsymbol{e}_l, l = 1, 2, \ldots, L$, and a vector $\bar{\boldsymbol{\mu}}$ is a mean vector of S supervectors, a new supervector $\hat{\boldsymbol{\mu}}$ is calculated as follows:

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{\mu}} + \sum_{l=1}^{L} \lambda_l \boldsymbol{e}_l. \tag{1}$$

Then a new speaker dependent HMM set is reconstructed from the generated supervector $\hat{\boldsymbol{\mu}}$. It is noted that eigenvoices for speech synthesis should be different from those in speech recognition because they should capture not only spectrum but also $F0$ parameters. Various voice qualities can be synthesized by setting arbitrary weight coefficients λ_l.

978-1-4244-4296-6/10/$25.00

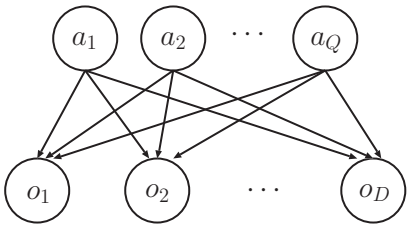

Fig. 1. Factor analysis model

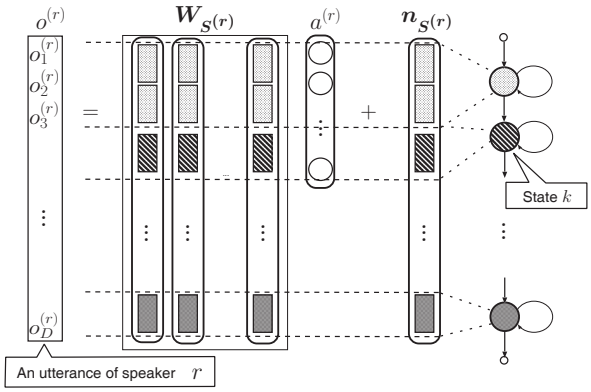

Fig. 2. Structure of the FA voice model

3. FACTOR ANALYZED VOICE MODELS

3.1. Factor analysis

The factor analysis (FA) is one of a statistical method for modeling the covariance structure of high dimensional data using a small number of latent variables. In the FA model, observation variable $\boldsymbol{o}$ is generated by following equation:

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{a} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{a}$ is a Q dimensional latent variable, $\boldsymbol{n}$ is a noise vector, and $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_Q]$, $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \ldots, w_{iD}]^\top$ is a $D \times Q$ matrix known as a factor loading matrix. The structure of the FA model is shown in Figure 1.

Applying the FA model as speech generation model, acoustic feature vectors of speaker dependent utterance $\boldsymbol{o} = [o_1, o_2, \ldots, o_D]^\top$ are generated from a common latent variable $\boldsymbol{a}$ and noise vector $\boldsymbol{n}$ which are inherent in each variable given by

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{I}$ is identity matrix. In this condition, observation sequence $\boldsymbol{o}$ is generated from

$$P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) = \int P(\boldsymbol{o} \mid \boldsymbol{a}, \boldsymbol{\Lambda}) P(\boldsymbol{a} \mid \boldsymbol{\Lambda}) d\boldsymbol{a}, \tag{4}$$

$$P(\boldsymbol{o} \mid \boldsymbol{a}, \boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{o} \mid \boldsymbol{W}\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{5}$$

where Λ is model parameter.

3.2. FA voice models

In the factor analyzed voice models, the loading matrix $\boldsymbol{W}$ corresponds to eigenvoice vectors and we assume that the loading matrix and noise vector are generated from HMM sets. Figure 2 shows the proposed model structure. In the proposed models, eigenvoice vectors can convert their structure according to various input data. Hence, this model receives states changing similarly to HMM. Furthermore, HMM parameters are generated from liner combination of eigenvoice vectors. As a result, this model is feature generating model with liner transformation.

In the FA voice model, an utterance observation sequence of speaker r is generated from following expression:

$$\boldsymbol{o}^{(r)} = \boldsymbol{W}_{\boldsymbol{S}^{(r)}} \boldsymbol{a}^{(r)} + \boldsymbol{n}_{\boldsymbol{S}^{(r)}}, \tag{6}$$

where $\boldsymbol{o}^{(r)}$ is an observation sequence of speaker r, $\boldsymbol{a}^{(r)}$ is latent variable, and $\boldsymbol{S}^{(r)}$ is HMM path. The loading matrix $\boldsymbol{W}_{\boldsymbol{S}^{(r)}}$ and a noise vector $\boldsymbol{n}_{\boldsymbol{S}^{(r)}}$ depend on state transition. The latent variable is prepared for each speaker and represents speaker qualities.

In this model, the likelihood function for utterances of all speakers is written as

$$\begin{aligned} P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) = \prod_r \sum_{\boldsymbol{S}^{(r)}} \int & P(\boldsymbol{o}^{(r)} \mid \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}, \boldsymbol{\Lambda}) \\ & \times P(\boldsymbol{a}^{(r)} \mid \boldsymbol{\Lambda}) P(\boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda}) d\boldsymbol{a}^{(r)}, \end{aligned} \tag{7}$$

$$\begin{aligned} & P(\boldsymbol{o}^{(r)} \mid \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}, \boldsymbol{\Lambda}) \\ & \quad = \mathcal{N}(\boldsymbol{o}^{(r)} \mid \boldsymbol{W}_{\boldsymbol{S}^{(r)}} \boldsymbol{a}^{(r)} + \boldsymbol{\mu}_{\boldsymbol{S}^{(r)}}, \boldsymbol{\Sigma}_{\boldsymbol{S}^{(r)}}), \end{aligned} \tag{8}$$

where $\boldsymbol{\Lambda}$ is model parameter and $\boldsymbol{\mu}_{\boldsymbol{S}^{(r)}}$, $\boldsymbol{\Sigma}_{\boldsymbol{S}^{(r)}}$ are mean and variance of noise vector $\boldsymbol{n}_{\boldsymbol{S}^{(r)}}$.

3.3. EM algorithm for FA voice models

In the proposed, model parameters are estimated by the expectation maximization (EM) algorithm to maximize the log-likelihood for given training data. $\mathcal{Q}$-function of the EM algorithm is given by

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}') = \sum_r \sum_{\boldsymbol{S}^{(r)}} \int & P(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{o}^{(r)}, \boldsymbol{\Lambda}) \\ & \times \ln P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda}') d\boldsymbol{a}^{(r)}, \end{aligned} \tag{9}$$

where $\boldsymbol{\Lambda}'$ is estimated model parameter. Maximizing $\mathcal{Q}$-function, the log-likelihood is guaranteed to increase or remain unchanged. However, calculating $\mathcal{Q}$-function is very hard, because it needs the summation for all paths $\boldsymbol{S}^{(r)}$ and the integral over the latent variable $\boldsymbol{a}^{(r)}$. Therefore, we introduce the variational methods to the EM algorithm.

In the variational EM algorithm, we approximate the posterior distribution in Eq. (9) by using an arbitrary distribution $Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)})$ and define the lower bound of the log-likelihood:

$$\begin{aligned} & \ln P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) \\ & = \ln \sum_r \sum_{\boldsymbol{S}^{(r)}} \int P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda}) d\boldsymbol{a}^{(r)} \\ & = \ln \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \frac{P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda})}{Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)})} d\boldsymbol{a}^{(r)} \\ & \geq \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \ln \frac{P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda})}{Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)})} d\boldsymbol{a}^{(r)} \\ & = \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \ln P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda}) d\boldsymbol{a}^{(r)} \\ & \quad - \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \ln Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) d\boldsymbol{a}^{(r)} \\ & = \mathcal{F}(Q, \boldsymbol{\Lambda}), \end{aligned} \tag{10}$$

where Jensen's inequality is applied. The difference between $\ln P(\boldsymbol{o} \mid \boldsymbol{\Lambda})$ and $\mathcal{F}$ is given by Kullback-Leibler divergence between $Q(\boldsymbol{a}, \boldsymbol{S})$ and the true posterior distribution $P(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{o}^{(r)}, \boldsymbol{\Lambda})$. The E-step computes the posterior probabilities $Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)})$ which maximizes $\mathcal{F}$.

In this paper, we assume the following constraints:

$$Q(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) = Q(\boldsymbol{a}^{(r)})Q(\boldsymbol{S}^{(r)}), \tag{11}$$

$$\sum_{\boldsymbol{S}^{(r)}} Q(\boldsymbol{S}^{(r)}) = 1\ , \ \int Q(\boldsymbol{a}^{(r)})d\boldsymbol{a} = 1. \tag{12}$$

Based on these constraints, we use Lagrange multiplier method and obtain following $Q(\boldsymbol{a}^{(r)})$ and $Q(\boldsymbol{S}^{(r)})$:

$$Q(\boldsymbol{a}^{(r)}) \propto P(\boldsymbol{a}^{(r)} \mid \boldsymbol{\Lambda}) \exp \left\langle \log P(\boldsymbol{o}^{(r)} \mid \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{S}^{(r)})}, \tag{13}$$

$$Q(\boldsymbol{S}^{(r)}) \propto P(\boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda}) \exp \left\langle \log P(\boldsymbol{o}^{(r)} \mid \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{a}^{(r)})}, \tag{14}$$

where $\langle \cdot \rangle_Q$ represents an expectation with respect to a distribution Q. The structure of Eq. (14) includes an expectation of a latent variable. Furthermore, Eq. (13) can be transformed into a Gaussian distribution. The mean $\boldsymbol{\mu}_{\boldsymbol{a}}^{(r)}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{a}}^{(r)}$ of Eq. (13) are given by

$$\boldsymbol{\mu}_{\boldsymbol{a}}^{(r)} = \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(r)} \left(\sum_k \boldsymbol{W}_k^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{U}_k^{(r)} - N_k^{(r)} \boldsymbol{\mu}_k) \right), \tag{15}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{a}}^{(r)} = \left(\boldsymbol{I} + \sum_k N_k^{(r)} \boldsymbol{W}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{W}_k \right)^{-1}. \tag{16}$$

We define following equations:

$$N_k^{(r)} = \sum_t \left\langle S_t^{(r)}, k \right\rangle\ ,\ \boldsymbol{U}_k^{(r)} = \sum_t \left\langle S_t^{(r)}, k \right\rangle \boldsymbol{o}_t^{(r)}, \tag{17}$$

where $\langle S_t^{(r)}, k \rangle$ is a probability of staying in state k at time t given observation $\boldsymbol{o}$.

In the M-step, the model parameter $\boldsymbol{\Lambda}$ is calculated for maximizing the lower bound $\mathcal{F}$. For simplifying, we define the following equations.

$$\bar{\boldsymbol{a}}^{(r)} = [\ 1 \quad \boldsymbol{a}^{(r)\top}\ ]^\top, \tag{18}$$

$$\bar{\boldsymbol{W}}_{\boldsymbol{s}^{(r)}} = [\ \boldsymbol{\mu}_{\boldsymbol{S}^{(r)}} \quad \boldsymbol{W}_{\boldsymbol{S}^{(r)}}\ ], \tag{19}$$

The lower bound $\mathcal{F}(Q, \boldsymbol{\Lambda})$ is partial differentiated about each model parameter and we obtain following equations for re-estimation:

$$\bar{\boldsymbol{W}}_k = \left(\sum_r \boldsymbol{U}_k^{(r)} \left\langle \bar{\boldsymbol{a}}^{(r)\top} \right\rangle \right) \left(\sum_r N_k^{(r)} \left\langle \bar{\boldsymbol{a}}^{(r)} \bar{\boldsymbol{a}}^{(r)\top} \right\rangle \right)^{-1}, \tag{20}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_r N_k^{(r)}} \sum_r \left(\boldsymbol{V}_k^{(r)} - \boldsymbol{U}_k^{(r)} \left\langle \bar{\boldsymbol{a}}^{(r)} \right\rangle^\top \bar{\boldsymbol{W}}_k^\top \right), \tag{21}$$

where

$$\boldsymbol{V}_k^{(r)} = \sum_t \left\langle S_t^{(r)}, k \right\rangle \boldsymbol{o}_t^{(r)} \boldsymbol{o}_t^{(r)\top}, \tag{22}$$

$$\langle \bar{\boldsymbol{a}}^{(r)} \rangle = \left[1 \quad \boldsymbol{\mu}_{\boldsymbol{a}}^{(r)\top} \right]^\top, \tag{23}$$

$$\langle \bar{\boldsymbol{a}}^{(r)} \bar{\boldsymbol{a}}^{(r)\top} \rangle = \begin{bmatrix} 1 & \boldsymbol{\mu}_{\boldsymbol{a}}^{(r)\top} \\ \boldsymbol{\mu}_{\boldsymbol{a}}^{(r)} & \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(r)} + \boldsymbol{\mu}_{\boldsymbol{a}}^{(r)} \boldsymbol{\mu}_{\boldsymbol{a}}^{(r)\top} \end{bmatrix}. \tag{24}$$

The variational EM algorithm iteratively maximizes $\mathcal{F}$ with respect to the $\mathcal{Q}$ and $\boldsymbol{\Lambda}$ holding the other parameters fixed:

$$(\text{E step}) \quad : \quad Q^{(k+1)} = \arg\max_Q \mathcal{F}(Q, \boldsymbol{\Lambda}^{(k)}), \tag{25}$$

$$(\text{M step}) \quad : \quad \boldsymbol{\Lambda}^{(k+1)} = \arg\max_{\boldsymbol{\Lambda}} \mathcal{F}(Q^{(k+1)}, \boldsymbol{\Lambda}), \tag{26}$$

In this procedure, the lower bound $\mathcal{F}$ is guaranteed to increase instead of the value of the $\mathcal{Q}$-function.

3.4. Context clustering

The proposed models can make arbitrary shared structure for each model. Therefore, we apply some questions to each node for increasing the log-likelihood and make clusters like HMM. In this paper, we make each model has same structure. In this clustering method, it is easy to re-estimate model parameters simultaneously because each model have a same structure. Calculating the log-likelihood directly costs a lot of computational cost. Hence, we use the lower bound of the log-likelihood $\mathcal{F}$ in Eq. (10).

4. ANNEALING BASED ESTIMATION ALGORITHM FOR FA VOICE MODELS

In this paper, we propose the DAEM algorithm [5] as annealing based estimation algorithm. Applying the DAEM algorithm to the FA voice model, the free energy function $\mathcal{L}_\beta$ can be defined as

$$\mathcal{L}_\beta = -\frac{1}{\beta} \ln \sum_r \sum_{\boldsymbol{S}^{(r)}} \int P\left(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} | \boldsymbol{\Lambda}\right)^\beta d\boldsymbol{a}^{(r)}, \tag{27}$$

where $\frac{1}{\beta}$ is called "temperature" and β is temperature parameter. Since the negative free energy function corresponds to the log-likelihood function, the lower bound of $-\mathcal{L}_\beta$ is defined as

$$\begin{aligned} &-\mathcal{L}_\beta \\ &\geq \frac{1}{\beta} \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q_\beta(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \ln P(\boldsymbol{o}^{(r)}, \boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)} \mid \boldsymbol{\Lambda})^\beta d\boldsymbol{a}^{(r)} \\ &\quad -\frac{1}{\beta} \sum_r \sum_{\boldsymbol{S}^{(r)}} \int Q_\beta(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) \ln Q_\beta(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)}) d\boldsymbol{a}^{(r)} \\ &= \mathcal{F}_\beta(Q, \boldsymbol{\Lambda}), \end{aligned} \tag{28}$$

where $Q_\beta(\boldsymbol{a}^{(r)}, \boldsymbol{S}^{(r)})$ is the approximated posterior distribution similarly to the variational EM algorithm. $Q_\beta(\boldsymbol{a}^{(r)})$ can be derived as the Gaussian distribution which has following mean $\boldsymbol{\mu}_{\boldsymbol{a}\beta}^{(r)}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{a}\beta}^{(r)}$:

$$\boldsymbol{\mu}_{\boldsymbol{a}\beta}^{(r)} = \beta \boldsymbol{\Sigma}_{\boldsymbol{a}\beta}^{(r)} \left(\sum_k \boldsymbol{W}_k^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{U}_k^{(r)} - N_k^{(r)} \boldsymbol{\mu}_k) \right), \tag{29}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{a}\beta}^{(r)} = \frac{1}{\beta} \left(\boldsymbol{I} + \sum_k N_k^{(r)} \boldsymbol{W}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{W}_k \right)^{-1}. \tag{30}$$

These equations have the structure which adds the temperature parameter β to the approximated posterior distribution $Q(\boldsymbol{a}^{(r)})$ in the variational EM algorithm. Furthermore, $\mathcal{F}_\beta$ is partial differentiated by model parameter and we can obtain a following equation:

$$\frac{\partial \mathcal{F}_\beta}{\partial \boldsymbol{\Lambda}} = \frac{\partial \mathcal{F}}{\partial \boldsymbol{\Lambda}} = \boldsymbol{0}. \tag{31}$$

This equation equivalent to the updating equation of the variational EM algorithm. In the DAEM algorithm, the temperature parameter β is gradually increased while iterating the EM-steps at each temperature. When $\frac{1}{\beta}$ is set to an initial temperature $\beta \simeq 0$, the EM-steps may achieve a single global maximum of $\mathcal{F}_\beta$. Finally at the $\beta = 1$, $\mathcal{F}_\beta$ is identical with the variational EM algorithm.

5. EXPERIMENTS

5.1. Experimental conditions

For training speaker independent HMM sets, we used ATR Japanese speech database B-set uttered by six male and four female speakers. Nine speakers excluding one male speaker were used for the experiment. Each speaker uttered 503 sentences: different 50 sentences were used for each speaker in the training, and the remaining 53 sentences were used for evaluation. The speech data was down-sampled from 20kHz to 16kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window. Feature vectors consisted of spectral and F_0 feature vectors. Each spectral feature vector consisted of 24 mel-cepstral coefficients and their delta and delta-delta coefficients. The F_0 parameter vectors consisted of log F_0, its delta and delta-delta. A left-to-right, 5-state, MSD-HMM with no skip structure was used. The number of eigenvoice vectors and the number of factors in the proposed method are both two. For the DAEM algorithm, the temperature parameter β was updated by

$$\beta(i) = \left(\frac{i}{20}\right)^2, \ (i = 0, 1, \ldots, 20), \tag{32}$$

where i denotes the iteration number. At each temperature, 20 EM-steps were conducted.

To evaluate the performance of the FA voice models, the following four training methods were compared:

- "PCA": the PCA based eigenvoice model with the model structure obtained by speaker independent HMMs.
- "PCA_STC": the PCA based eigenvoice model with the model structure obtained by STC [6].
- "FA_EM": the FA voice model initialized by speaker independent HMMs, and the EM algorithm was used.
- "FA_DAEM": the FA voice model initialized by speaker independent HMMs, and the DAEM algorithm was used.

5.2. Experimental results

A subjective listening test was conducted to evaluate quality of synthesized speech. The test compared the naturalness of synthesized speech by the mean opinion score (MOS) test method. The subjects were 10 Japanese graduate students. Speech samples were randomly chosen from the evaluation sentences. Voice characteristics of speech samples were average voice and that of nine speakers in the training data, and two sentences for each characteristic, in total 20 sentences were prepared for each subject. In the MOS test, after listening to each test sample, the subjects were asked to assign it a 5-point naturalness score (5:excellent, 4:good, 3:fair, 2:poor, 1:bad).

Figure 3 plots the experimental results. It can be seen from the figure that the proposed methods "FA_EM" and "FA_DAEM" achieved better subjective scores than the conventional methods "PCA" and "PCA_STC." Since the FA voice models were estimated from training data directly, appropriate model parameters were obtained for representing speech utterances accurately. It is also the reason of the improvement that the context clustering was performed based on the same criterion as the parameter estimation. It can also be seen that "FA_DAEM" obtained a better subjective score than "FA_EM." Because the DAEM algorithm can improve the local maxima problem. These results clearly show the effectiveness of the proposed method in speech synthesis and the DAEM algorithm is more effective than the EM algorithm for estimating model parameters.

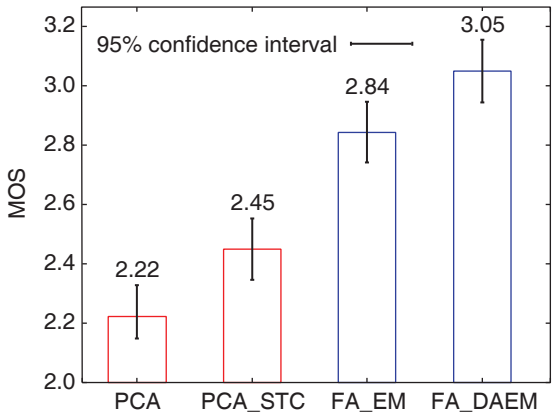

Fig. 3. Mean opinion scores of synthesized speech with 95% confidence intervals by the conventional and proposed methods

6. CONCLUSION

This paper proposed the FA voice model for HMM-based speech synthesis. It can synthesize speech with various voice characteristics. This method can estimate more appropriate model parameters than eingevoice method based on PCA. We also derived the EM and DAEM algorithm for the proposed method. In the experiments, the proposed method achieved a higher performance than the conventional eigenvoice method. Furthermore, the DAEM algorithm improved the performance of the proposed method. Experiments on larger datasets and evaluation of synthesized speech with various voice characteristics will be future work.

7. ACKNOWLEDGEMENT

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

8. REFERENCES

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. 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," Proc. of EUROSPEECH, pp.2347–2350, 1999.

[3] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," Proc. of ICSLP, vol.1, pp.1269–1272, 2002.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for Graphical models," Machine Learing, vol.37, pp.183–233, 1997.

[5] N. Ueda and R. Nakano, "Determinisic annealing EM algorithm," Neural Networks, vol.11, no.2, pp.271–282, 1998

[6] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice model in HMM-based speech synthesis," Proc. of ICSLP, pp. 133–136, 2002.