Deterministic Annealing Based Training Algorithm for Bayesian Speech Recognition

Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

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

Abstract

This paper proposes a deterministic annealing based training algorithm for Bayesian speech recognition. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters. However, the local maxima problem in the Bayesian method is more serious than in the ML-based approach, because the Bayesian method treats not only state sequences but also model parameters as latent variables. The deterministic annealing EM (DAEM) algorithm has been proposed to improve the local maxima problem in the EM algorithm, and its effectiveness has been reported in HMM-based speech recognition using ML criterion. In this paper, the DAEM algorithm is applied to Bayesian speech recognition to relax the local maxima problem. Speech recognition experiments show that the proposed method achieved a higher performance than the conventional methods.

Index Terms: variational Bayesian method, cross validation, speech recognition, deterministic annealing

1. Introduction

In HMM-based speech recognition, the expectation maximization (EM) algorithm is widely used for parameter estimation. The EM algorithm provides a simple iterative procedure to obtain approximate maximum likelihood (ML) estimates. However, it sometimes suffers from the local maxima problem. To relax this problem, the deterministic annealing EM (DAEM) algorithm has been proposed [1]. In the DAEM algorithm, the problem of maximizing the log-likelihood is reformulated as minimizing the thermodynamic free energy. It's posterior distribution derived includes a "temperature" parameter which controls the influence of unreliable model parameters. It has been reported that the DAEM algorithm is effective for HMM-based speech recognition using the ML criterion [2].

The ML criterion has been usually used for training HMMs. However, the ML criterion produces a point estimate of HMM parameters and the estimation accuracy may be degraded when little training data is available. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters, and it can accurately estimate observation distributions even though 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 [3].

The Bayesian approach uses prior information which is represented by the prior distributions. Since prior distributions affect the estimation of posterior distributions and model selection, the determination of prior distributions is an important problem for estimating of appropriate acoustic models. To overcome this problem, the prior distribution determination technique using cross validation has been proposed [4]. By using cross valid prior distributions, an appropriate model structure can be selected in the context clustering without tuning parameters. In this papaer, we use the prior determination technique based on the cross validation as a baseline system of the Bayesian approach.

Although the Bayesian approach achieves higher performance than the ML approach, the local maxima problem in the Bayesian method is more serious than in the ML-based approach, because the Bayesian method treats not only state sequences but also model parameters as latent variables. The combination of many latent variables makes the likelihood function complicated. Therefore, the optimization algorithm is important for the Bayesian approach. Furthermore, the VB method assumes the independence between the posterior distributions of state sequences and model parameters, and these factorized distributions are iteratively updated. This means that the VB method requires reliable initial posterior distributions. To overcome this problem, we applied the DAEM algorithm to the VB method to improve the performance of the VB speech recognition. The proposed method provides a theoretically well defined algorithm, because the update equations of the posterior distributions are straightforwardly derived from the DAEM free energy function using the integrated manner based on the variational approximation.

The rest of this paper is organized as follows. Section 2 describes the deterministic annealing EM algorithm method, and Section 3 describes the speech recognition based on variational Bayesian. Section 4 describes the DAEM algorithm for the Bayesian speech recognition. In Section 5, results of the continuous phoneme recognition experiments are presented, and concluding remarks and future work are presented in the final section.

2. Deterministic annealing EM algorithm

The objective of the EM algorithm is to estimate a set of model parameters which maximizes the incomplete log-likelihood function:

$$\mathcal{L}(\boldsymbol{O}) = \log \sum_{\boldsymbol{Z}} P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}). \tag{1}$$

where $\boldsymbol{\Lambda}$ denotes a set of model parameters and $\boldsymbol{O} = (\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_T)$ and $\boldsymbol{Z} = (z_1, z_2, ..., z_T)$ are the observation and state sequences, respectively. The EM algorithm iteratively maximizes the auxiliary function so called $\mathcal{Q}$-function:

$$\mathcal{Q}(\mathbf{\Lambda}, \mathbf{\Lambda'}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z} \mid \boldsymbol{O}, \mathbf{\Lambda}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda'}), \quad (2)$$

where $P(\boldsymbol{Z} \mid \boldsymbol{O}, \mathbf{\Lambda})$ is the posterior probability of $\boldsymbol{Z}$. It can be obtained by the Bayes rule as follows:

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

In the DAEM algorithm [1], the problem of maximizing the log-likelihood function is reformulated as the problem of minimizing the following free energy function:

$$\begin{aligned} \mathcal{F}_\beta(\mathbf{\Lambda}) &= -\frac{1}{\beta} \log \sum_{\boldsymbol{Z}} P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) \\ &= -\sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{O}, \mathbf{\Lambda}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) \\ &\qquad - \frac{1}{\beta} I[f(\boldsymbol{Z} \mid \boldsymbol{O}, \mathbf{\Lambda})], \end{aligned} \quad (4)$$

where $I[x]$ denotes the entropy of x and $1/\beta$ is called as "temperature." If $\beta = 1$, the negative free energy $-\mathcal{F}_\beta(\mathbf{\Lambda})$ becomes equal to the log-likelihood function $\mathcal{L}(\boldsymbol{O})$. In the deterministic annealing approach, the new posterior distribution f is derived so as to minimize the free energy under the constraint of $\sum_{\boldsymbol{Z}} f = 1$. To solve this problem, we can use elementary calculus of variations to take functional derivatives of Eq. (4) with respect to f, and the optimal distribution can be derived as

$$f(\boldsymbol{Z} \mid \boldsymbol{O}, \mathbf{\Lambda}) = \frac{P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda})}{\sum_{\boldsymbol{Z}} P^\beta(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda})}. \quad (5)$$

In the DAEM algorithm, the temperature parameter β is gradually increased while iterating the EM-steps at each temperature. This process is called "annealing." When $1/\beta$ is set to an initial temperature $\beta^{(0)} \simeq 0$, the EM-steps may achieve a single global minimum of $\mathcal{F}_\beta(\mathbf{\Lambda})$. At the initial temperature, the posterior distribution f takes a form nearly uniform distribution. While the temperature is decreasing, the form of f changes from uniform to the original posterior. Finally at the temperature $1/\beta = 1$, the DAEM algorithm is identical with the original EM algorithm. the reliable model parameters can be estimated usin the DAEM algorithm and it has been reported that the DAEM algorithm is effective for HMM-based speech recognition using the ML criterion[2].

3. Speech recognition based on variational Bayesian method

3.1. Bayesian approach

Let $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ be a set of training data of D dimensional feature vectors, and T is used to denote the frame number. The likelihood function of an HMM is represented by:

$$P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) = \prod_{t=1}^{T} a_{z_{t-1}z_t} \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{z_t}, \boldsymbol{S}_{z_t}^{-1}), \quad (6)$$

where $\boldsymbol{Z} = (z_1, z_2, \cdots, z_T)$ is a sequence of HMM states, $z_t \in \{1, \ldots, N\}$ denotes a state at frame t and N is the number of states in an HMM. A set of model parameters $\mathbf{\Lambda} = \{a_{ij}, \boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i,j=1}^N$ consists of 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 $\mathcal{N}(\cdot \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1})$.

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

$$P(\mathbf{\Lambda} \mid \boldsymbol{O}) = \frac{P(\boldsymbol{O} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda})}{P(\boldsymbol{O})}, \quad (7)$$

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

$$P(\boldsymbol{X} \mid \boldsymbol{O}) = \int P(\boldsymbol{X} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda} \mid \boldsymbol{O}) d\mathbf{\Lambda}. \quad (8)$$

The model parameters are integrated out in Eq. (8), so that the effect of over-fitting is mitigated. However, it is difficult to solve the integral and expectation calculations. Especially, when a 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 and it showed good performance in the HMM-based speech recognition [3], [5].

3.2. Variational Bayesian method

The variational Bayesian method maximizes a lower bound of log marginal likelihood $\mathcal{F}$ instead of the true likelihood. A lower bound of log marginal likelihood is defined by using Jensen's inequality:

$$\begin{aligned} \mathcal{L}(\boldsymbol{O}) &= \log \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda})\, d\mathbf{\Lambda} \\ &= \log \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}) Q(\mathbf{\Lambda}) \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda})}{Q(\boldsymbol{Z}) Q(\mathbf{\Lambda})}\, d\mathbf{\Lambda} \\ &\geq \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}) Q(\mathbf{\Lambda}) \log \frac{P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) P(\mathbf{\Lambda})}{Q(\boldsymbol{Z}) Q(\mathbf{\Lambda})}\, d\mathbf{\Lambda} \\ &= \mathcal{F}. \end{aligned} \quad (9)$$

In the VB method, VB posterior distributions $Q(\mathbf{\Lambda})$ and $Q(\boldsymbol{Z})$ are introduced to approximate the true posterior distributions. The optimal VB posterior distributions can be obtained by maximizing the objective function $\mathcal{F}$ with the variational method as follows:

$$Q(\mathbf{\Lambda}) = C_{\mathbf{\Lambda}} P(\mathbf{\Lambda}) \exp\left\{ \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda}) \right\}, \quad (10)$$

$$Q(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp\left\{ \int Q(\mathbf{\Lambda}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \mathbf{\Lambda})\, d\mathbf{\Lambda} \right\}, \quad (11)$$

where $C_{\mathbf{\Lambda}}$ and $C_{\boldsymbol{Z}}$ are the normalization terms of $Q(\mathbf{\Lambda})$ and $Q(\boldsymbol{Z})$, respectively. Since equations (10) and (11) are depend on each other, these updates should be iterated as the EM algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence.

3.3. Bayesian context clustering using cross validation

In the Bayesian approach, prior distributions are usually determined heuristically. However, hyper-parameters (parameters of prior distributions) affect the model selection as tuning parameters. Therefore, to automatically select an apropriate model structure, a determination technique of prior distribution is required. One possible approach is to optimize the hyper-parameters using training data so as to maximize the marginal likelihood. However, it still needs tuning parameters which control influences of prior distributions, and often leads to the over-fitting problem as the ML criterion. To overcome this problem, the prior distribution determination technique using cross validation has been proposed [4]. The cross validation is known as a straightforward and useful method for model structure optimization. By using cross valid prior distributions, an appropriate model structure can be selected in the Bayesian context clustering without tuning parameters. We apply the prior determination technique based on K-fold cross validation as a baseline system of the Bayesian approach.

4. DAEM algorithm for variational Bayes method

In the VB method, the free energy function for Bayesian approach can be rewritten as follows:

$$\mathcal{F}_\beta(\boldsymbol{\Lambda}) = -\frac{1}{\beta}\log\sum_{\boldsymbol{Z}}\int P^\beta(\boldsymbol{O},\boldsymbol{Z}\mid\boldsymbol{\Lambda})P^\beta(\boldsymbol{\Lambda})\,d\boldsymbol{\Lambda}. \quad (12)$$

An upper bound of log marginal likelihood $\bar{\mathcal{F}}_\beta(\boldsymbol{\Lambda})$ is defined by using Jensen's inequality:

$$\begin{aligned}\mathcal{F}_\beta(\boldsymbol{\Lambda}) &= -\frac{1}{\beta}\log\sum_{\boldsymbol{Z}}\int \hat{Q}(\boldsymbol{Z})\hat{Q}(\boldsymbol{\Lambda})\frac{P^\beta(\boldsymbol{O},\boldsymbol{Z}\mid\boldsymbol{\Lambda})P^\beta(\boldsymbol{\Lambda})}{\hat{Q}(\boldsymbol{Z})\hat{Q}(\boldsymbol{\Lambda})}\,d\boldsymbol{\Lambda}\\ &\le -\frac{1}{\beta}\sum_{\boldsymbol{Z}}\int \hat{Q}(\boldsymbol{Z})\hat{Q}(\boldsymbol{\Lambda})\log\frac{P^\beta(\boldsymbol{O},\boldsymbol{Z}\mid\boldsymbol{\Lambda})P^\beta(\boldsymbol{\Lambda})}{\hat{Q}(\boldsymbol{Z})\hat{Q}(\boldsymbol{\Lambda})}\,d\boldsymbol{\Lambda}\\ &= \bar{\mathcal{F}}_\beta(\boldsymbol{\Lambda}) \quad (13)\end{aligned}$$

The optimal VB posterior distributions can be obtained by minimizing the objective function $\bar{\mathcal{F}}_\beta(\boldsymbol{\Lambda})$ with the variational method as follows:

$$\hat{Q}(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}}P^\beta(\boldsymbol{\Lambda})\exp\left\{\sum_{\boldsymbol{Z}}\hat{Q}(\boldsymbol{Z})\log P^\beta(\boldsymbol{O},\boldsymbol{Z}\mid\boldsymbol{\Lambda})\right\}, \quad (14)$$

$$\hat{Q}(\boldsymbol{Z}) = C_{\boldsymbol{Z}}\exp\left\{\int\hat{Q}(\boldsymbol{\Lambda})\log P^\beta(\boldsymbol{O},\boldsymbol{Z}\mid\boldsymbol{\Lambda})\,d\boldsymbol{\Lambda}\right\}. \quad (15)$$

Since equations (14) and (15) are dependent each other, these updates should be iterated in the E-step of the DAEM algorithm. At the initial temperature $\beta^{(0)} \simeq 0$, the VB posterior distributions $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ take a form nearly uniform distribution. While the temperature is decreasing, the form of $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ change from uniform to each original posterior distribution. Finally the temperature $\beta = 1$, $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ take each original posterior distribution and the reliable posterior distributions can be estimated.

5. Experiments

To evaluate the effectiveness of the proposed method, speaker independent continuous phoneme recognition experiments were conducted.

Table 1: Experimental conditions

Training data	JNAS 20,000 utterances
Test data	JNAS 100 utterances
Sampling rate	16 kHz
Feature vector	12-order MFCC + ΔMFCC + ΔEnergy
Window	Hamming
Frame size	25ms
Frame shift	10ms
Number of HMM state	3 (left-to-right)
Number of phoneme categories	43

5.1. Experimental conditions

The experimental conditions are summarized in Table 1. The training data of about 20,000 Japanese sentences and testing data of 100 sentences were prepared from Japanese Newspaper Article Sentences (JNAS). Three-state left-to-right HMMs were used to model 43 Japanese phonemes, and 144 questions were prepared for the decision tree context clustering. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix.

In these experiments, the following five algorithms were compared.

- "ML" : Acoustic models trained by ML criterion and model structures selected by MDL criterion [6] and 50 EM-steps was conducted in the EM algorithm. HMMs were initialized by the segmental k-means algorithm.
- "CV-Bayes(f-EM50)" : Acoustic models trained by the Bayesian criterion and model structures selected by the Bayesian criterion using cross validation and 50 EM-steps was conducted in the EM algorithm. The posterior distributions were initialized by the flat start training.
- "CV-Bayes(EM5)" : Acoustic models trained by the Bayesian criterion and model structures selected by the Bayesian criterion using cross validation and 5 EM-steps was conducted in the EM algorithm. The posterior distributions were initialized by the k-means algorithm.
- "CV-Bayes(EM50)" : Acoustic models trained by the Bayesian criterion and model structures selected by the Bayesian criterion using cross validation and 50 EM-steps was conducted in the EM algorithm. The posterior distributions were initialized by the k-means algorithm.
- "CV-Bayes(DAEM)" : Acoustic models trained by the Bayesian criterion and model structures selected by the Bayesian criterion using cross validation and the DAEM algorithm was used for training algorithm.

The flat start training ("CV-Bayes(f-EM50)") assumes that initial posterior distributions of state sequences are uniform distribution. Once the posterior distributions of state sequences are given, the posterior distributions of model parameters can be estimated by the statistics of state sequences. In the initialization by the k-means algorithm, the posterior distribution of state sequences were initialized by the segmental k-means algorithm using phoneme boundary labels. In the Bayesian approaches, the posterior distribution of model parameters are also updated in the segmental k-means algorithm. Although the DAEM algorithm includes the initialization process, the DAEM algorithm ("CV-Bayes(DAEM)") with $\beta = 0$ is equivalent to the initial values of the flat start training. This means that the DAEM algorithm uses no phoneme boundary labels in the initialization of posterior distributions. However, even though the flat start training updates the posterior distributions immediately at the

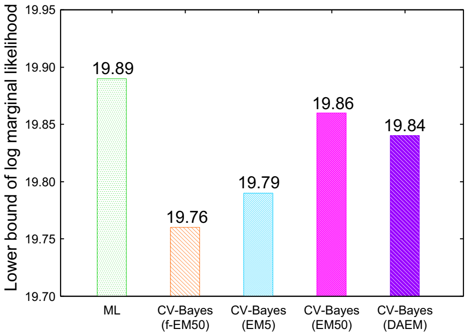

Figure 1: Log marginal likelihood

first iteration based on unreliable initial parameters (this corresponds to the DAEM with $\beta = 0$ at the 1st iteration and $\beta = 1$ at the nd iteration), the DAEM algorithm gradually increase the temperature parameter β, and updates the posterior distributions slowly based on the annealing process.

The model structure based on MDL criterion has 5400 states and based on the Bayesian approach using cross validation has 16205 states. In "CV-Bayes" methods, the cross validation uses 10 folds. The temperature parameter β for the DAEM algorithm was updated by

$$\beta(i) = \frac{i}{I}, (i = 0, \ldots, I) \tag{16}$$

where i denotes the iteration number. The number of temperature update steps was set to 10 ($I = 10$), and 5 EM-steps were conducted at each temperature, in total 50 EM-steps were conducted.

5.2. Experimental results

Figure 1 compares the lower bound of the log marginal likelihood $\mathcal{F}$ for the training data, though the value of "ML" shows the log likelihood of the ML parameters (not marginal). Since the marginal likelihood is defined as the weighted sum of the likelihood function (equation (9)), the marginal likelihoods of the Bayesian approaches were lower than the likelihood of "ML." The marginal likelihood of "CV-Bayes(f-EM50)" was the lowest among Bayesian methods. This is because of the local maxima problem caused by the inappropriate initial posterior distributions obtained without using phoneme boundary information. Although "CV-Bayes(DAEM)" also uses no phoneme boundaries, the marginal likelihood of "CV-Bayes(DAEM)" was improved than that of "CV-Bayes(f-EM50)." This result confirmed that the local maxima problem can be relaxed by the DAEM algorithm. Comparing "CV-Bayes(EM5)" with "CV-Bayes(EM50)," "CV-Bayes(EM50)" obtained the higher likelihood. This means that 5 EM-steps are not enough to converge the marginal likelihood. "CV-Bayes(DAEM)" also iterated the EM-steps 5 times at the last temperature ($\beta = 1$), and this may be the reason that the marginal likelihood of "CV-Bayes(DAEM)" was lower than that of "CV-Bayes(EM50)." However, the likelihood of "CV-Bayes(DAEM)" was higher than that of "CV-Bayes(EM5)." This means that the DAEM algorithm obtained reliable posterior distributions by using annealing process, even though no phoneme boundary information was used.

Figure 2 shows the phoneme accuracy of acoustic models. Contrary to the marginal likelihood, the Bayesian approaches outperformed "ML." This result confirmed that the Bayesian approach is useful for HMM-based speech recognition. Comparing the Bayesian approaches, "CV-Bayes(f-EM50)" was the lowest recognition performance, because of the local maxima problem. Although "CV-Bayes(EM50)" achieved the highest likelihood, "CV-Bayes(EM50)" obtained no significant improvement as compared with "CV-Bayes(EM5)" in phoneme accuracy. Comparing the EM and DAEM algorithm, "CV-Bayes(DAEM)" achieved the higher phoneme accuracy than the EM algorithm using phoneme boundary information. This result indicated that the DAEM algorithm is effective to relax the serious local maxima problem in the VB speech recognition.

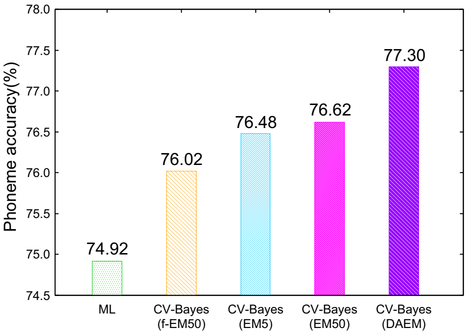

Figure 2: Phoneme accuracy

6. Conclusions

This paper proposed a deterministic annealing based training algorithm for Bayesian speech recognition. The local maxima problem in the Bayesian method is more serious than in the ML-based approach, because the Bayesian method treats not only state sequences but also model parameters as latent variables. In this paper, the DAEM algorithm was applied to the Bayesian speech recognition to improve the recognition performance. The results of speech recognition experiments showed that the proposed method achieved higher performance than the conventional methods. As future work, we will apply this proposed framework to the simultaneous optimization of state sequences and model structures [7].

7. References

[1] N. Ueda and R. Nakano, "Deterministic Annealing EM Algorithm, " Neural Networks, (11), pp.271–282, 1998.

[2] 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.

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

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

[5] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. SAP, vol.12, pp.365–381, 2004.

[6] K. Shinoda and T. Watanabe, "Acoustic Modeling Based on the MDL Criterion for speech recognition," in Proc. of Eurospeech, pp.99–102, 1997.

[7] S. Shiota, K. Hashimoto, Y. Nankaku, A. Lee, K. Tokuda, "Acoustic Modeling Based on Model Structure Annealing for Speech Recognition," in Proc. Interspeech, pp.932–935, 2008.