A MODEL STRUCTURE INTEGRATION BASED ON A BAYESIAN FRAMEWORK FOR SPEECH RECOGNITION

Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan

ABSTRACT

This paper proposes an acoustic modeling technique based on Bayesian framework using multiple model structures for speech recognition. The Bayesian approach is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters, and its effectiveness in HMM-based speech recognition has been reported. Although the basic idea underlying the Bayesian approach is to treat all parameters as random variables, only one model structure is still selected in the conventional method. Multiple model structures are treated as latent variables in the proposed method and integrated based on the Bayesian framework. Furthermore, we applied deterministic annealing to the training algorithm to estimate appropriate acoustic models. 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.

Index Terms— Speech recognition, Hidden Markov model, Bayesian methods, Deterministic annealing

1. INTRODUCTION

The maximum likelihood (ML) criterion has been used for training HMMs in conventional hidden Markov model (HMM)-based speech recognition systems. However, since the ML criterion produces point estimates of model parameters, the accuracy of estimation may decrease when only a small number of training data is available. The Bayesian approach is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters, and it can be used to accurately estimate observation distributions even if there are few training data. However, calculations become complicated due to a combination of latent variables (i.e., state sequences and model parameters). The variational Bayesian (VB) method has been proposed as an effective method of approximation for the Bayesian approach [1] to solve this problem, and it performed well in HMM-based speech recognition [2].

There have been many efforts in the conventional speech recognition based on generative models 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., random forest [3], ROVER [4], and model structure annealing [5], 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 VB method.

The conventional VB method sometimes suffers from the local maxima problem because of the combination of latent variables. To overcome this problem, we have proposed a training algorithm applying the deterministic annealing EM (DAEM) algorithm [6] to Bayesian speech recognition, and reported its effectiveness in the local maxima problem [7]. Since the proposed method uses the multiple model structures, the model structures are additionally treated as a latent variable in the VB method. This means that the proposed framework might cause a serious local maxima problem. Therefore, to improve the optimization algorithm, the 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.

The rest of this paper is organized as follows. Section 2 describes speech recognition based on the variational Bayesian method. Bayesian speech recognition using multiple model structures obtained by applying the DAEM algorithm is described in Section 3. Section 4 presents results obtained from continuous phoneme recognition experiments, and the final section draws conclusions and introduces future work.

2. SPEECH RECOGNITION BASED ON VARIATIONAL BAYESIAN METHOD

The Bayesian approach assumes that model parameter $\mathbf{\Lambda}$ is a random variable, while the ML approach estimates a constant model parameter. Let $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ be a set of training data and T denotes the number of frames; the log marginal likelihood is represented by:

$$\log P(\boldsymbol{O}) = \log \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda})\, d\mathbf{\Lambda}, \qquad (1)$$

where $\boldsymbol{Z} = (z_1, z_2, \ldots, z_T)$ is a sequence of HMM states. Since the model parameters are marginalized out, the effect of over-fitting is mitigated. However, it is difficult to solve integral and expectation calculations. Calculation becomes much more complicated, especially when the model includes latent variables. To overcome this problem, the variational Bayesian (VB) method has been proposed as a tractable method of approximation in the Bayesian approach [1]. The lower bound of the log marginal likelihood is maximized in the VB method instead of the true likelihood. The lower bound of the

log marginal likelihood is defined by using Jensen's inequality:

$$\log P(\boldsymbol{O}) = \log \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})} \, d\boldsymbol{\Lambda}$$
$$\geq \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})} \, d\boldsymbol{\Lambda}, \quad (2)$$

where $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ is an approximate distribution of posterior distribution $P(\boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{O})$. Optimal posterior distribution $Q(\boldsymbol{Z}, \boldsymbol{\Lambda})$ is estimated by maximizing the lower bound. However, the calculation becomes complicated because of the combination of latent variables. Thus, the random variables in the variational method are assumed to be conditionally independent of one another, i.e., $Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) = Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda})$. Under this assumption, the optimal VB posterior distributions that maximize the lower bound are obtained. Since the VB posterior distributions $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ that are obtained are dependent on each other, these updates should be iterated. It has been reported that speech recognition based on the VB method outperformed the ML approach in speech recognition [2].

3. A MODEL STRUCTURE INTEGRATION BASED ON A BAYESIAN FRAMEWORK

Some approaches using multiple model structures have recently been proposed to increase model complexity (e.g., random forest [3], ROVER [4], and model structure annealing [5]). Although various integration techniques and criteria can be considered, this paper focuses on a model structure integration based on the Bayesian framework.

3.1. Statistical model including multiple model structures

We define a marginal likelihood function treating model structures as latent variables to consider the framework using multiple model structures in Bayesian speech recognition.

$$P(\boldsymbol{O}) = \sum_{m} \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z}, m, \boldsymbol{\Lambda}_m) d\boldsymbol{\Lambda}_m, \quad (3)$$

$$P(\boldsymbol{O}, \boldsymbol{Z}, m, \boldsymbol{\Lambda}_m) = P(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m) P(\boldsymbol{\Lambda}_m \mid m) P(m), \quad (4)$$

where $m \in \{1, \ldots, M\}$ indexes the model structures, M is the number of the model structures, and $\boldsymbol{\Lambda}_m \in \{\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_M\}$ denotes a set of model parameters for the m-th model structure. Prior distribution $P(\boldsymbol{\Lambda}_m \mid m)$ is prepared for each model structure m. Note that this paper assumes structures of a phonetic decision tree is treated as the model structure. In Eq. (4), the state sequence $\boldsymbol{Z}$ is not dependent on 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. Although the proposed model can be trained in the same manner as the variational Bayesian method, it has been confirmed [7] that even conventional Bayesian speech recognition using a single model structure suffers from the local maxima problem. Since the proposed method not only treats state sequences and model parameters but also model structures as latent variables, the local maxima problem is more serious than conventional Bayesian speech recognition. Deterministic annealing was adopted in the proposed framework to overcome this problem.

3.2. Training algorithm based on deterministic annealing

The problem of maximizing the log likelihood function is reformulated in the DAEM algorithm [6] as the problem of minimizing a free energy function. To adopt deterministic annealing for the proposed method, we redefine the free energy function based on the marginal likelihood function in Eq. (3) as:

$$\bar{\mathcal{F}}_\beta = -\frac{1}{\beta} \sum_{m} \sum_{\boldsymbol{Z}} \int \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m)$$
$$\times P^\beta(\boldsymbol{\Lambda}_m \mid m) P^\beta(m) d\boldsymbol{\Lambda}_m, \quad (5)$$

where β is called a temperature parameter. The upper bound of the free energy function is defined by using Jensen's inequality:

$$\bar{\mathcal{F}}_\beta \leq -\frac{1}{\beta} \sum_{m} \sum_{\boldsymbol{Z}} \int \tilde{Q}(\boldsymbol{Z}, m, \boldsymbol{\Lambda}_m)$$
$$\times \log \frac{P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m) P^\beta(\boldsymbol{\Lambda}_m \mid m) P^\beta(m)}{\tilde{Q}(\boldsymbol{Z}, m, \boldsymbol{\Lambda}_m)} d\boldsymbol{\Lambda}_m. \quad (6)$$

Since approximate distribution $\tilde{Q}(\boldsymbol{Z}, m, \boldsymbol{\Lambda}_m)$ is a joint distribution of the three latent variables, calculating the upper bound becomes more complicated than that with the conventional VB method using only one model structure. To obtain the minimum upper bound, we assume the constraint: $\tilde{Q}(\boldsymbol{Z}, m, \boldsymbol{\Lambda}_m) = \tilde{Q}(\boldsymbol{Z})\tilde{Q}(m)\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$. Note that the dependence between model parameters and model structures remains as a prior distribution in Eq. (4). Under this constraint, optimal posterior distributions $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ are obtained as:

$$\tilde{Q}(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp\Big\langle \Big\langle \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m) \Big\rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)} \Big\rangle_{\tilde{Q}(m)}, \quad (7)$$

$$\tilde{Q}(m) = C_m P^\beta(m) \exp\Big\langle \Big\langle \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m) \Big\rangle_{\tilde{Q}(\boldsymbol{Z})}$$
$$+ \log \frac{P^\beta(\boldsymbol{\Lambda}_m \mid m)}{\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)} \Big\rangle_{\tilde{Q}(\boldsymbol{\Lambda}_m)}, \quad (8)$$

$$\tilde{Q}(\boldsymbol{\Lambda}_m \mid m) = C_{\boldsymbol{\Lambda}_m} P^\beta(\boldsymbol{\Lambda}_m \mid m)$$
$$\times \exp\Big\langle \log P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid m, \boldsymbol{\Lambda}_m) \Big\rangle_{\tilde{Q}(\boldsymbol{Z})}, \quad (9)$$

where $C_{\boldsymbol{Z}}$, C_m and $C_{\boldsymbol{\Lambda}_m}$ correspond to the normalization terms of $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ and $\langle \cdot \rangle_Q$ denotes the expectation with respect to Q. Since optimal variational posterior distributions $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ depend on one another, from Eqs. (7), (8), and (9), these distributions should be iteratively updated. Although the number of the model structures are considered infinity theoretically, the finite number is set to M practically. In the proposed framework, after the multiple model structures are prepared, the posterior distribution of the model structure is estimated auomatically. Since the proposed framework adopts the deterministic annealing process for training, the temperature parameter β is gradually increasing from 0 to 1. At the initial temperature ($\beta \simeq 0$), the variational posterior distributions $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ take a form that has a nearly uniform distribution. This means that all model structures are uniformly used for estimating the posterior distribution of the model parameters and the state sequences in the initial step. While the temperature is decreasing ($\beta \to 1$), the form of $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ change to each original posterior distribution. The factorized posterior distributions at this stage gradually interact with one another while taking into account the reliability of their estimates, and this process leads to a good solution as a joint posterior distribution. The $\tilde{Q}(\boldsymbol{Z})$, $\tilde{Q}(m)$, and $\tilde{Q}(\boldsymbol{\Lambda}_m \mid m)$ at the final temperature ($\beta = 1$) take each original posterior distribution. Through this process, the optimal posterior probability of each model structure can be automatically estimated.

3.3. Related topics

We reported a method of approximating the joint optimization of state sequences and model structures based on ML-based speech recognition [5]. There was a problem in the ML-based framework in that an accurate posterior probability for the model structures could not be automatically estimated. This is because the ML criterion selected the largest model structure, and this was inappropriate due to the over-fitting problem. The proposed method, on the other hand, can be used to automatically estimate an adequate posterior distribution for the model structures because the Bayesian criterion can be used for model selection.

The random forest (RF) [3] is one technique that uses multiple model structures. However, there are some differences between RF and the proposed method. One difference is how the model structures are constructed. The RF method changes the data set or question set used for constructing the model structures. Although the proposed approach can also use these methods, we used the Bayesian framework to construct adequate model structures. Another difference is how multiple model structures can be used. Several ways of combining models have been tried in the RF method because there are no criteria for estimating combined weights. The proposed method can be used to automatically estimate the posterior probabilities of model structures based on the consistent Bayesian criterion.

4. EXPERIMENTS

4.1. Experimental condition

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 [5]. 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 [8], [7]. Prior distributions and model selection of the Bayesian methods are automatically optimized by using the cross validation. Two tree structures were used for the approaches utilizing a single model structure (**Flat-start**, **DAEM**, **Label10**, and **Label200**).

Table 1. Upper bound of log marginal likelihood $\bar{\mathcal{F}}_\beta$

without phoneme boundaries			with phoneme boundaries	
Flat-start	**DAEM**	**Mtree**	**Label10**	**Label200**
-77.39	-77.19	-76.39	-77.24	-77.07

- 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 [9]. This structure had 18,099 leaf nodes (CV-Bayes).

We also prepared a model structure representing monophone models for **Mtree(ML)** and **Mtree(Bayes)**. The monophone structure had 129 leaf nodes. The number of temperature parameter updates in the DAEM algorithm was set to 20 ($I = 20$), and EM-steps were iterated 10 times at each temperature. Temperature parameter β was updated by using $\beta(i) = (i/I)^n,\ i = 0, ..., I$, 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: $(Q_{ML}(Monophone) = 0.5(1 - i/I), Q_{ML}(MDL) = 0.5(1 + i/I))$. Note that **Mtree(Bayes)** does not require predetermined posterior probabilities of the model structures.

4.2. Experimental results

Table 1 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 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 1 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

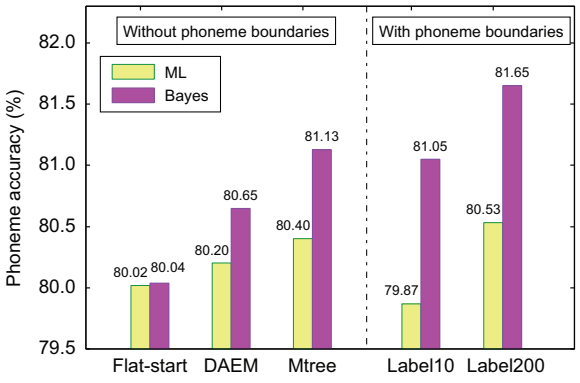

Fig. 1. *Phoneme accuracy*

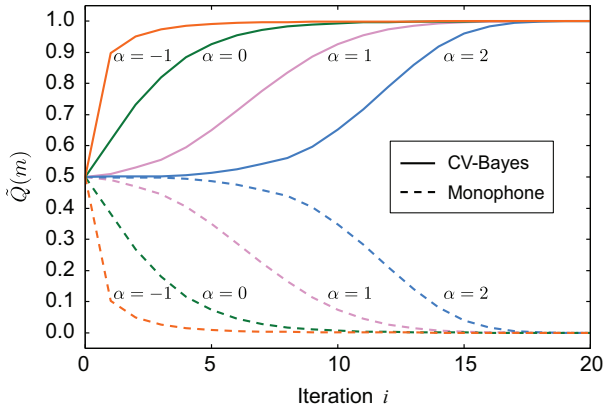

Fig. 2. *Posterior distributions of model structures.*

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 over-fitting 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 2 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. CONCLUSION

This paper proposed integrating model structures based on the Bayesian framework for speech recognition. The proposed method not only treated state sequences and model parameters but also model structures as latent variables. Furthermore, deterministic annealing was applied to the proposed framework for relaxing the local maxima problem. The speech recognition experiment demonstrated the proposed method could automatically estimate reliable posterior distributions of model parameters and an adequate posterior distribution of model structures. We intend to investigate what effect increasing the number of model structures will have in future work and consider optimizing the training process.

6. ACKNOWLEDGEMENTS

The research leading to these results was partly funded by 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).

7. REFERENCES

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

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

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

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

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

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

[7] S. Shiota, K. Hashimoto, Y. Nankaku, A. Lee, and K. Tokuda, "Deterministic annealing based training algorithm for bayesian speech recognition," *Proceedings of Interspeech 2009*, pp. 680–683, 2009.

[8] 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, 2005.

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