PAPER

# A Bayesian Framework Using Multiple Model Structures for Speech Recognition

**Sayaka SHIOTA**[†a)], **Kei HASHIMOTO**[†b)], ***Nonmembers***, **Yoshihiko NANKAKU**[†c)], ***and*** **Keiichi TOKUDA**[†d)], ***Members***

**SUMMARY** This paper proposes an acoustic modeling technique based on Bayesian framework using multiple model structures for speech recognition. The aim of the Bayesian approach is to obtain good prediction of observation by marginalizing all variables related to generative processes. Although the effectiveness of marginalizing model parameters was recently reported in speech recognition, most of these systems use only "one" model structure, e.g., topologies of HMMs, the number of states and mixtures, types of state output distributions, and parameter tying structures. However, it is insufficient to represent a true model distribution, because a family of such models usually does not include a true distribution in most practical cases. One of solutions of this problem is to use multiple model structures. Although several approaches using multiple model structures have already been proposed, the consistent integration of multiple model structures based on the Bayesian approach has not seen in speech recognition. This paper focuses on integrating multiple phonetic decision trees based on the Bayesian framework in HMM based acoustic modeling. The proposed method is derived from a new marginal likelihood function which includes the model structures as a latent variable in addition to HMM state sequences and model parameters, and the posterior distributions of these latent variables are obtained using the variational Bayesian method. Furthermore, to improve the optimization algorithm, the deterministic annealing EM (DAEM) algorithm is applied to the training process. The proposed method effectively utilizes multiple model structures, especially in the early stage of training and this leads to better predictive distributions and improvement of recognition performance.

***key words:*** *speech recognition, acoustic modeling, Bayesian approach, model structure integration, deterministic annealing*

## 1. Introduction

The maximum likelihood (ML) criterion has usually been used for training statistical models for speech recognition systems. However, since the ML criterion produces a point estimate of model parameters, the estimation accuracy may degrade when little training data is available. The aim of the Bayesian approach is to obtain good prediction of observation by marginalizing all variables related to generative processes, and it can accurately estimate observation distributions even if the amount of training data is small. However, the calculation becomes complicated due to the combination of latent variables, i.e., state sequences and model parameters. To solve this problem, the variational Bayesian (VB) method has been proposed as an effective approximation method of the Bayesian approach [1], [2], and the effectiveness of marginalizing model parameters was recently reported in speech recognition [3]–[7].

In conventional speech recognition based on generative models, there are many efforts to find appropriate model structures to predict observation vector sequences (e.g., multi-mixture models, clustering techniques and more complicated models). However, most of these systems use only "one" model structure, e.g., topologies of HMMs, the number of states and mixtures, types of state output distributions, and parameter tying structures. In most practical cases, it is insufficient to represent a true model distribution because a family of such models usually does not include a true distribution. One of solutions of this problem is to use multiple model structures. Although several approaches using multiple model structures have already been proposed, e.g., ROVER [8], random forest [9] and model structure annealing [10], the consistent integration of the multiple model structures based on the Bayesian approach has not seen in speech recognition. This paper focuses on integrating the multiple model structures based on the Bayesian framework in HMM based acoustic modeling. The proposed method is derived from a new marginal likelihood function which includes the model structures as a latent variable in addition to HMM state sequences and model parameters, and the posterior distributions of these latent variables are obtained using the VB method.

The conventional VB method sometimes suffers from the local maxima problem because of the combination of the latent variables. To overcome this problem, some approaches have been reported [11], [12], and we have also reported the training algorithm applying the deterministic annealing EM (DAEM) algorithm [13] to the conventional VB method for speech recognition system [14]. Since the proposed method also treats the multiple model structures as a latent variable additionally, the local maxima problem becomes more serious than the conventional VB method. Therefore, to improve the training algorithm of the proposed method, the deterministic annealing framework is applied to the training process. The proposed method which applied the deterministic annealing process effectively utilizes multiple model structures, especially in the early stage of training and this leads to better predictive distributions and improvement of recognition performance.

The proposed method has a similarity to the non-

Manuscript received June 15, 2012.

Manuscript revised November 14, 2012.

[†] The author are with the Department of Computer Science, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

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

b) E-mail: bonanza@sp.nitech.ac.jp

c) E-mail: nankaku@sp.nitech.ac.jp

d) E-mail: tokuda@nitech.ac.jp

DOI: 10.1587/transinf.E96.D.939

parametric Bayesian method [15] because both methods use multiple model structures and integrate them based on the Bayesian framework. The main difference between these methods is that the non-parametric Bayesian method assumes generating processes of multiple model structures for each data sample. Although the proposed method simply prepared multiple model structures in advance, it still has the effect of model structure marginalization and can be performed without increasing the complexity of the training process.

The rest of this paper is organized as follows. Section 2 describes the speech recognition based on the variational Bayesian approach. The Bayesian speech recognition including multiple model structures and the training algorithm using the DAEM algorithm are described in Sect. 3. Section 4 illustrates results of the continuous phoneme recognition experiments, and the final section presents conclusions and future work.

## 2. Speech Recognition Based on Variational Bayesian Method

### 2.1 Bayesian Approach

The output distribution is obtained based on a left-to-right HMM which has been widely used to represent an acoustic model for speech recognition. Let $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_T)$ be a set of training data, and $T$ denotes the frame number. The likelihood function of an HMM are represented by

$$\begin{aligned} P(\boldsymbol{O} \mid \boldsymbol{\Lambda}) &= \sum_{\boldsymbol{Z}} P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) & (1) \\ &= \sum_{\boldsymbol{Z}} \prod_{t=1}^{T} a_{z_{t-1}z_t} \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{z_t}, \boldsymbol{S}_{z_t}^{-1}), & (2) \end{aligned}$$

where $\boldsymbol{Z} = (z_1, z_2, \dots, z_T)$ is a sequence of HMM states, $z_t \in \{1, \dots, N\}$ denotes a state at frame $t$, and $N$ is the number of states in an HMM. A set of model parameters $\boldsymbol{\Lambda} = \{\pi_i, a_{ij}, \boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i,j=1}^{N}$ consists of the intial state probability $\pi_i$ of state $i$, the state transition probability $a_{ij}$ from state $i$ to state $j$, the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{S}_i^{-1}$ of a Gaussian distribution $\mathcal{N}(\cdot \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1})$, Note that the initial state probability of an HMM is represented by $a_{z_0z_1}$. Although Gaussian mixture model is typically used for output probabilities in many systems, this paper assumes a single Gaussian distribution for the simplicity of description. However, assuming that $\boldsymbol{Z}$ includes the index sequences of the mixture components as well as the state index sequences, the following equations can be easily extended to those of the Gaussian mixture models.

The Bayesian approach assumes that the model parameter $\boldsymbol{\Lambda}$ is a latent variable, while the ML approach estimates constant model parameters. The posterior distribution for a set of model parameters $\boldsymbol{\Lambda}$ is obtained with the Bayes theorem as follows:

$$P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) = \sum_{\boldsymbol{Z}} \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda})}{P(\boldsymbol{O})}\ , \qquad (3)$$

where $P(\boldsymbol{\Lambda})$ is a prior distribution for $\boldsymbol{\Lambda}$. Once the posterior distribution $P(\boldsymbol{\Lambda} \mid \boldsymbol{O})$ is estimated, a predictive distribution for input data $\boldsymbol{X}$ is represented by

$$P(\boldsymbol{X} \mid \boldsymbol{O}) = \sum_{\boldsymbol{Z}_x} \int P(\boldsymbol{X}, \boldsymbol{Z}_x \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda} \mid \boldsymbol{O})d\boldsymbol{\Lambda}. \qquad (4)$$

Since the model parameters are marginalized out in Eq. (4), the effect of over-fitting is mitigated. However, it is difficult to solve the integral and expectation calculations. Especially, when the model includes latent variables, the calculation becomes more complicated. To overcome this problem, the variational Bayesian (VB) method has been proposed as a tractable approximation method of the Bayesian approach [1].

### 2.2 Variational Bayesian Method

In the Bayesian method, the marginal likelihood[†] is represented by

$$\log P(\boldsymbol{O}) = \log \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda})\, d\boldsymbol{\Lambda}. \qquad (5)$$

However, the calculation of Eq. (5) requires averaging over all configurations of the latent variables. Therefore, in the variational Bayesian (VB) method, a lower bound of the log marginal likelihood $\mathcal{F}$ is maximized instead of the true likelihood. The lower bound of the log marginal likelihood $\mathcal{F}$ is defined by using Jensen's inequality:

$$\begin{aligned} \log P(\boldsymbol{O}) &= \log \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})}\, d\boldsymbol{\Lambda} \\ &\geq \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})}\, d\boldsymbol{\Lambda} \\ &= \mathcal{F}, \qquad (6) \end{aligned}$$

where $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ is an arbitrary distribution. Then, the relation between the log marginal likelihood and the lower bound $\mathcal{F}$ is represented from Eq. (6).

$$\begin{aligned} \log P(\boldsymbol{O}) - \mathcal{F} &= \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})}{P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})} d\boldsymbol{\Lambda}, \\ &= \mathrm{KL}[Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \mid P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})], \qquad (7) \end{aligned}$$

where $\mathrm{KL}[Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \mid P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})]$ denotes the Kullback-Leibler (KL) divergence [16] between $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ and the true posterior distribution $P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})$. As the difference between the true log marginal likelihood and the lower bound is reduced, $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ approximates the true posterior distribution $P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})$. Therefore, by maximizing the lower bound $\mathcal{F}$, the optimal posterior distribution $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ is estimated. However, the calculation becomes complicated because of the combination of latent variables and $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ include the integration of the model parameter. To obtain

[†] The marginal likelihood is corresponded to the free energy function in statistical mechanics.

approximate posterior distributions $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$, the variational method is applied. In the variational method, the latent variables are assumed conditionally independent[†] each other as follows:

$$Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) = Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda}). \tag{8}$$

Under this assumption, the optimal VB posterior distributions which maximize the objective function $\mathcal{F}$ are given by the variational method:

$$Q(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp\left\{\langle \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})\rangle_{Q(\boldsymbol{\Lambda})}\right\}, \tag{9}$$

$$Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) \exp\left\{\langle \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})\rangle_{Q(\boldsymbol{Z})}\right\}, \tag{10}$$

where $\langle \cdot \rangle_Q$ denotes the expectation with respect to $Q$, $C_{\boldsymbol{Z}}$ and $C_{\boldsymbol{\Lambda}}$ are the normalization terms of $Q(\boldsymbol{Z})$ and $Q(\boldsymbol{\Lambda})$, respectively. Since the obtained VB posterior distributions $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ are dependent on each other, these updates should be iterated as the EM algorithm, which increases the value of the objective function $\mathcal{F}$ at each iteration until convergence.

Although the Bayesian approach achieved higher performance than the ML approach [3], the local maxima problem in the Bayesian approach is more serious than in the ML-based approach because the Bayesian approach treats not only state sequences but also model parameters as latent variables. Therefore, the optimization algorithm is important for the VB method. To optimize the training algorithm, we applied the DAEM algorithm to the VB method [14].

### 2.3 DAEM Algorithm for Variational Bayesian Method

To adopt the DAEM algorithm [13] to the variational Bayesian method, the lower bound of the log marginal likelihood $\mathcal{F}$ (Eq. (6)) can be written as follows:

$$\mathcal{F} = \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} - \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) d\boldsymbol{\Lambda}. \tag{11}$$

From the view of statistical physics, the lower bound F corresponds to the negative of the Helmholtz free energy with temperature $\mathcal{T} = 1$:

$$\mathcal{F}_{Helmholtz} = E - \mathcal{T}\mathcal{S}, \tag{12}$$

where $E$ and $\mathcal{S}$ denote the internal energy and the entropy, corresponding to the first and second term of Eq. (11) respectively. For the DAEM algorithm, the inverse temperature parameter $\beta = 1/\mathcal{T}$ is introduced and a new objective function $\mathcal{F}_\beta$ is defined:

$$\mathcal{F}_\beta = \frac{1}{\beta} \sum_{\boldsymbol{Z}} \int \hat{Q}(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) P^\beta(\boldsymbol{\Lambda})}{\hat{Q}(\boldsymbol{Z}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda}. \tag{13}$$

This function can be regarded as the lower bound of the following function by using Jensen's inequality. Therefore, the marginal likelihood function based on the DAEM algorithm $\mathcal{L}_\beta$ can also be redefined:

$$\mathcal{L}_\beta = \frac{1}{\beta} \log \sum_{\boldsymbol{Z}} \int P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) P^\beta(\boldsymbol{\Lambda})\, d\boldsymbol{\Lambda}. \tag{14}$$

To obtain the VB posterior distributions, the constraint ($\hat{Q}(\boldsymbol{Z}, \boldsymbol{\Lambda}) = \hat{Q}(\boldsymbol{Z})\hat{Q}(\boldsymbol{\Lambda})$) is applied to the lower bound $\mathcal{F}_\beta$. Under the constraint, the optimal VB posterior distributions which maximize the lower bound can be obtained as follows:

$$\hat{Q}(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp\left\{\left\langle \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})\right\rangle_{\hat{Q}(\boldsymbol{\Lambda})}\right\}, \tag{15}$$

$$\hat{Q}(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P^\beta(\boldsymbol{\Lambda}) \exp\left\{\left\langle \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})\right\rangle_{\hat{Q}(\boldsymbol{Z})}\right\}. \tag{16}$$

By applying the deterministic annealing framework to the VB method, the temperature parameter $\beta$ is attached to the original VB posterior distributions (Eqs. (9) and (10)). In the deterministic annealing process, since temperature denotes $1/\beta$, the temperature parameter $\beta$ is gradually increased and the form of the VB posterior distributions depends on each temperature parameter. When $\beta$ is set to the initial temperature $\beta^{(0)} \simeq 0$, the VB posterior distributions $\hat{Q}(\boldsymbol{Z})$ and $\hat{Q}(\boldsymbol{\Lambda})$ take a form nearly uniform distribution. While the temperature is decreasing, the form of $\hat{Q}(\boldsymbol{Z})$ and $\hat{Q}(\boldsymbol{\Lambda})$ become close to each original posterior distribution. Finally at the temperature $\beta = 1$, $\hat{Q}(\boldsymbol{Z})$ and $\hat{Q}(\boldsymbol{\Lambda})$ take each original posterior distribution, and the reliable model parameters can be estimated without the effect of the local maxima problem.

## 3. Bayesian Speech Recognition Using Multiple Model Structures

Recently, to improve the model complexity, some approaches were reported using multiple model structures (e.g., random forest [9], ROVER [8], and the model structure annealing [10]). Although various integration techniques and criteria can be considered, this paper focuses on model structure integration based on the Bayesian framework in acoustic modeling.

### 3.1 Marginal Likelihood Function Including Multiple Model Structures

For considering the proposed framework of using multiple model structures based on the Bayesian approach for speech recognition, we define a marginal likelihood function which includes model structures as a latent variable as follows:

$$\log P(\boldsymbol{O}) = \log \sum_m \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z}, m, \boldsymbol{\Lambda}_m) d\boldsymbol{\Lambda}_{\boldsymbol{m}}, \tag{17}$$

[†] "Conditionally independent" means that the distribution is assumed to be independent under the condition that observation $\boldsymbol{O}$ and structure $m$ are given. Even the variables of the prior distribution (distribution without the condition) are independent, the variables of the posterior distribution (distribution with the condition) are usually dependent each other by the condition. Therefore the conditional independent assumption (Eq. (8)) can be regarded as a kind of approximation.

$$P(\boldsymbol{O}, \mathbf{Z}, m, \boldsymbol{\Lambda}_m) = P(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) P(\boldsymbol{\Lambda}_m \mid m) P(m), \quad (18)$$

where $m \in \{1, \ldots, M\}$ indexes model structures, $\boldsymbol{\Lambda}_m \in \{\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_M\}$ denotes a set of model parameters for the $m$-th model structure, and the prior distribution $P(\boldsymbol{\Lambda}_m \mid m)$ is calculated from each model structure $m$. Note that this paper regards a structure of a phonetic decision tree as a model structure. In Eq. (18), the state sequence $\mathbf{Z}$ is not dependent of the model structures $m$. This means that the state sequences are estimated from a combination of the multiple model structures, and it is expected reliable posterior distributions of state sequences are estimated. However, the proposed method also treats the model structures as a latent variable, the local maxima problem is more serious than the conventional Bayesian method. Therefore, the proposed framework should adopt the deterministic annealing process for the training algorithm.

## 3.2 Applying DAEM Algorithm to Proposed Framework

We redefine the free energy function which based on the marginal likelihood function (Eq. (17)) as follows:

$$\bar{\mathcal{L}}_\beta = \frac{1}{\beta} \sum_m \sum_{\mathbf{Z}} \int \log P^\beta(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) \times P^\beta(\boldsymbol{\Lambda}_m \mid m) P^\beta(m) d\boldsymbol{\Lambda}_{\boldsymbol{m}}. \quad (19)$$

The difference of the new free energy function from Eq. (14) is that model structure $m$ is added as a latent variable. The lower bound of the free energy function $\bar{\mathcal{L}}_\beta$ is defined by using Jensen's inequality:

$$\bar{\mathcal{F}}_\beta = \frac{1}{\beta} \sum_m \sum_{\mathbf{Z}} \int \tilde{Q}(\mathbf{Z}, m, \boldsymbol{\Lambda}_m) \times \log \frac{P^\beta(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) P^\beta(\boldsymbol{\Lambda}_m \mid m) P^\beta(m)}{\tilde{Q}(\mathbf{Z}, m, \boldsymbol{\Lambda}_m)} d\boldsymbol{\Lambda}_{\boldsymbol{m}}. \quad (20)$$

An arbitrary distribution $\tilde{Q}(\mathbf{Z}, m, \boldsymbol{\Lambda}_m)$ has a combination of the three latent variables, and it makes the objective function more complicated than the conventional method which uses only one model structure. To obtain approximate posterior distributions, we assume the following constraint:

$$\tilde{Q}(\mathbf{Z}, m, \boldsymbol{\Lambda}_m) = \tilde{Q}(\mathbf{Z}) \tilde{Q}(m) \tilde{Q}(\boldsymbol{\Lambda}_m \mid m). \quad (21)$$

Note that the dependency between model parameters and model structures remains as the prior distribution in Eq. (18). Under this constraint, by maximizing the lower bound $\bar{\mathcal{F}}_\beta$, the optimal posterior distributions $\tilde{Q}(\mathbf{Z}), \tilde{Q}(m)$ and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ are obtained:

$$\tilde{Q}(\mathbf{Z}) = C_{\mathbf{Z}} \exp\left\{ \left\langle \left\langle \log P^\beta(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) \right\rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)} \right\rangle_{\tilde{Q}(m)} \right\}, \quad (22)$$

$$\tilde{Q}(m) = C_m P^\beta(m) \exp\left\{ \left\langle \left\langle \left\langle \log P^\beta(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) \right\rangle_{\tilde{Q}(\mathbf{Z})} + \log \frac{P^\beta(\boldsymbol{\Lambda}_m \mid m)}{\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)} \right\rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m)} \right. \right\}, \quad (23)$$

$$\tilde{Q}(\boldsymbol{\Lambda}_m \mid m) = C_{\boldsymbol{\Lambda}_m} P^\beta(\boldsymbol{\Lambda}_m \mid m) \times \exp\left\{ \left\langle \log P^\beta(\boldsymbol{O}, \mathbf{Z} \mid m, \boldsymbol{\Lambda}_m) \right\rangle_{\tilde{Q}(\mathbf{Z})} \right\}. \quad (24)$$

From Eqs. (22), (23) and (24), since the optimal variational posterior distributions $\tilde{Q}(\mathbf{Z})$, $\tilde{Q}(m)$ and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ depend on each other, these distributions should be updated iteratively in the deterministic annealing framework. Since a set of model parameter $\boldsymbol{\Lambda}_m$ consist of the parameters of transition probability $\boldsymbol{\Lambda}^{(a)}$ and output probability $\boldsymbol{\Lambda}_m^{(b)}$, Eqs. (22) and (24) are rewritten as:

$$\tilde{Q}(\mathbf{Z}) = C_{\mathbf{Z}} \exp\left\langle \log P^\beta(\mathbf{Z} \mid \boldsymbol{\Lambda}^{(a)}) \right\rangle_{\tilde{Q}(\boldsymbol{\Lambda}^{(a)})} \times \prod_m \left[ \exp\left\langle \log P^\beta(\boldsymbol{O} \mid \mathbf{Z}, m, \boldsymbol{\Lambda}_m^{(b)}) \right\rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m)} \right]^{\tilde{Q}(m)}, \quad (25)$$

$$\tilde{Q}(\boldsymbol{\Lambda}^{(a)}) = C_{\boldsymbol{\Lambda}^{(a)}} P^\beta(\boldsymbol{\Lambda}^{(a)}) \exp\left\langle \log P^\beta(\mathbf{Z} \mid \boldsymbol{\Lambda}^{(a)}) \right\rangle_{Q(\boldsymbol{\Lambda}^{(a)})}, \quad (26)$$

$$\tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) = C_{\boldsymbol{\Lambda}_m^{(b)}} P^\beta(\boldsymbol{\Lambda}_m^{(b)} \mid m) \times \exp\left\langle \log P^\beta(\boldsymbol{O} \mid \mathbf{Z}, m, \boldsymbol{\Lambda}_m^{(b)}) \right\rangle_{\tilde{Q}(\mathbf{Z})}, \quad (27)$$

where the state transition probability $\boldsymbol{\Lambda}^{(a)}$ is independent of the model structure $m$. The VB posterior distribution $\tilde{Q}(\mathbf{Z})$ takes the same form of the posterior distribution based on the ML criterion: $\exp \langle \log P^\beta(\mathbf{Z} \mid \boldsymbol{\Lambda}^{(a)}) \rangle_{\tilde{Q}(\boldsymbol{\Lambda}^{(a)})}$ and $\exp \langle \log P^\beta(\boldsymbol{O} \mid \mathbf{Z}, m, \boldsymbol{\Lambda}_m^{(b)}) \rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m)}$ correspond to the state transition probability and the output probability respectively, and $\tilde{Q}(m)$ can be regarded as a stream weight of multi-stream HMMs. Thus, the state occupancy is calculated from the Forward-Backward algorithm as the standard multi-stream HMMs (This step is called VB-E step). However, contrary to the standard multi-stream HMMs, the proposed method can estimate the stream weights as the update of posterior distribution $\tilde{Q}(m)$ automatically. Updates of $\tilde{Q}(\boldsymbol{\Lambda}^{(a)})$ and $\tilde{Q}(\boldsymbol{\Lambda}^{(b)} \mid m)$ correspond to the M-step in the standard EM algorithm (This step is called VB-M step). Additionally, the concrete forms of $\tilde{Q}(\mathbf{Z})$, $\tilde{Q}(m)$, $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ and the normalization terms is written in the appendix. In the proposed framework, the multiple model structures are previously constructed and the context clustering is not conducted during the annealing process. If infinite number of model structures can be used, the proposed method is theoretically regarded as performing the context clustering at each annealing step. Although the proposed method can use only a finite number of the model structures in practice, the reliable model parameters can be estimated by using the multiple model structures.

Since the proposed framework adopts the deterministic annealing process for the training algorithm, similar to the Sect. 2.3, the temperature parameter $\beta$ is gradually increasing from 0 to 1, and at each temperature the posterior distributions are estimated. Figure 1 shows the training process in the proposed method. As this figure, at the initial temperature ($\beta \simeq 0$), the variational posterior distributions $\tilde{Q}(\mathbf{Z})$, $\tilde{Q}(m)$ and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ take a form nearly uniform distribution. This means that all model structures can be used almost uniformly used for estimating the model parameters

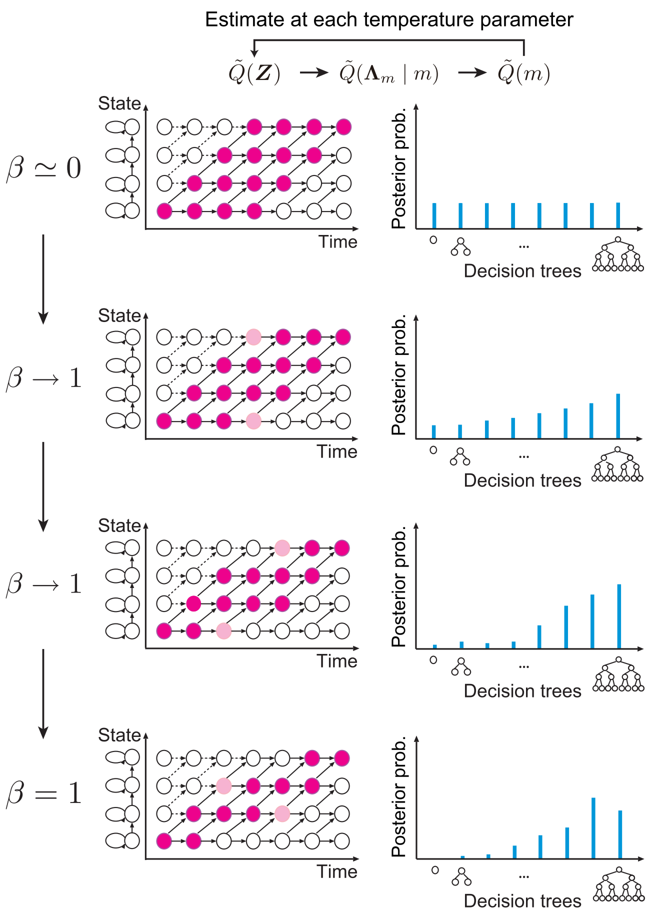

Estimate at each temperature parameter
$\tilde{Q}(\boldsymbol{Z}) \rightarrow \tilde{Q}(\boldsymbol{\Lambda}_m \mid m) \rightarrow \tilde{Q}(m)$
$\beta \simeq 0$
$\beta \rightarrow 1$
$\beta \rightarrow 1$
$\beta = 1$
State
Time
Posterior prob.
Decision trees
...

**Fig. 1** Training process.

and the state sequences in the initial step. While the temperature is decreasing ($\beta \rightarrow 1$), the form of $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$ and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ change to each original posterior distribution. At the final temperature ($\beta = 1$), $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$ and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ take each original posterior distribution. Through this process, the optimal posterior probability of each model structure can be automatically estimated.

### 3.3 Related Approaches

We reported the approximation method of the joint optimization of the state sequences and model structures based on the ML-based speech recognition [10]. This framework depends on the negative free energy function which is defined as follows:

$$\mathcal{L}_{\beta}(\Lambda) = \frac{1}{\beta} \log \sum_{\boldsymbol{Z}} \sum_{m} P^{\beta}(\boldsymbol{O}, \boldsymbol{Z} \mid m, \Lambda) P^{\beta}(m). \qquad (28)$$

Comparing this function with the function of the proposed method, the previous work can be regarded as the the proposed method which produced a point estimation of model parameters. In the ML-based framework, there was a serious problem that accurate posterior probabilities of the model structures cannot be selected automatically. This is because the ML criterion selects the largest model structure, and the largest model structure is not always adequate. On the other hand, the Bayesian criterion can be used to select the adequate model structure [7]. Thus, the proposed method can estimate the adequate posterior distributions of the model structures and be expected to improve the speech recognition performance.

From another point of view, the proposed method has a similarity to the non-parametric Bayesian method [6], [15] because both methods use multiple model structures with different complexities and are integrated based on the Bayesian framework. The main difference between them is that the non-parametric Bayesian method assumes processes to generate multiple model structures for each data sample. Although the proposed method simply prepared multiple model structures, it still has the effect of model structure marginalization and can be performed without increasing the complexity of the training process.

Random forest [9] is one of the techniques using multiple model structures. There are some different points between the random forest (RF) method and the proposed method. One of the different points is how to construct the model structures. The RF method is changing the data set or question set for constructing other model structures. Although the proposed method can also use these methods, we use the Bayesian framework for constructing the adequate model structures. Another point is how to use multiple model structures. In the RF method, several methods of model combination have been tried, because there is no criteria for deciding the combination weights. The proposed method can automatically estimate the posterior probability of each model structure.

In recent state-of-the-art speech recognition systems, discriminative approaches have been used [17], [18]. Contrary to this, the proposed method is based on a generative model of the observations as the conventional HMM based speech recognition. However, the most discriminative approaches use structures of generative statistical models, and finding the appropriate model structures is still essential problem of speech recognition. Therefore, the authors think that the idea of using multiple model structures and integration based on the consistent statistical criterion are useful and available for various approaches including discriminative approaches in future work.

## 4. Experiments

### 4.1 Experimental Conditions

We conducted speaker independent experiments on continuous phoneme recognition to evaluate the effectiveness of the proposed method, where training data from 18,823 Japanese sentences and testing data from 100 sentences were prepared from Japanese Newspaper Article Sentences (JNAS). Speech signals were sampled at a frequency of 16 kHz and windowed at 10-ms frame rates using a 25-ms Hamming window. The spectrum parameter vectors consisted of 12-order MFCC and their delta and delta-delta coefficients. Three-state left-to-right HMMs were used to model triphones consisting of 43 Japanese phonemes and 204 questions were prepared for context clustering. All state output probability distributions were modeled by using a Gaussian distribution with a diagonal covariance matrix. The five algorithms below were compared in this experiment.

- **Flat-start** : HMMs were initialized by flat-start training and trained with the EM algorithm (the EM-steps were iterated 200 times).
- **DAEM** : HMMs were initialized by flat-start training and trained with the DAEM algorithm.
- **Mtree** : HMMs were initialized by flat-start training and trained with the DAEM algorithm with multiple model structures.
- **Label10** : HMMs were initialized with the segmental $k$-means algorithm using phoneme boundary labels and trained with the EM algorithm (the EM-steps were iterated 10 times).
- **Label200** : HMMs were initialized with the segmental $k$-means algorithm using phoneme boundary labels and trained with the EM algorithm (the EM-steps were iterated 200 times).

The ML and Bayes criteria could be applied to all five algorithms, and comparative methods were represented by combining the algorithms and criteria. **Mtree(Bayes)** is the new proposed method and **Mtree(ML)** is the previous method we proposed using the ML criterion reported in [10]. DAEM methods using a single model structure **DAEM(ML)** and **DAEM(Bayes)** were also compared with the proposed method and their details have been reported [14], [19]. Prior distributions and model selection of the Bayesian methods are automatically optimized by using the cross validation [7]. It is desirable to use multiple model structures. However, when the several model structures are used, we need to determine many conditions (e.g., the size and structure of trees, and the number of trees). Although how to determine the number of model structures and how to construct multiple model structures are essential problems for the proposed method, in this experiment, we only focus on the evaluation of the integration part of multiple model structures. Therefore, this experiment simply used only two kinds of model structures for the proposed framework. At first, to prepare a single model structure for the approaches utilizing a single model structure (**Flat-start**, **DAEM**, **Label10**, and **Label200**), two model structures based on the ML and Bayes criteria are constructed as follows:

- ML : a model structure was selected by using the minimum description length (MDL) criterion. This structure had 4,021 leaf nodes.
- Bayes : a model structure was selected by using the Bayesian criterion utilizing 200-folds cross validation [7]. This structure had 18,099 leaf nodes (CV-Bayes).

A model structure representing each monophone model is also prepared for **Mtree(ML)** and **Mtree(Bayes)**. The monophone structure had 129 leaf nodes. Since the annealing process is sensitive to the temperature update function, there are many reports how to determine the annealing schedule [20], [21]. On the other hand, it is empirically known that the exponential function works well without preliminary examinations. Therefore, the exponential function (Eq. (29)) was used in this experiment. The number of temperature parameter updates was set to 20 ($I = 20$), and EM-steps were iterated 10 times at each temperature. The temperature parameter $\beta$ was updated by

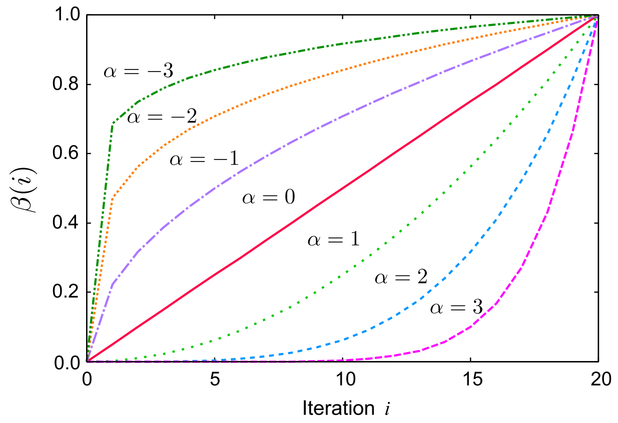


**Fig. 2** Schedule of temperature parameter $\beta$.

$$\beta(i) = \left(\frac{i}{I}\right)^n, \quad i = 0, \ldots, I, \tag{29}$$

where $i$ denotes the number of iterations of temperature updates, and $n$ was varied to $n = 2^\alpha, (\alpha = -3, \ldots, 3)$. Because the EM-steps in **DAEM** were iterated a total of 200 times, the EM-steps in **Flat-start** and **Label200** were iterated 200 times. Since it is difficult to estimate the accurate posterior probabilities of the model structures in **Mtree(ML)**, we heuristically assumed that $Q_{ML}(m)$ would be updated by the following linear functions:

$$Q_{ML}(Monophone) = 0.5\left(1 - \frac{i}{I}\right), \tag{30}$$

$$Q_{ML}(MDL) = 0.5\left(1 + \frac{i}{I}\right). \tag{31}$$

Figures 2 and 3 show plots of the schedules of the temperature parameter $\beta$ and the update schedules of $Q_{ML}(m)$. Note that the proposed method does not require pre-determined posterior probabilities of the model structures such as Eqs. (30) and (31). Note that **Mtree(Bayes)** does not require pre-determined posterior probabilities of the model structures.

### 4.2 Experimental Results

### 4.2.1 Single Mixture Experiment

Figure 4 summarized the upper bounds of the log marginal likelihood $\bar{\mathcal{F}}_\beta$ for the training data. The temperature update schedules were adjusted to obtain the highest marginal

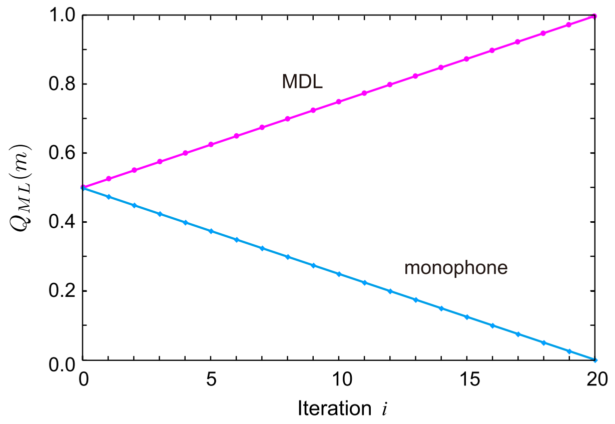


**Fig. 3** Schedule of update $Q_{ML}(m)$.

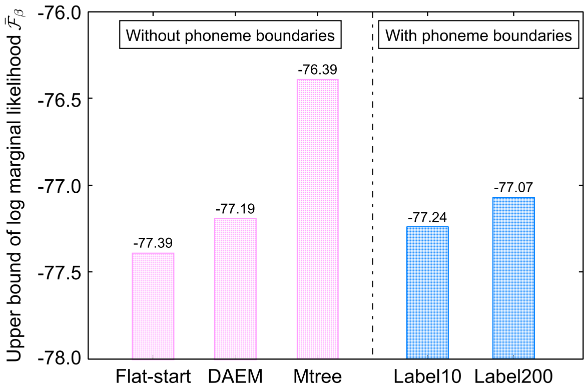


**Fig. 4** Upper bound of log marginal likelihood $\bar{\mathcal{F}}_\beta$.

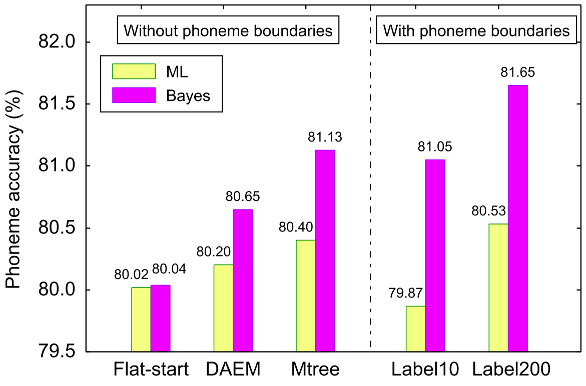


**Fig. 5** Phoneme accuracy.

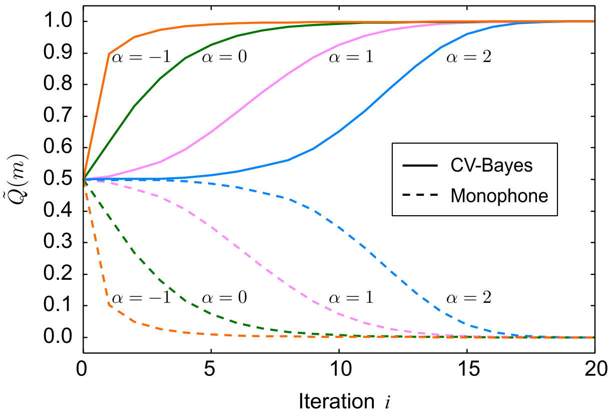


**Fig. 6** Posterior distributions of model structures.

likelihood ($\alpha = 0$). The table indicates that the marginal likelihood of **Flat-start** was lowest for the Bayesian methods. This is because HMMs were initialized by inappropriate initial posterior distributions using no phoneme boundaries. Although **DAEM** also used no phoneme boundaries, the marginal likelihood of **DAEM** was improved from that of **Flat-start**. This indicates the DAEM algorithm effectively solved the local maxima problem. **Mtree** obtained the highest marginal likelihood of the Bayesian methods. Moreover, **Mtree** could achieve a higher marginal likelihood than the methods using label information (**Label10** and **Label200**). This demonstrates that the method using multiple model structures could estimate more reliable posterior distributions than the conventional Bayesian methods.

Figure 5 shows the phoneme accuracy for each method. The temperature schedules were adjusted to obtain the best phoneme accuracy (**DAEM(ML)**: $\alpha = 0$, **Mtree(ML)**: $\alpha = 1$, **DAEM(Bayes)**: $\alpha = 0$, **Mtree(Bayes)**: $\alpha = 0$). Comparing the ML-based methods with the Bayesian methods, all Bayesian methods were more accurate than those that were ML-based. This confirmed the effectiveness of the Bayesian approach for speech recognition. Similar to the comparison of marginal likelihoods, **Mtree** achieved the highest accuracy of methods using no phoneme boundaries (**Flat-start**, **DAEM** and **Mtree**) in both criteria. Moreover, the improvement for **Mtree** was higher than that for **DAEM** by comparing the improvements from the ML criterion to the Bayesian criterion between **DAEM** and **Mtree** methods. This means that consistently optimizing the model parameters and model structures based on the Bayesian criterion effectively improved recognition. While **Mtree(Bayes)** yielded higher accuracy than **Label10(Bayes)**, **Mtree(Bayes)** could not achieve the accuracy of **Label200(Bayes)**. Since **Label200** obtained higher accuracy than **Label10** in both criteria, **Mtree(Bayes)** might be able to obtain higher accuracy when we adjust the number of iterations or the schedule for temperature updates.

The posterior probabilities of the model structures in **Mtree(ML)** were in proportion to the likelihoods obtained by the ML estimates in all model structures. Since a larger model structure obtained a higher likelihood in the ML criterion, the largest model structure was always selected. However, this was inappropriate in most cases due to the overfitting problem. A heuristic approach to control the posterior probabilities of model structures is required to avoid this problem. However, when the number of model structures increases, it is difficult to use such heuristics to obtain an appropriate posterior distribution. In contrast, **Mtree(Bayes)** could automatically estimate accurate posterior distributions of model structures. Figure 6 plots the posterior distribution of model structures with all temperature schedules during the training process. It can be seen that the posterior probability of the larger model structure (CV-Bayes) gradually increased begin dependent on the temperature parameter to estimate the posterior distributions of the model parameters and state sequences in the early stages. Since the posterior distribution of the model structures was automatically estimated based on the Bayesian criterion, we could easily increase the number of model structures without heuristics, and we intend to investigate the effectiveness of using more than two model structures in future work.

## 5. Conclusions

This paper proposed a Bayesian framework using multiple model structures for speech recognition. For integrating the multiple model structures, the proposed method treated not only the state sequences and the model parameters but also the model structures as latent variables. Furthermore, for estimating the appropriate acoustic models, the DAEM algorithm was applied to the proposed framework. The speech recognition experiment showed the optimal posterior distributions of the model structures can be estimated automatically and a higher performance can be obtained.

As future work, we will investigate the effect of increasing the number of model structures and consider the optimization of the annealing schedules. We will also perform the word recognition experiments and using Gaussian mixture models.

## Acknowledgements

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) of the Ministry of Internal Affairs and Communication, Japan. Part of this research was supported by the Japan Society for the Promotion of Science (JSPS) Research Fellowships for Young Scientists (23-5301).

### References

[1] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. UAI 15, 1999.

[2] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to variational methods for graphical models," Mach. Learn., vol.37, no.2, pp.183–233, 2005.

[3] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. Speech Audio Process., vol.12, no.4, pp.365–381, 2004.

[4] T. Jitsuhiro and S. Nakamura, "Automatic generation of non-uniform and context-dependent HMMs based on the variational Bayesian approach," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.391–400, March 2005.

[5] K. Yu and M. Gales, "Bayesian adaptation and adaptively trained systems," Proc. Automatic Speech Recognition and Understanding Workshop (ASRU) 2005, pp.209–214, 2005.

[6] N. Ding and Z. Ou, "Variational nonparametric Bayesian hidden markov model," Proc. ICASSP 2010, pp.2098–2101, 2005.

[7] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for HMM-based speech recognition," Proc. Interspeech 2008, pp.936–939, 2008.

[8] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output error reduction (rover)," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.347–352, 1997.

[9] J. Xue and Y. Zhao, "Random forest of phonetic decision trees for acoustic modeling in conversational speech recognition," IEEE Trans. Audio Speech Language Process., vol.16, no.3, pp.519–528, 2008.

[10] S. Shiota, K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Acoustic modeling based on model structure annealing for speech recognition," Proc. Interspeech 2008, pp.932–935, 2008.

[11] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.

[12] K. Katahira, K. Watanabe, and M. Okada, "Deterministic annealing variant of variational Bayes method," J. Physics: Conference Series, vol.95, no.1, 012015, 2008.

[13] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," Neural Netw., vol.11, pp.271–282, 1998.

[14] S. Shiota, K. Hashimoto, Y. Nankaku, A. Lee, and K. Tokuda, "Deterministic annealing based training algorithm for Bayesian speech recognition," Proc. Interspeech 2009, pp.680–683, 2009.

[15] N. Ueda and T. Yamada, "Nonparametric Bayes," J. Japanese Applied Mathematics, vol.17, no.3, pp.196–214, 2007.

[16] S. Kullback and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol.22, pp.79–86, 1951.

[17] D. Povey and P.C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," Proc. ICASSP 2002, vol.1, pp.13–17, 2002.

[18] E. McDermott, T. Hazen, J.L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," IEEE Trans. Audio Speech Language Process., vol.15, no.1, pp.203–223, 2007.

[19] Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Deterministic annealing EM algorithm in parameter estimation for acoustic model," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.425–431, March 2005.

[20] J. Lam and J. marc Delosme, An efficient simulated annealing schedule: implementation and evaluation, Ph.D thesis, 1988.

[21] M. Miki, M. Shibata, and T. Hiroyasu, "106 automatic determination of the temperature schedule in simulated annealing programming," Japan Society of Mechanical Engineers, vol.2006, no.7, pp.27–32, Dec. 2006.

## Appendix: The concrete forms of the posterior distributions and the normalization terms

### A.1 The Concrete Form of $\tilde{Q}$

$$\tilde{Q}(\boldsymbol{\pi}) = C_{\boldsymbol{\pi}} \exp\Big\{\sum_{i=1}^{N}\Big\langle Z_1^i\Big\rangle \log \pi_i\Big\} \qquad \text{(A· 1)}$$

$$\tilde{Q}(\boldsymbol{\Lambda}^{(a)}) = \tilde{Q}(\boldsymbol{\alpha}_i)$$

$$= C_{\alpha_i} P^{\beta}(\alpha_i) \exp\Big\{\sum_{j=1}^{N}\sum_{t=1}^{T-1}\Big\langle Z_t^i Z_{t+1}^j\Big\rangle \log \alpha_{ij}^{\beta}\Big\} \qquad \text{(A· 2)}$$

$$\tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) = \tilde{Q}(\boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}) =$$

$$C_{\boldsymbol{\mu}_{im},\boldsymbol{S}_{im}} P^{\beta}(\boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}) \exp\Big\{\sum_{t=1}^{T}\Big\langle Z_t^i\Big\rangle \log \mathcal{N}^{\beta}(o_t \mid \boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}^{-1})\Big\} \qquad \text{(A· 3)}$$

$$\begin{aligned}\tilde{Q}(\boldsymbol{Z}) = C_Z &\prod_{i=1}^{N} \exp\left\{z_1^i \langle \log \pi_i \rangle_{\tilde{Q}(\boldsymbol{\pi})}\rangle\right\} \\ &\times \prod_{t=1}^{T-1}\prod_{i=1}^{N}\prod_{j=1}^{N} \exp\left\{z_t^i z_{t+1}^j \langle \log a_{ij}\rangle_{\tilde{Q}(\boldsymbol{a}_i)}\right\} \\ &\times \prod_{t=1}^{T}\prod_{i=1}^{N} \exp\left\{z_t^i \langle \log \mathcal{N}(o_t \mid \boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}^{-1})\rangle_{\tilde{Q}(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im}^{-1})}\right\}\end{aligned} \tag{A$\cdot$4}$$

$$\begin{aligned}\tilde{Q}(m) = C_m P^{\beta}(m) &\\ \times \exp\Big\{&\sum_{\boldsymbol{Z}} \int \tilde{Q}(\boldsymbol{Z})\tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) \\ &\times \log P^{\beta}(\boldsymbol{O} \mid \boldsymbol{Z}, m, \boldsymbol{\Lambda}_m^{(b)}) d\boldsymbol{\Lambda}_m^{(b)} \\ &+ \int \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) \log P^{\beta}(\boldsymbol{\Lambda}_m^{(b)} \mid m) d\boldsymbol{\Lambda}_m^{(b)} \\ &- \int \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) \log \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) d\boldsymbol{\Lambda}_m^{(b)}\Big\}.\end{aligned} \tag{A$\cdot$5}$$

$$\begin{aligned}&\int \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) \log P^{\beta}(\boldsymbol{\Lambda}_m^{(b)} \mid m) d\boldsymbol{\Lambda}_m^{(b)} \\ &= \beta\Big\langle \log \mathcal{N}(\boldsymbol{\mu}_{im} \mid \boldsymbol{\nu}_{im}, (\xi_{im}\boldsymbol{S}_{im})^{-1})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &\quad + \beta\Big\langle \log \mathcal{W}(\boldsymbol{S}_{im} \mid \eta_{im}, \boldsymbol{B}_{im})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})}\end{aligned} \tag{A$\cdot$6}$$

$$\begin{aligned}&\Big\langle \log \mathcal{N}(\boldsymbol{\mu}_{im} \mid \boldsymbol{\nu}_{im}, (\xi_{im}\boldsymbol{S}_{im})^{-1})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &\quad + \Big\langle \log \mathcal{W}(\boldsymbol{S}_{im} \mid \eta_{im}, \boldsymbol{B}_{im})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &= \frac{d}{2}\log|\xi_{im}| - \frac{d^2+d}{2}\log 2 - \frac{d^2+d}{4}\log\pi \\ &\quad + \frac{\eta_{im}-d}{2}\log|\bar{\boldsymbol{B}}_{im}| \\ &\quad + \frac{\eta_{im}}{2}\log|\boldsymbol{B}_{im}| - \sum_{j=1}^{d}\log\Gamma\left(\frac{\eta_{im}+1-j}{2}\right) \\ &\quad - \frac{1}{2}Tr\Big(\bar{\eta}_{im}\bar{\boldsymbol{B}}_{im}^{-1}\xi_{im}(\bar{\boldsymbol{\nu}}_{im}-\boldsymbol{\nu}_{im})(\bar{\boldsymbol{\nu}}_{im}-\boldsymbol{\nu}_{im})^{\mathrm{T}} + \xi_{im}\bar{\xi}_{im}^{-1}I\Big) \\ &\quad + \frac{1}{2}(\eta_{im}-d)\sum_{j=1}^{d}\Psi\left(\frac{\bar{\eta}_{im}+1-j}{2}\right) - \frac{1}{2}Tr\Big(\boldsymbol{B}_{im}\bar{\eta}_{im}\bar{\boldsymbol{B}}_{im}^{-1}\Big)\end{aligned} \tag{A$\cdot$7}$$

$$\begin{aligned}&\int \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) \log \tilde{Q}(\boldsymbol{\Lambda}_m^{(b)} \mid m) d\boldsymbol{\Lambda}_m^{(b)} \\ &= \Big\langle \log \mathcal{N}(\boldsymbol{\mu}_{im} \mid \bar{\boldsymbol{\nu}}_{im}, (\bar{\xi}_{im}\boldsymbol{S}_{im})^{-1})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &\quad + \Big\langle \log \mathcal{W}(\boldsymbol{S}_{im} \mid \bar{\eta}_{im}, \bar{\boldsymbol{B}}_{im})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})}\end{aligned} \tag{A$\cdot$8}$$

$$\begin{aligned}&\Big\langle \log \mathcal{N}(\boldsymbol{\mu}_{im} \mid \bar{\boldsymbol{\nu}}_{im}, (\bar{\xi}_{im}\boldsymbol{S}_{im})^{-1})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &\quad + \Big\langle \log \mathcal{W}(\boldsymbol{S}_{im} \mid \bar{\eta}_{im}, \bar{\boldsymbol{B}}_{im})\Big\rangle_{Q(\boldsymbol{\mu}_{im},\boldsymbol{S}_{im})} \\ &= \frac{d}{2}\log|\bar{\xi}_{im}| - \frac{d^2+d}{2}\log 2 - \frac{d^2+d}{4}\log\pi + \frac{d}{2}\log|\bar{\boldsymbol{B}}_{im}| \\ &\quad - \sum_{j=1}^{d}\log\Gamma\left(\frac{\bar{\eta}_{im}+1-j}{2}\right) \\ &\quad + \frac{1}{2}(\bar{\eta}_{im}-d)\sum_{j=1}^{d}\Psi\left(\frac{\bar{\eta}_{im}+1-j}{2}\right) \\ &\quad - \frac{1}{2} - \frac{1}{2}Tr\Big(\bar{\eta}_{im}I\Big)\end{aligned} \tag{A$\cdot$9}$$

### A.2 Prior Distribution

In the Bayesian approach, a conjugate prior distribution is widely used as a prior distribu- tion. Prior distributions are respectively represented as follows.

$$P(\pi) = \mathcal{D}(\{\pi_i\}_{i=1}^{N} \mid \{\phi_i\}_{i=1}^{N}), \tag{A$\cdot$10}$$

$$P(\alpha_i) = \mathcal{D}(\{a_{ij}\}_{j=1}^{N} \mid \{\alpha_{ij}\}_{j=1}^{N}), \tag{A$\cdot$11}$$

$$P(\boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}) = \mathcal{N}(\boldsymbol{\mu}_{im} \mid \nu_i(\xi_i\boldsymbol{S}_{im})^{-1})\mathcal{W}(\boldsymbol{S}_{im} \mid \eta_i, \boldsymbol{B}_i), \tag{A$\cdot$12}$$

$$P(m) = \frac{1}{m} \tag{A$\cdot$13}$$

where $\mathcal{D}(\dot{})$ is a Dirichlet distribution, and $\mathcal{N}(\dot{})\mathcal{W}(\dot{})$ is a Gauss-Wishart distribution. $\{\phi_i, \alpha_{ij}, \xi_i, \eta_i, \nu_i, \boldsymbol{B}_i\}_{i,j=1}^{N}$ is a set of hyper-parameters.

### A.3 Update of Posterior Distribution

The posterior distribution of model parameters $\tilde{Q}(\boldsymbol{\Lambda})$ can be updated by sufficient statis- tics of the training data as follows.

$$\bar{\phi}_i = \phi_i + \Big\langle Z_1^i \Big\rangle \tag{A$\cdot$14}$$

$$\bar{\alpha}_{ij} = \alpha_{ij} + \bar{T}_{ij} \tag{A$\cdot$15}$$

$$\bar{\xi}_{im} = \xi_{im} + \bar{T}_i \tag{A$\cdot$16}$$

$$\bar{\eta}_{im} = \eta_{im} + \bar{T}_i \tag{A$\cdot$17}$$

$$\bar{\nu}_{im} = \frac{\bar{T}_i\bar{o}_i + \xi_{im}\nu_{im}}{\bar{T}_i + \xi_{im}} \tag{A$\cdot$18}$$

$$\bar{\boldsymbol{B}}_{im} = \bar{T}_i\bar{C}_i + \boldsymbol{B}_{im} + \frac{\bar{T}_i\xi_{im}}{\bar{T}_i + \xi_{im}}(\bar{o}_i - \nu_{im})(\bar{o}_i - \nu_{im})^T, \tag{A$\cdot$19}$$

where the sufficient statistics $\bar{T}_i, \bar{T}_{ij}, \bar{o}_i$ and $\bar{C}_i$ are represented as follows:

$$\bar{T}_i = \sum_{t=1}^{T}\Big\langle Z_t^i \Big\rangle \tag{A$\cdot$20}$$

$$\bar{T}_{ij} = \sum_{t=1}^{T-1}\Big\langle Z_t^i Z_{t+1}^j \Big\rangle \tag{A$\cdot$21}$$

$$\bar{o}_i = \frac{1}{\bar{T}_i}\sum_{t=1}^{T}\Big\langle Z_t^i \Big\rangle o_t \tag{A$\cdot$22}$$

$$\bar{C}_i = \frac{1}{\bar{T}_i}\sum_{t=1}^{T}\Big\langle Z_t^i \Big\rangle (o_t - \bar{o}_i)(o_t - \bar{o}_i)^T \tag{A$\cdot$23}$$

### A.4 Normalization Terms

$$C_{\mathbf{\Lambda}} = C_{\pi} \prod_{i=1}^{N} C_{a_i} \prod_{i=1}^{N} C_{\boldsymbol{\mu}_{im}, \boldsymbol{S}_{im}} \tag{A$\cdot$24}$$

$$C_{\pi} = \frac{\bar{C}_{\mathcal{D}_\pi}}{C_{\mathcal{D}_\pi}}, \bar{C}_{\mathcal{D}_\pi} = \frac{\Gamma^{\beta}(\sum_{i=1}^{N} \bar{\phi}_i)}{\prod_{i=1}^{N} \Gamma^{\beta}(\bar{\phi}_i)} \tag{A$\cdot$25}$$

$$C_{\alpha_i} = \frac{\bar{C}_{\mathcal{D}_i}}{C_{\mathcal{D}_i}}, \bar{C}_{\mathcal{D}_i} = \frac{\Gamma^{\beta}(\sum_{j=1}^{N} \bar{\alpha}_{ij})}{\prod_{j=1}^{N} \Gamma^{\beta}(\bar{\alpha}_{ij})} \tag{A$\cdot$26}$$

$$C_{\mu_i, S_i} = \frac{\bar{C}_{\mathcal{N}} \bar{C}_{\mathcal{W}_i}}{C_{\mathcal{N}} C_{\mathcal{W}_i}} (2\pi)^{\frac{\beta \bar{N} D}{2}} \tag{A$\cdot$27}$$

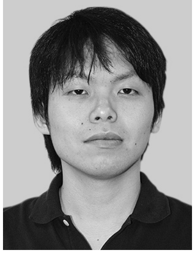
$$C_{\mathcal{N}_i} = (2\pi)^{-\frac{\beta D}{2}} \xi_{im}^{\frac{\beta D}{2}} \tag{A$\cdot$28}$$

$$C_{\mathcal{W}_i} = \frac{|B_{im}|^{\frac{\eta_{im}}{2}}}{2^{\frac{\beta \eta_{im} D}{2}} \pi^{\frac{\beta D(D-1)}{4}} \prod_{j=1}^{D} \Gamma^{\beta}(\frac{\eta_{im}+1-j}{2})} \tag{A$\cdot$29}$$

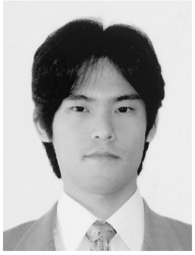
The normalization term $C_m$ can be defined by using $\sum_m \tilde{Q}(m) = 1$,

$$C_m = \frac{1}{\sum_m \tilde{Q}(m)} \tag{A$\cdot$30}$$

**Sayaka Shiota** 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 2007, and 2009, respectively. She is currently a Doctor's candidate at Nagoya Institute of Technology. From October 2009 to January 2010, she was an intern researcher at National Institute of Information and Communications Technology (NICT), Kyoto, Japan. Her research interests include statistical speech recognition and synthesis. She is a student member of the Acoustical Society of Japan and ISCA.



**Kei Hashimoto** received 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 is a Research Fellow of the Japan Society for the Promotion of Science (JSPS) in the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan. From May 2010 to September 2010, he was a visiting researcher in University of Edinburgh and Cambridge University. 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 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 speech coding, speech synthesis and recognition, and statistical machine learning.