社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (38)
2008 / 12 / 10

複数の音素決定木構造を含む統計モデルに基づく音声認識

塩田 さやか† 橋本 佳† 南角 吉彦† 李 晃伸† 徳田 恵一†

† 名古屋工業大学 大学院工学研究科 情報工学専攻
〒 466-8555 愛知県名古屋市昭和区御器所町

あらまし 近年，連続音声認識システムにおける音響モデルとして，コンテキスト，音素環境を考慮したコンテキスト依存モデルが広く利用されている．また，与えられた学習データ量に応じて適切なパラメータ共有構造を得るために，決定木に基づくコンテキストクラスタリングが用いられる．しかし，音素決定木の構築には，あらかじめ推定した状態系列に関する情報が必要となる．一方，最適な状態系列の推定にはモデルのパラメータ共有構造をあらかじめ決定しておく必要がある．このように，決定木構造と状態系列は，その推定に相互を必要とするため，同時最適化は困難であった．そこで我々はこれまでに，複数の決定木構造を含む確率モデルを定義し，確定的アニーリング EM(Deterministic Annealing Expectation Maximization; DAEM) アルゴリズムを用いることで同時最適化を近似する手法を提案してきた．しかし，通常のデコーダに適用可能とするため，認識においては決定木構造を 1 つだけ選択して用いてきた．そこで，本稿では，学習時と同様の枠組を用いて認識時にも複数の決定木構造を用いることを検討する．

キーワード 連続音素認識，音響モデル，コンテキストクラスタリング，音素決定木構造，確定的アニーリング

Speech Recognition Based on Statistical Models Including Multiple Decision Trees

Sayaka SHIOTA†, Kei HASHIMOTO†, Yoshihiko NANKAKU†, Akinobu LEE†, and Keiichi TOKUDA†

† Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showak-ku, Nagoya, Aichi, 466-8555 Japan

Abstract This paper proposes a speech recoginition technique using multiple decision trees. 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 overcome this problem, we have proposed a simultaneous optimization technique of state sequences and decision trees based on annealing process using multiple decision trees. However, this technique uses a single decision tree in decoding. In this paper, we use multiple decision trees not only in training process but also in decoding within the same annealing framework.

Key words Continuous speech recognition, Acoustic modeling, Context clustering, Phonetic decision tree, 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. The posterior distribution derived in the DAEM algorithm 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 [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 overfitting. 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 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 model parameters. That is, 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, the exact solution of this optimization is computationally intractable. To overcome this problem, we reformulated this optimization problem as maximizing a newly defined likelihood function which includes the parameter tying structure as a hidden variable. Furthermore, the variational approximation and the DAEM algorithm were adopted to derive a tractable algorithm for achieving a good sub-optimal solution. However, this technique uses a single decision tree in decoding, because of the complexity and applicability to standard decoders [7]. In this paper, we proposed a speech recognition technique using multiple decision trees not only in training process but also in decoding within the same annealing framework.

The rest of this paper is organized as follows. Section 2 describes the DAEM algorithm, and Section 3 describes speech recognition based on multiple phonetic decision trees. 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:

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

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

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

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:

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

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

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

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 $\sum_{\boldsymbol{q}} f = 1$. To solve this problem, we can use elementary calculus of variations to take functional derivatives of equation (4) with respect to f, and the optimal distribution can be derived as

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

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 equation (5) is written as

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

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 equation. (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. Speech Recognition Based on Multiple Phonetic Decision Trees

3.1 Acoustic modeling based on model structure annealing

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:

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

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

where $m \in \{1, \ldots, M\}$ indexes parameter tying structures and $\Lambda \in \{\Lambda_1, \ldots, \Lambda_M\}$ 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 multiple decision tree model for the DAEM algorithm also can be written as

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

However, estimating the DAEM posterior distribution $f(\boldsymbol{q}, m \mid \boldsymbol{o}, \Lambda)$ 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 minimize an upper bound of the free energy function. The upper bound of the free energy function $\bar{\mathcal{F}}_{\beta}(\Lambda)$ is defined as

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

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

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

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. Hence, we assume the following constraint:

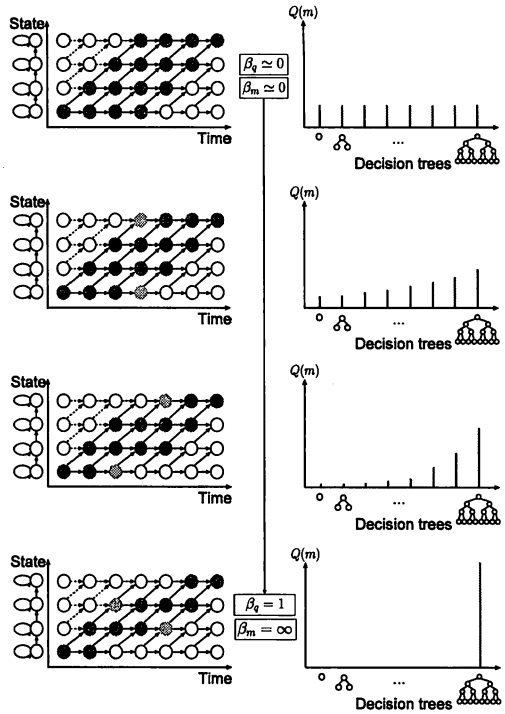

Fig. 1 Joint optimization process

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

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

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

It can be seen that the temperature parameter β changes the ratio between the value of 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

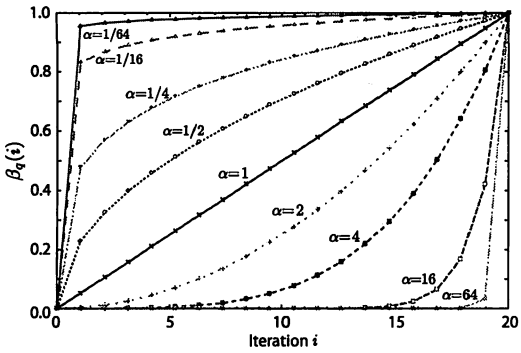

Fig. 2 Schedule of temperature parameter β_q

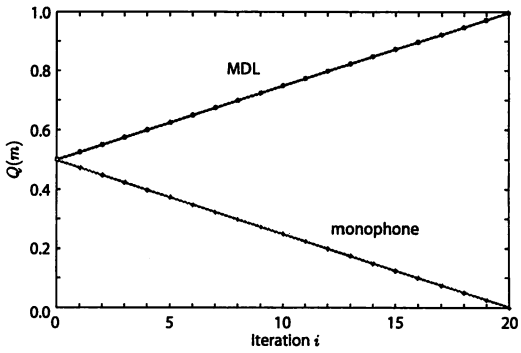

Fig. 3 Schedule of temperature parameter β_m

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

The optimal variational posterior distribution of $Q(q)$ 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:

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

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

where $\langle\cdot\rangle_{Q(\cdot)}$ denotes the expectation with respect to the distribution $Q(\cdot)$. Since eqations (15) and (16) are dependent each other, these updates should be iterated in the E-step. Figure 1 illustrates the joint optimization process based on the DAEM algorithm. At the initial temperature ($\beta_q^{(0)}, \beta_m^{(0)} \simeq 0$), the variational posterior distributions $Q(q)$ and $Q(m)$ take a form nearly uniform distribution. While the temperature is decreasing, the form of $Q(q)$ and $Q(m)$ change from uniform to each original posterior distribution, and at the final temperature ($\beta_q, \beta_m = 1$), $Q(q)$ 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, the multiple decision tree models are inapplicable to standard decoders. Hence, we choose a single model structure by setting the temperature β_m to ∞ (the DAEM algorithm with $\beta_q = \infty$ becomes the Viterbi training, however the final temperature is fixed as $\beta_q = 1$ 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.

3.2 Speech decoding based on multiple model structures

In the conventional model structure annealing, a single model structure was chosen because of the complexity and applicability to standard decoders. In this paper, we proposes a speech recognition technique using multiple model structures not only in training process but also in decoding. Although there are many approaches of using multiple model structures in decoding, we use a method which stops the annealing in the early stage of training. Using $Q(m)$, the criterion for decoding can be written as:

$$\max_q P(q \mid \Lambda) \prod_m P^{Q(m)}(o,q,m \mid \Lambda). \tag{17}$$

By inspection, this criterion can be calculated by the output probabilities of a multi stream HMM where $Q(m)$ becomes the weight of each stream.

4. Experiments

To evaluate the effectiveness of the proposed method, a 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 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. 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

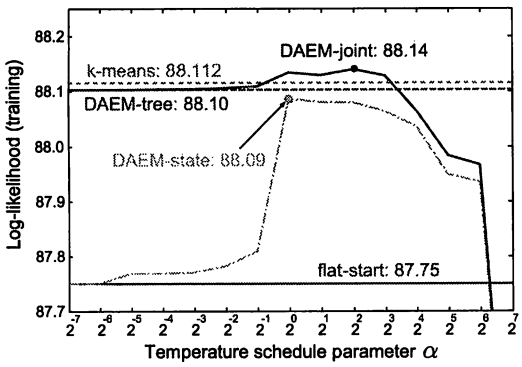

Fig. 4 Log-likelihood of training data

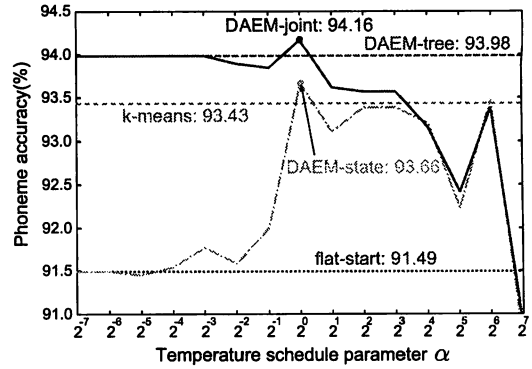

Fig. 5 Phoneme accuracy for each temperature schedule

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 only to state sequence. A single decision tree was used.
- "DAEM-tree": The DAEM algorithm was applied only 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 methods using a single decision tree ("flat-start," "k-means" and "DAEM-state"), a decision tree obtained by the context clustering based on the MDL criterion was used. In addition to this model, "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). Monophone has 111 leaves and MDL has 1097 leaves. The temperature parameter β_q was updated by

$$\beta_q(i) = (i/I)^{\alpha}, (i = 0, \ldots, I) \tag{18}$$

where i denotes the iteration number of temperature updates, and α was varied as $\alpha = 2^n (n = -7, \ldots, 7)$ and $\alpha = \infty$. 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 $Q(m)$ was updated by the following linear functions $Q(1) = 0.5(1 - i/I), Q(2) = 0.5(1 + i/I)$. Figures 2 and 3 plot each the schedule of the temperature parameter β_q and β_m, respectively. The number of EM-steps was set to 200 for the standard EM algorithm. In the DAEM algorithm, the number of temperature update steps was set to 20 ($I = 20$), and 10 EM-steps were conducted at each temperature.

4.2 Experimental results using a single decision tree in training

Figure 4 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 $\alpha = 2^2$. These results show that the structure annealing can estimate reliable state sequences using multiple decision trees.

Figure 5 shows the phoneme accuracy of each method. It is noted that only one decision tree (MDL) is used for decoding. 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 11.1% 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.

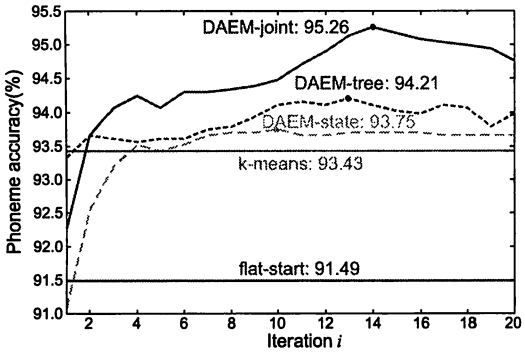

Fig. 6 Phoneme accuracy (without insertion penalty)

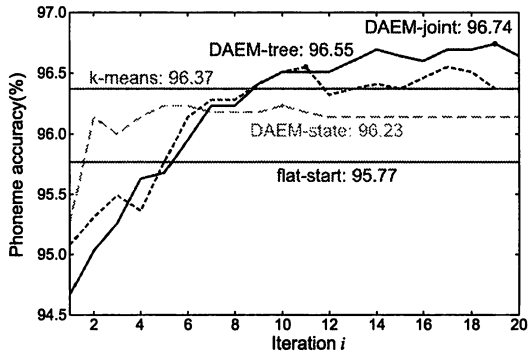

Fig. 7 Phoneme accuracy (with insertion penalty)

4.3 Experimental results using multiple decision trees in decoding

For the decoding using multiple decision trees, the models obtained at $i < 20$ were used and the performance was evaluated at each iteration. The schedule of temperature parameter β_q was updated by equation. (18) with $\alpha = 1$. Figures 6 and 7 show the phoneme accuracy without and with insertion penalty, respectively, while varying the temperature in decoding. The results at $i = 20$ of "DAEM-tree" and "DAEM-joint" indicate using a single decision tree in decoding, even though multiple decision trees were used in the training process. It can be seen from figure 5 that the decoding using multiple decision trees ($i < 20$) improves the accuracy by setting the appropriate temperature. Although "DAEM-state" was also evaluated at $i < 20$, no improvement was obtained because it used only one decision tree. In figure 7, although the improvement decreased by using the insertion penalty, the decoding using multiple decision trees still improved the performance as compared with that at $i = 20$. These results indicate that not only in training process but also in speech decoding, the use of multiple decision trees is effective for improving the performance of speech recognition.

5. Conclusion

This paper proposed the speech recognition technique using multiple decision trees in decoding. In continuous phoneme recognition experiments, the proposed technique improved the performance of speech recognition. As future works, we will investigate other methods of using multiple decision trees in decoding. We will also perform experiments on a larger database.

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.
p

[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.

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