interspeech 2008 incorporating SST 08

Acoustic Modeling Based on Model Structure Annealing for Speech Recognition

Sayaka Shiota, Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

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

Abstract

This paper proposes an HMM training technique using multiple phonetic decision trees and evaluates it in speech recognition. In the use of context dependent models, the decision tree based context clustering is applied to find a parameter tying structure. However, the clustering is usually performed based on statistics of HMM state sequences which are obtained by unreliable models without context clustering. To avoid this problem, we optimize the decision trees and HMM state sequences simultaneously. In the proposed method, this is performed by maximum likelihood (ML) estimation of a newly defined statistical model which includes multiple decision trees as hidden variables. Applying the deterministic annealing expectation maximization (DAEM) algorithm and using multiple decision trees in early stage of model training, state sequences are reliably estimated. In continuous phoneme recognition experiments, the proposed method can improve the recognition performance.

Index Terms: Continuous speech recognition, Acoustic modeling, Context clustering, Phonetic decision tree, Deterministic annealing

10.21437/Interspeech.2008-111

1. Introduction

The expectation maximization (EM) algorithm is widely used for parameter estimation of statistical models with hidden variables. 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 defined by the principle of maximum entropy and a statistic mechanics analogy. The posterior distribution derived in the DAEM algorithm includes a "temperature" parameter which controls the influence of unreliable model parameters. Gradually decreasing the temperature while iterating the EM-steps, the annealing process can reduce the dependency on initial model parameters. Finally the reliable model parameters can be obtained by the DAEM algorithm. It has been reported that the DAEM algorithm is effective for HMM-based speech recognition [2].

In large vocabulary continuous speech recognition systems, context-dependent models (e.g. triphone HMMs) are widely used. Although a large number of triphones can capture variations in speech data, too many parameters cause over-fitting. Therefore, maintaining a good balance between the model complexity and robustness is important to achieve good recognition performance. The phonetic decision tree-based context clustering technique [3] is one of good solutions for this problem. This technique constructs the parameter tying structure which can assign a sufficient amount of training data to each HMM state. The embedded training followed by the context clustering can estimate reliable model parameters based on the appropriate model structure. However, the technique requires statistics of HMM state sequences obtained from unreliable model parameters (without parameter tying). Although we need reliable model parameters to construct the appropriate parameter tying structure, estimating reliable model parameters requires the appropriate tying structure. Hence, model parameters and parameter tying structure should be jointly optimized. However, exact solution of this optimization is computationally intractable. In this paper, we reformulate this optimization problem as maximizing a newly defined likelihood function which includes the parameter tying structure as a hidden variable. Furthermore, we relax the problem using the variational approximation and solve it based on the DAEM algorithm. Both hidden variables (state sequence and parameter tying structures) are optimized in the annealing process.

The rest of this paper is organized as follows. Section 2 describes the DAEM algorithm, and Section 3 describes the acoustic modeling based on model structure annealing. Experimental results are presented in Section 4, and concluding remarks and future works are presented in the final section.

2. Deterministic annealing EM algorithm in parameter estimation

2.1. 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:

$$\mathcal{L}(\Lambda) = \log \sum_{\forall \boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda), \tag{1}$$

where Λ denotes a set of model parameters and $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ and $\boldsymbol{q} = (q_1, q_2, ..., q_T)$ are the observation and state sequences, respectively. The EM algorithm iteratively maximizes the auxiliary function so called $\mathcal{Q}$-function:

$$\mathcal{Q}(\Lambda, \Lambda') = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{q} \mid \boldsymbol{o}, \Lambda) \log P(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda') \tag{2}$$

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

$$P(\boldsymbol{q} \mid \boldsymbol{o}, \Lambda) = \frac{P(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda)}{\sum_{\forall \boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda)}. \tag{3}$$

In the DAEM algorithm [1], the problem of maximizing the log-likelihood function is reformulated as the problem of

Accepted after peer review of full paper

minimizing the following free energy function:

$$\mathcal{F}_\beta(\Lambda) = -\frac{1}{\beta}\log\sum_{\forall \boldsymbol{q}} P^\beta(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda)$$
$$= -\sum_{\forall \boldsymbol{q}} f(\boldsymbol{q}; \boldsymbol{o}, \Lambda)\log P(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda) - \frac{1}{\beta} I[f(\boldsymbol{q}; \boldsymbol{o}, \Lambda)], \tag{4}$$

where $I[x]$ denotes the entropy of x and $1/\beta$ is called as "temperature." If $\beta = 1$, the negative free energy $-\mathcal{F}_\beta(\Lambda)$ becomes equal to the log-likelihood function $\mathcal{L}(\Lambda)$. In the deterministic annealing approach, the new posterior distribution f is derived so as to minimize the free energy under the constraint of $\sum_{\forall \boldsymbol{q}} f = 1$. 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

$$f(\boldsymbol{q}; \boldsymbol{o}, \Lambda) = \frac{P^\beta(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda)}{\sum_{\forall \boldsymbol{q}} P^\beta(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda)}. \tag{5}$$

In the DAEM algorithm, the temperature parameter β is gradually increased while iterating the EM-steps at each temperature. When $1/\beta$ is set to an initial temperature $\beta^{(0)} \simeq 0$, the EM-steps may achieve a single global minimum of $\mathcal{F}_\beta(\Lambda)$. 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.

2.2. Optimization of state sequences

In the HMM case, the DAEM posterior distribution f can be calculated by the forward-backward algorithm. The numerator of the posterior distribution in Eq. (5) is written as

$$P^\beta(\boldsymbol{o}, \boldsymbol{q} \mid \Lambda) = P^\beta(\boldsymbol{o} \mid \boldsymbol{q}, \Lambda) P^\beta(\boldsymbol{q} \mid \Lambda)$$
$$= \prod_{t=1}^{T} P^\beta(\boldsymbol{o}_t \mid q_t, \Lambda) \prod_{t=1}^{T} P^\beta(q_t \mid q_{t-1}, \Lambda), \tag{6}$$

where $P(\boldsymbol{o}_t \mid q_t, \Lambda)$ and $P(q_t \mid q_{t-1}, \Lambda)$ indicate state output and transition probabilities, respectively. It can be observed that Eq. (6) has the same form as the likelihood function of HMMs. Therefore, the expectations with respect to the DAEM posterior distribution f can be calculated by replacing the state output and transition probabilities with $P^\beta(\boldsymbol{o}_t \mid q_t, \Lambda)$ and $P^\beta(q_t \mid q_{t-1}, \Lambda)$, respectively.

3. Acoustic modeling based on model structure annealing

In this paper to derive the algorithm of model structure annealing, we define a new likelihood function which includes parameter tying structure as a hidden variable as follows:

$$P(\boldsymbol{o} \mid \Lambda) = \sum_{\forall \boldsymbol{q}} \sum_{\forall m} P(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda), \tag{7}$$

$$P(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda) = P(m) P(\boldsymbol{q} \mid \Lambda) P(\boldsymbol{o} \mid \boldsymbol{q}, m, \Lambda), \tag{8}$$

where $m \in \{1, ..., M\}$ indexes tying structure and $\Lambda \in \{\Lambda_1, ..., \Lambda_M\}$ denotes a set of model parameters for the m-th tying structure. We construct each parameter tying structure by a phonetic decision tree. In the EM algorithm, the ML estimate of the model parameters is obtained using the posterior distribution of hidden variables estimated in the E-step. Therefore, solving the ML problem for the newly defined model is regarded as the simultaneous optimization of state sequences and parameter tying structure. The free energy function of the proposed model for the DAEM algorithm also can be written as

$$\mathcal{F}_\beta(\Lambda) = -\frac{1}{\beta}\log\sum_{\forall \boldsymbol{q}}\sum_{\forall m} P^\beta(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda). \tag{9}$$

However, estimating the posterior distribution $f(\boldsymbol{q}, m; \boldsymbol{o}, \Lambda)$ is intractable due to the combination of hidden variables. To solve this problem, we apply the variational EM algorithm [4]. The objective of the algorithm is to maximize an upper bound of the free energy function. The upper bound $\bar{\mathcal{F}}_\beta(\Lambda)$ is defined as

$$\mathcal{F}_\beta(\Lambda) = -\frac{1}{\beta}\log\sum_{\forall \boldsymbol{q}}\sum_{\forall m} Q(\boldsymbol{q}, m)\frac{P^\beta(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda)}{Q(\boldsymbol{q}, m)}$$
$$\leq -\frac{1}{\beta}\sum_{\forall \boldsymbol{q}}\sum_{\forall m} Q(\boldsymbol{q}, m)\log\frac{P^\beta(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda)}{Q(\boldsymbol{q}, m)}$$
$$= \bar{\mathcal{F}}_\beta(\Lambda) \tag{10}$$

where $Q(\boldsymbol{q}, m)$ is an arbitrary distribution. The upper bound $\bar{\mathcal{F}}_\beta(\Lambda)$ can be transformed as follows:

$$\bar{\mathcal{F}}_\beta(\Lambda) = \frac{1}{\beta} KL(Q \mid\mid f) - \log P(\boldsymbol{o} \mid \Lambda) + \text{const} \tag{11}$$

The above equation shows that minimizing $\bar{\mathcal{F}}_\beta(\Lambda)$ with respect to $Q(\boldsymbol{q}, m)$ is equivalent to minimizing the KL-divergence between Q and f. Although if there is no constraint with distribution Q, minimizing $\bar{\mathcal{F}}_\beta(\Lambda)$ results $f = Q$. Assuming a constraint to reduce the complexity, the distribution Q which minimizes $\bar{\mathcal{F}}_\beta(\Lambda)$ becomes an approximate distribution of f. In this paper, we assume the following constraint:

$$Q(\boldsymbol{q}, m) = Q(\boldsymbol{q})Q(m) \tag{12}$$

where $\sum_{\forall \boldsymbol{q}} Q(\boldsymbol{q}) = 1$ and $\sum_{\forall m} Q(m) = 1$. Using these factorized distributions, the upper bound $\bar{\mathcal{F}}_\beta(\Lambda)$ can be rewritten as

$$\bar{\mathcal{F}}_\beta(\Lambda) = -\sum_{\forall \boldsymbol{q}}\sum_{\forall m} Q(\boldsymbol{q})Q(m)\log P(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda)$$
$$- \frac{1}{\beta} I[Q(\boldsymbol{q})] - \frac{1}{\beta} I[Q(m)]. \tag{13}$$

It can be seen that the temperature parameter β changes the ratio between the value of $\mathcal{Q}$-function and the entropy of hidden variables in $\bar{\mathcal{F}}_\beta(\Lambda)$. Extending this interpretation, we can control the annealing process of decision trees and state sequences individually. By introducing $\beta_{\boldsymbol{q}}$ and β_m, $\bar{\mathcal{F}}_\beta(\Lambda)$ is rewritten by

$$\bar{\mathcal{F}}_\beta(\Lambda) = -\sum_{\forall \boldsymbol{q}}\sum_{\forall m} Q(\boldsymbol{q})Q(m)\log P(\boldsymbol{o}, \boldsymbol{q}, m \mid \Lambda)$$
$$- \frac{1}{\beta_{\boldsymbol{q}}} I[Q(\boldsymbol{q})] - \frac{1}{\beta_m} I[Q(m)]. \tag{14}$$

The optimal variational posterior distribution of $Q(\boldsymbol{q})$ and $Q(m)$ are derived by minimizing $\bar{\mathcal{F}}_\beta(\Lambda)$. This functional optimization can be solved by the variational method, and the following formulae are obtained:

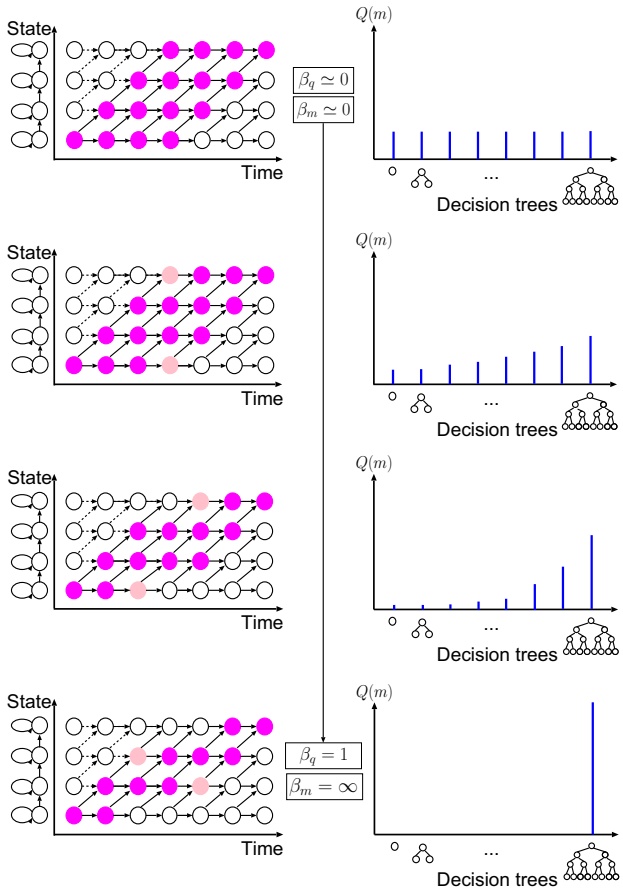

Figure 1: Joint optimization process

$$Q(\boldsymbol{q}) \propto \left[P(\boldsymbol{q} \mid \Lambda) \exp \Big\langle \log P(\boldsymbol{o} \mid \boldsymbol{q}, m, \Lambda) \Big\rangle_{Q(m)} \right]^{\beta_{\boldsymbol{q}}} \tag{15}$$

$$Q(m) \propto \left[P(m) \exp \Big\langle \log P(\boldsymbol{o} \mid \boldsymbol{q}, m, \Lambda) \Big\rangle_{Q(\boldsymbol{q})} \right]^{\beta_m} \tag{16}$$

where $\langle \cdot \rangle_{Q(\cdot)}$ denotes the expectation with respect to the distribution $Q(\cdot)$. Since Eqs. (15) and (16) are dependent each other, these updates are iterated in the E-step. Figure 1 illustrates the proposed joint optimization process based on the DAEM algorithm. At the initial temperature ($\beta_{\boldsymbol{q}}^{(0)}, \beta_m^{(0)} \simeq 0$), the variational posterior distributions $Q(\boldsymbol{q})$ and $Q(m)$ take a form nearly uniform distribution. While the temperature is decreasing, the form of $Q(\boldsymbol{q})$ and $Q(m)$ change from uniform to each original posterior distribution, and at the final temperature ($\beta_{\boldsymbol{q}}, \beta_m = 1$), $Q(\boldsymbol{q})$ and $Q(m)$ take each original posterior distribution. Then, the posterior probability of model structures is in proportion to the likelihood of each model structure. However, it is computationally expensive to use multiple model structures in decoding. Hence we choose a single model structure by setting the temperature β_m to ∞ (the DAEM algorithm with $\beta_{\boldsymbol{q}} = \infty$ becomes the Viterbi training, however the final temperature is fixed as $\beta_{\boldsymbol{q}} = 1$ in this paper). Although the model structure with the largest decision tree is selected at $\beta_m = \infty$ in most cases, reliable state sequences can be obtained by using multiple model structures in the early stage of training procedure.

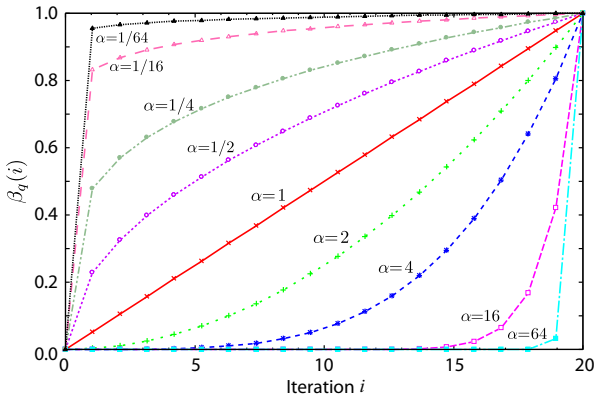

Figure 2: Schedule of temperature parameter $\beta_{\boldsymbol{q}}$.

4. Experiments

To evaluate the effectiveness of the proposed model structure annealing technique, continuous phoneme recognition experiment was conducted.

4.1. Experimental condition

We used phonetically balanced 503 sentences uttered by a single male speaker MHT from the ATR Japanese speech database b-set. The 450 sentences were used for training HMMs and the remaining 53 sentences were used for testing. The speech data was down-sampled from 20kHz to 16kHz, windowed at a 25-ms Blackman window, and parameterized into 19 mel-cepstral coefficients with the mel-cepstral analysis technique [5]. Static coefficients including the zero-th coefficients and their first and second derivatives were used as feature parameters. Three-state left-to-right HMMs were used to model 43 Japanese phonemes, and 118 questions were prepared for decision tree clustering. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix.

In this experiment, the following five training methods were compared.

- "flat-start": HMMs were initialized by equal mean and variance for all states, and re-estimated using the EM algorithm.
- "k-means": HMMs were initialized by the segmental k-means algorithm using phoneme boundary labels and re-estimated using the EM algorithm.
- "DAEM-state": The DAEM algorithm was applied to state sequence. A single decision tree was used.
- "DAEM-tree": The DAEM algorithm was applied to decision trees. For state sequences, it is equivalent to "flat-start."
- "DAEM-joint": The DAEM algorithm was applied to both state sequences and to decision trees.

For the conventional methods ("flat-start," "k-means" and "DAEM-state"), decision tree clustering based on the MDL criterion [6] was used. Monophone has 111 leaves and MDL has 1097 leaves. In addition to this model, the proposed methods ("DAEM-tree" and "DAEM-joint") used a decision tree representing monophone HMMs. In total two decision trees were used for model structure annealing ($m = 1$: monophone, $m = 2$: MDL). It is noted that only one decision tree (MDL)

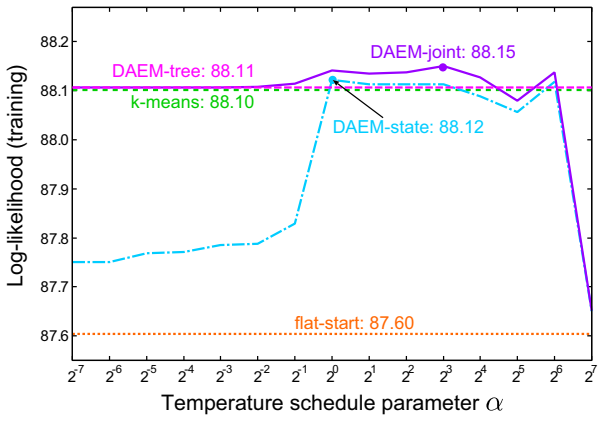

Figure 3: Log-likelihood of training data

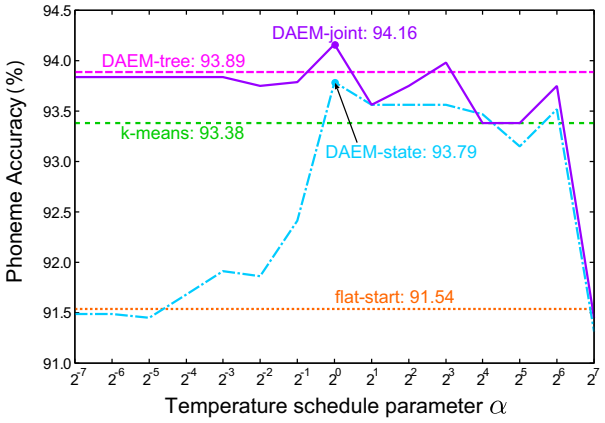

Figure 4: Phoneme accuracy

is used for decoding. The temperature parameter for $\boldsymbol{q}$ was updated by

$$\beta_{\boldsymbol{q}}(i) = \left(\frac{i}{I}\right)^{\alpha}, i = 0, ..., I \tag{17}$$

where i denotes the iteration number to update temperature parameter, and α was varied as $\alpha = 2^n (n = -7, ..., 7)$, and $\alpha = \infty$. Figure 2 plots the temperature parameter $\beta_{\boldsymbol{q}}$. Since only two decision trees were used in this experiment, determining the temperature parameter β_m is equivalent to setting the variational posterior probabilities $Q(m)$ directly. Therefore, it was assumed that the posterior distribution of the decision trees was updated by the following linear functions $Q(1) = 0.5(1 - i/I), Q(2) = 0.5(1 + i/I)$. The number of EM-steps was set to 30 for the conventional methods. In the DAEM algorithm, the number of temperature update steps was set to 20 ($I = 20$), and 30 EM-steps were conducted at each temperature.

4.2. Experimental results

Figure 3 compares the log-likelihood of the training data. It can be seen that the likelihood of "flat-start" was lower than that of "k-means." This is because "flat-start" uses no phoneme boundary information for initializing HMMs and inappropriate initial model parameters cause the local maxima problem. Although "DAEM-state" also uses no phoneme boundaries, the likelihood of "DAEM-state" was close to that of "k-means" when an appropriate temperature scheduling was used. This result confirmed that the local maxima problem can be relaxed by the DAEM algorithm. Comparing the proposed structure annealing with the conventional methods, "DAEM-tree" achieved the similar likelihood of "k-means" and "DAEM-state." Furthermore, "DAEM-joint" obtained the highest likelihood at $\alpha = 2^3$. These results show that the structure annealing can estimate reliable state sequences using multiple decision trees.

Figure 4 shows the phoneme accuracy of each method. Similar to the likelihood, "flat-start" was worse than the other methods because of local maxima problem. It can also be seen that the methods using the DAEM algorithm outperformed "k-means," even though phoneme boundary information is not used in the DAEM algorithm. Moreover, "DAEM-tree" and "DAEM-joint" improved the performance as compared with the conventional "DAEM-state," and "DAEM-joint" achieved 11.7% relative error reduction over "k-means" at $\alpha = 2^0$. This result indicates that reliably estimated HMM parameters using the structure annealing are effective for improving the speech recognition performance.

5. Conclusion

This paper proposed the HMM training technique using multiple phonetic decision trees for speech recognition. In the proposed method, this was performed by maximum likelihood (ML) estimation of the newly defined statistical model which includes multiple decision trees as hidden variables. Applying the deterministic annealing expectation maximization (DAEM) algorithm and using multiple decision trees in early stage of model training, state sequences were reliably estimated. In continuous phoneme recognition experiments, the proposed method achieved higher accuracy than the standard EM algorithm with segmental k-means initialization and the DAEM algorithm using a single decision tree.

As a future work, we will investigate the effect of increasing the number of the decision trees. We will also perform the experiments with a different amount of training data and with various update schemes of temperature parameter in the DAEM algorithm.

6. 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] J. J. Odel, "The Use of Context in Large Vocabulary Speech Recognition, " Ph.D. dissertation, Cambridge University, 1995.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," Machine Learning, vol.37, pp.183–233, 1999.

[5] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," Proc. of ICASSP, vol.1, pp.137–140, 1992.

[6] K. Shinoda and T. Watanabe, "Acoustic Modeling Based on the MDL Principle for Speech Recognition," Proc. of Eurospeech, vol.1, pp.99–102, 1997.