IEICE TRANS. INF. & SYST., VOL.E94–D, NO.3 MARCH 2011

PAPER

Bayesian Context Clustering Using Cross Validation for Speech Recognition

Kei HASHIMOTO[†a)], **Heiga ZEN**[†∗b)], ***Nonmembers***, **Yoshihiko NANKAKU**[†c)], **Akinobu LEE**[†d)], ***and*** **Keiichi TOKUDA**[†e)], ***Members***

SUMMARY This paper proposes Bayesian context clustering using cross validation for hidden Markov model (HMM) based speech recognition. The Bayesian approach is a statistical technique for estimating reliable predictive distributions by treating model parameters as random variables. The variational Bayesian method, which is widely used as an efficient approximation of the Bayesian approach, has been applied to HMM-based speech recognition, and it shows good performance. Moreover, the Bayesian approach can select an appropriate model structure while taking account of the amount of training data. Since prior distributions which represent prior information about model parameters affect estimation of the posterior distributions and selection of model structure (e.g., decision tree based context clustering), the determination of prior distributions is an important problem. However, it has not been thoroughly investigated in speech recognition, and the determination technique of prior distributions has not performed well. The proposed method can determine reliable prior distributions without any tuning parameters and select an appropriate model structure while taking account of the amount of training data. Continuous phoneme recognition experiments show that the proposed method achieved a higher performance than the conventional methods.

key words: *Bayesian approach, speech recognition, HMM, context clustering, cross validation*

1. Introduction

In hidden Markov model (HMM) based speech recognition systems [1], accurate acoustic modeling is necessary for reducing recognition error rate. The maximum likelihood (ML) criterion is one of the standard criteria for training acoustic models in speech recognition. The ML criterion guarantees to estimate the true values of the parameters as the amount of training data infinitely increases. However, the performance of current speech recognition systems is still far from satisfactory. In a real environment, there are many fluctuations originating from various factors such as the speaker, speaking style, and noise. A mismatch between the training and testing conditions often brings a drastic degradation in performance. However, since the ML criterion produces a point estimate of model parameters, the estimation accuracy may be degraded due to the over-fitting problem when the amount of training data is insufficient.

On the other hand, the Bayesian approach considers the posterior distribution of all variables [2]. That is, all the variables introduced when models are parameterized, such as model parameters and latent variables, are regarded as random variables, and their posterior distributions are obtained based on the Bayes theorem. The difference between the Bayesian and ML approaches is that the target of estimation is the distribution function in the Bayesian approach whereas it is the parameter value in the ML approach. Based on this posterior distribution estimation, the Bayesian approach can generally achieve more robust model construction and classification than the ML approach [3]–[5]. However, the Bayesian approach requires complicated integral and expectation computations to obtain posterior distributions when models have latent variables. Since the acoustic models used in speech recognition (e.g., HMMs) have the latent variables, it is difficult to apply the Bayesian approach to speech recognition directly with no approximation. Recently, the Variational Bayesian (VB) approach has been proposed in the field of learning theory to avoid complicated computations by employing the variational approximation technique [6]. With this VB approach, approximate posterior distributions are obtained effectively by iterative calculations similar to the Expectation-Maximization (EM) algorithm used in the ML approach. The VB approach has been applied to speech recognition and it shows good performance [7].

The VB approach has also been applied to the context clustering [7], [8]. It is well known that contextual factors affect speech. Therefore, context-dependent acoustic models (e.g., triphone HMMs) are widely used in HMM-based speech recognition [9], [10]. Although a large number of context-dependent acoustic models can capture variations in speech data, too many model parameters lead to the over-fitting problem. Consequently, maintaining a good balance between model complexity and the amount of training data is very important for obtaining high generalization performance. The decision tree based context clustering [11] is an efficient method for dealing with the problem of data sparseness, for both estimating robust model parameter of context-dependent acoustic models and obtaining predictive distributions of unseen contexts. This method constructs a model parameter tying structure which can assign a sufficient amount of training data to each HMM state. The tree

Manuscript received June 30, 2010.
Manuscript revised October 25, 2010.

[†]The authors are with the Department of Computer Science, Naogya Institute of Technology, Nagoya-shi, 466–8555 Japan.
[∗]Presently, with the Speech Technology Group, Toshiba Europe Research Ltb. Cambridge Research laboratory, Cambridge, UK.
a) E-mail: bonanza@sp.nitech.ac.jp
b) E-mail: heiga.zen@crl.toshiba.co.uk
c) E-mail: nankaku@sp.nitech.ac.jp
d) E-mail: ri@nitech.ac.jp
e) E-mail: tokuda@nitech.ac.jp
DOI: 10.1587/transinf.E94.D.668

is grown step by step, choosing questions that divide the set of contexts using a greedy strategy to maximize an objective function.

The ML criterion is inappropriate as a model selection criterion because it increases monotonically as the number of states increases. Some heuristic thresholding is therefore necessary to stop splitting nodes in the context clustering. To solve this problem, the minimum description length (MDL) criterion has been employed to select the model structure [12]. However, the MDL criterion is based on an asymptotic assumption, therefore it is ineffective when the amount of training data is small. On the other hand, the Bayesian information criterion (BIC) [13] has been proposed as an approximated Bayesian criterion. However, since the BIC is practically the same as the MDL criterion, The BIC is also ineffective when the amount of training data is small. In contrast to the BIC, the model selection based on the VB method has been proposed [7], [8]. The VB method can select an appropriate model structure, even when there are insufficient amounts of data, because it does not use an asymptotic assumption. Therefore, the speech recognition framework which consistently applis the VB method is effective for estimating appropriate acoustic models and model structures.

The Bayesian approach has an advantage that it can utilize prior distributions which represent the prior information of model parameters. In the Bayesian approach, since prior distributions of model parameters affect the estimation of posterior distributions and model selection, the determination of prior distributions is an important problem for estimating appropriate acoustic models. As the determination technique of prior distributions, some techniques have been proposed in the filed of machine learning, e.g., using uninformative (uniform) prior distributions, hierarchical Bayesian methods, and empirical Bayesian methods [14]. However, it has not been thoroughly investigated in speech recognition, and the determination technique of prior distributions has not performed well. This paper proposes a prior distribution determination technique using cross validation and applies it to the context clustering for the speech recognition framework based on Bayesian approach. The cross validation method is known as a straightforward and useful method for model structure optimization [15], [16]. The main idea behind cross validation is to split data for estimating the risk of each model. Part of data is used for training each model, and the remaining part is used for estimating the risk of the model. Then, the cross validation method selects the model with the smallest estimated risk. The cross validation method avoids the over-fitting problem because the training data is independent from the validation data. The context clustering based on the ML criterion using cross validation has been proposed, and it can select a more appropriate model structure than the conventional ML criterion [17]. The proposed method can be regarded as an extension of context clustering using cross validation to the Bayesian approach. Using prior distributions determined by the cross validation, it is expected that a higher generalization ability is achieved and an appropriate model structure can be selected in the context clustering without any tuning parameters. Comparing to [18], this paper describes the detailed derivation of the proposed method and shows further experimental results.

The rest of the present paper is organized as follows. Section 2 describes speech recognition based on the variational Bayesian method. Section 3 derives the prior distribution determination technique using cross validation and apply it to the context clustering. Results of the continuous phoneme recognition experiments are shown in Sect. 4. Concluding remarks and future plans are presented in the final section.

2. Speech Recognition Based on Variational Bayesian Method

2.1 Bayesian Approach

The output distribution is obtained based on a left-to-right HMM which has been widely used to represent an acoustic model for speech recognition. Let $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ be a set of training data of D dimensional feature vectors, and T is used to denote the number of frames. The log output distribution is represented by

$$\begin{aligned}\log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) &= \sum_{i=1}^{N} z_1^i \log \pi_i + \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} z_t^i z_{t+1}^j \log a_{ij} \\ &\quad + \sum_{t=1}^{T} \sum_{i=1}^{N} z_t^i \log \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1}) \end{aligned} \tag{1}$$

where $\boldsymbol{Z} = (z_1, z_2, \ldots, z_T)$ is a sequence of latent variables which represent HMM states, $z_t \in \{1, \ldots, N\}$ denotes a state at frame t, and N is the number of states in an HMM.

$$z_t^i = \delta(z_t, i) = \begin{cases} 1 & \text{if } z_t = i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

A set of model parameters $\boldsymbol{\Lambda} = \{\pi_i, a_{ij}, \boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i,j=1}^{N}$ consists of the initial state probability π_i of state i, the state transition probability a_{ij} from state i to state j, the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{S}_i^{-1}$ of a Gaussian distribution $\mathcal{N}(\cdot \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1})$.[†]

In HMM-based speech recognition, the ML criterion has typically been used to train HMMs. In the ML criterion, the optimal model parameters are estimated by maximizing the likelihood for given training data as follows.

$$\begin{aligned}\boldsymbol{\Lambda}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \mid \boldsymbol{\Lambda}) \\ &= \arg\max_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{Z}} P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) \end{aligned} \tag{3}$$

[†]Although a multi-mixture Gaussian is typically used as a state output probability distribution in recent HMM-based speech recognition systems, a single Gaussian is assumed as a state output probability distribution in this paper for simplification.

The model parameters can be estimated using an iterative procedure such as the EM algorithm [19] because it is difficult to obtain the model parameters $\mathbf{\Lambda}_{\mathrm{ML}}$ analytically. The ML criterion guarantees to estimate the true values of the model parameters as the amount of training data infinitely increases. However, the ML criterion produces a point estimate of model parameters. The use of point estimate will cause an over-fitting problem when the amount of training data is insufficient. A overfitted model will generally have poor predictive performance, because it captures minor fluctuations in the training data.

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 given by the famous Bayes theorem as follows.

$$P(\mathbf{\Lambda} \mid \boldsymbol{O}) = \frac{P(\boldsymbol{O} \mid \mathbf{\Lambda})P(\mathbf{\Lambda})}{P(\boldsymbol{O})} \tag{4}$$

where $P(\mathbf{\Lambda})$ is a prior distribution for $\mathbf{\Lambda}$, and $P(\boldsymbol{O})$ is an evidence.

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

$$P(\boldsymbol{X} \mid \boldsymbol{O}) = \int P(\boldsymbol{X} \mid \mathbf{\Lambda})P(\mathbf{\Lambda} \mid \boldsymbol{O})d\mathbf{\Lambda} \tag{5}$$

The model parameters are integrated out in Eq. (5) so that the effect of over-fitting is mitigated, and robust classification is achieved. However, the Bayesian approach requires complicated integral and expectation calculations to obtain posterior distributions when models include latent variables. To overcome this problem, maximum a posterior (MAP) approach has been proposed [20]. In the MAP approach, the optimal model parameters are estimated by maximizing the posterior probability. The MAP criterion can utilize the prior distribution $P(\mathbf{\Lambda})$, and can be seen as an extension of the ML criterion. However, it also produces a point estimate of HMM parameters. Consequently, it still has the effect of the over-fitting due to a point estimate.

On the other hand, the variational Bayesian (VB) method has been proposed as a tractable approximation method of the Bayesian approach [6]. The VB method avoids complicated computations by employing the variational approximation technique, and estimates approximate posterior distributions effectively by iterative calculations similar to the EM algorithm in the ML approach.

2.2 Variational Bayesian Method

In the variational Bayesian method, an approximate posterior distribution is estimated by maximizing 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.

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

where $Q(\boldsymbol{Z}, \mathbf{\Lambda})$ is an arbitrary distribution. The relation between the log marginal likelihood and the lower bound $\mathcal{F}$ is represented by using the Kullback-Leibler (KL) divergence [21] between $Q(\boldsymbol{Z}, \mathbf{\Lambda})$ and true posterior distribution $P(\boldsymbol{Z}, \mathbf{\Lambda} \mid \boldsymbol{O})$.

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

where $\mathrm{KL}[Q(\boldsymbol{Z}, \mathbf{\Lambda}) \mid P(\boldsymbol{Z}, \mathbf{\Lambda} \mid \boldsymbol{O})]$ denote a KL divergence. As the difference between the true log marginal likelihood and the lower bound is reduced, $Q(\boldsymbol{Z}, \mathbf{\Lambda})$ approximate the true posterior distribution $P(\boldsymbol{Z}, \mathbf{\Lambda} \mid \boldsymbol{O})$. Therefore, the optimal posterior distribution is estimated by the variational method, which results in minimizing the right hand side of Eq. (7).

To obtain approximate posterior distributions (VB posterior distributions) $Q(\boldsymbol{Z}, \mathbf{\Lambda})$, it is assumed that random variables are conditionally independent each other.

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

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

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

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

where $\langle \cdot \rangle_Q$ denotes the expectation with respect to Q, $C_{\mathbf{\Lambda}}$ and $C_{\boldsymbol{Z}}$ are the normalization terms of $Q(\mathbf{\Lambda})$ and $Q(\boldsymbol{Z})$, respectively. Moreover, it is assumed that the model parameters $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^N$, $\boldsymbol{a}_i = \{a_{ij}\}_{j=1}^N$, and $\{\boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i=1}^N$ are independent each other in the prior distribution $P(\mathbf{\Lambda})$. Therefore, the prior distribution $P(\mathbf{\Lambda})$ can be represented as follows.

$$P(\mathbf{\Lambda}) = P(\boldsymbol{\pi}) \prod_{i=1}^{N} P(\boldsymbol{a}_i) \prod_{i=1}^{N} P(\boldsymbol{\mu}_i, \boldsymbol{S}_i) \tag{11}$$

By using this assumption, the posterior distribution $Q(\mathbf{\Lambda})$ and its normalization term $C_{\mathbf{\Lambda}}$ can be written as follows.

$$Q(\mathbf{\Lambda}) = Q(\boldsymbol{\pi}) \prod_{i=1}^{N} Q(\boldsymbol{a}_i) \prod_{i=1}^{N} Q(\boldsymbol{\mu}_i, \boldsymbol{S}_i) \tag{12}$$

$$C_{\mathbf{\Lambda}} = C_{\boldsymbol{\pi}} \prod_{i=1}^{N} C_{\boldsymbol{a}_i} \prod_{i=1}^{N} C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i} \tag{13}$$

From Eqs. (1), (2) and (9)–(13), the posterior distributions of model parameters are given as follows.

$$Q(\boldsymbol{\pi}) = C_{\boldsymbol{\pi}} P(\boldsymbol{\pi}) \exp\left\{\sum_{i=1}^{N} \langle z_1^i \rangle \log \pi_i\right\} \tag{14}$$

$$Q(\boldsymbol{a}_i) = C_{\boldsymbol{a}_i} P(\boldsymbol{a}_i) \times \exp\left\{\sum_{j=1}^{N}\sum_{t=1}^{T-1} \langle z_t^i z_{t+1}^j \rangle \log a_{ij}\right\} \tag{15}$$

$$Q(\boldsymbol{\mu}_i, \boldsymbol{S}_i) = C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i} P(\boldsymbol{\mu}_i, \boldsymbol{S}_i) \times \exp\left\{\sum_{t=1}^{T} \langle z_t^i \rangle \log \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i)\right\} \tag{16}$$

where $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^N$ is a set of initial state probabilities, $\boldsymbol{a}_i = \{a_{ij}\}_{j=1}^N$ is a set of state transition probabilities from state i, and $\langle z_t^i \rangle$ and $\langle z_t^i z_{t+1}^j \rangle$ are the expectation value with respect to $Q(\boldsymbol{Z})$ as follows.

$$\langle z_t^i \rangle = \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) z_t^i \tag{17}$$

$$\langle z_t^i z_{t+1}^j \rangle = \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) z_t^i z_{t+1}^j \tag{18}$$

The posterior distribution $Q(\boldsymbol{Z})$ can be represented by using Eqs. (1), (2) and (10)–(16) as follows.

$$\begin{aligned} Q(\boldsymbol{Z}) = C_Z &\prod_{i=1}^{N} \exp\left\{z_1^i \left\langle \log \pi_i \right\rangle_{Q(\boldsymbol{\pi})}\right\} \\ &\times \prod_{t=1}^{\hat{T}-1}\prod_{i=1}^{N}\prod_{j=1}^{N} \exp\left\{z_t^i z_{t+1}^j \left\langle \log a_{ij} \right\rangle_{Q(\boldsymbol{a}_i)}\right\} \\ &\times \prod_{t=1}^{\hat{T}}\prod_{i=1}^{N} \exp\left\{z_t^i \left\langle \log \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1}) \right\rangle_{Q(\boldsymbol{\mu}_i, \boldsymbol{S}_i)}\right\} \end{aligned} \tag{19}$$

The posterior distribution $Q(\boldsymbol{Z})$ is similar to the likelihood function of an HMM when the terms $\exp\left\{\langle \log \pi_i \rangle_{Q(\boldsymbol{\pi})}\right\}$, $\exp\left\{\left\langle \log a_{ij} \right\rangle_{Q(\boldsymbol{a}_i)}\right\}$, and $\exp\left\{\left\langle \log \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1}) \right\rangle_{Q(\boldsymbol{\mu}_i, \boldsymbol{S}_i)}\right\}$ are respectively used as the initial state probability of state i, the state transition probability from state i to state j, and the output probability of state i. Therefore, Eqs. (17) and (18) can be computed efficiently by the Forward-Backward algorithm.

2.3 Prior Distribution

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

$$P(\boldsymbol{\pi}) = \mathcal{D}(\{\pi_i\}_{i=1}^N \mid \{\phi_i\}_{i=1}^N), \tag{20}$$

$$P(\boldsymbol{a}_i) = \mathcal{D}(\{a_{ij}\}_{j=1}^N \mid \{\alpha_{ij}\}_{j=1}^N), \tag{21}$$

$$P(\boldsymbol{\mu}_i, \boldsymbol{S}_i) = \mathcal{N}(\boldsymbol{\mu}_i \mid \boldsymbol{\nu}_i, (\xi_i \boldsymbol{S}_i)^{-1}) \mathcal{W}(\boldsymbol{S}_i \mid \eta_i, \boldsymbol{B}_i) \tag{22}$$

where $\mathcal{D}(\cdot)$ is a Dirichlet distribution, and $\mathcal{N}(\cdot)\mathcal{W}(\cdot)$ is a Gauss-Wishart distribution. Moreover, $\{\phi_i, \alpha_{ij}, \xi_i, \eta_i, \boldsymbol{\nu}_i, \boldsymbol{B}_i\}_{i,j=1}^N$ is a set of hyper-parameters. When these conjugate prior distributions are used, the posterior distributions are represented by the same set of parameters $\left\{\bar{\phi}_i, \bar{\alpha}_{ij}, \bar{\xi}_i, \bar{\eta}_i, \bar{\boldsymbol{\nu}}_i, \bar{\boldsymbol{B}}_i\right\}_{i,j=1}^N$.

2.4 Update of Posterior Distribution

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

$$\bar{\phi}_i = \phi_i + \langle z_1^i \rangle \tag{23}$$

$$\bar{\alpha}_{ij} = \alpha_{ij} + \bar{T}_{ij} \tag{24}$$

$$\bar{\xi}_i = \xi_i + \bar{T}_i \tag{25}$$

$$\bar{\eta}_i = \eta_i + \bar{T}_i \tag{26}$$

$$\bar{\boldsymbol{\nu}}_i = \frac{\bar{T}_i \bar{\boldsymbol{o}}_i + \xi_i \boldsymbol{\nu}_i}{\bar{T}_i + \xi_i} \tag{27}$$

$$\bar{\boldsymbol{B}}_i = \bar{T}_i \bar{\boldsymbol{C}}_i + \boldsymbol{B}_i + \frac{\bar{T}_i \xi_i}{\bar{T}_i + \xi_i} (\bar{\boldsymbol{o}}_i - \boldsymbol{\nu}_i)(\bar{\boldsymbol{o}}_i - \boldsymbol{\nu}_i)^\top \tag{28}$$

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

$$\bar{T}_i = \sum_{t=1}^{T} \langle z_t^i \rangle \tag{29}$$

$$\bar{T}_{ij} = \sum_{t=1}^{T-1} \langle z_t^i z_{t+1}^j \rangle \tag{30}$$

$$\bar{\boldsymbol{o}}_i = \frac{1}{\bar{T}_i} \sum_{t=1}^{T} \langle z_t^i \rangle \boldsymbol{o}_t \tag{31}$$

$$\bar{\boldsymbol{C}}_i = \frac{1}{\bar{T}_i} \sum_{t=1}^{T} \langle z_t^i \rangle (\boldsymbol{o}_t - \bar{\boldsymbol{o}}_i)(\boldsymbol{o}_t - \bar{\boldsymbol{o}}_i)^\top \tag{32}$$

These optimizations can be performed effectively by iterative calculations as the EM algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence.

2.5 Speech Recognition Based on Bayesian Approach

In the speech recognition based on the Bayesian approach, the test data $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{\hat{T}})$ are recognized by using the predictive distribution as follows.

$$\begin{aligned} \hat{\boldsymbol{H}} &= \arg\max_{\boldsymbol{H}} P(\boldsymbol{H} \mid \boldsymbol{X}, \boldsymbol{O}) \\ &= \arg\max_{\boldsymbol{H}} P(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H}) P(\boldsymbol{H}) \end{aligned} \tag{33}$$

where $\boldsymbol{H}$ is a hypothesis of a phoneme sequence. The acoustic likelihood $P(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H})$ can be approximated by the variational Bayesian method as model training described in Sect. 2.2.

$$\begin{aligned}
&\log P(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H}) \\
&\quad = \log \sum_{\hat{\boldsymbol{Z}}} \int P(\boldsymbol{X}, \hat{\boldsymbol{Z}} \mid \boldsymbol{\Lambda}, \boldsymbol{H}) P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) d\boldsymbol{\Lambda} \\
&\quad \geq \sum_{\hat{\boldsymbol{Z}}} \int \hat{Q}(\hat{\boldsymbol{Z}}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{X}, \hat{\boldsymbol{Z}} \mid \boldsymbol{\Lambda}, \boldsymbol{H}) P(\boldsymbol{\Lambda} \mid \boldsymbol{O})}{\hat{Q}(\hat{\boldsymbol{Z}}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\
&\quad = \hat{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H}) \qquad (34)
\end{aligned}$$

where $\hat{\boldsymbol{Z}} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{\hat{T}})$ is a sequence of HMM states for the test data $\boldsymbol{X}$, and $\hat{Q}(\hat{\boldsymbol{Z}}, \boldsymbol{\Lambda})$ is the VB posterior distribution which approximates the true posterior distribution $P(\hat{\boldsymbol{Z}}, \boldsymbol{\Lambda} \mid \boldsymbol{X})$. In the recognition process, the VB posterior distribution of model parameters $Q(\boldsymbol{\Lambda})$ estimated in the training part is used instead of $P(\boldsymbol{\Lambda} \mid \boldsymbol{O})$, and the same assumption as Eq. (8) is used. Moreover, it is assumed that the amount of test data is much smaller than the one of training data in this paper. Then, the VB posterior distribution $\hat{Q}(\boldsymbol{\Lambda})$ is approximated by $Q(\boldsymbol{\Lambda})$. Therefore, the lower bound $\hat{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H})$ is calculated by using $Q(\boldsymbol{\Lambda})$.

$$\begin{aligned}
&\hat{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H}) \\
&\quad = \log \sum_{\hat{\boldsymbol{Z}}} \Bigg\{ \prod_{i=1}^{N} \exp\left\{ \hat{z}_1^i \left\langle \log \pi_i \right\rangle_{Q(\boldsymbol{\pi})} \right\} \\
&\qquad \times \prod_{t=1}^{\hat{T}-1} \prod_{i=1}^{N} \prod_{j=1}^{N} \exp\left\{ \hat{z}_t^i \hat{z}_{t+1}^j \left\langle \log a_{ij} \right\rangle_{Q(\boldsymbol{a}_i)} \right\} \\
&\qquad \times \prod_{t=1}^{\hat{T}} \prod_{i=1}^{N} \exp\left\{ \hat{z}_t^i \left\langle \log \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1}) \right\rangle_{Q(\boldsymbol{\mu}_i, \boldsymbol{S}_i)} \right\} \Bigg\}
\end{aligned} \qquad (35)$$

Then, $\hat{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H})$ is similar to the likelihood function of an HMM as Eq. (19). Although the accurate $\hat{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{H})$ is computed by considering all possible sequences of HMM states $\hat{\boldsymbol{Z}}$ as the training part, the Viterbi algorithm is applied in decoding as the ML approach.

3. Bayesian Context Clustering Using Cross Validation

3.1 Bayesian Context Clustering

The decision tree based context clustering is a top-down clustering method to optimize the state tying structure for robust model parameter estimation [11]. A leaf node of the decision tree corresponds to a set of HMM states to be tied. The decision tree growing process begins with a root node that may have all HMM states, or all states associated with a particular phone, etc. Then, a question which divides the set of states into two subsets assigned respectively to two child nodes, "Yes" node and "No" node as illustrated in Fig. 1, is chosen so that the corresponding new HMM has the largest value of an objective function for training data. The decision tree is grown in a greedy fashion, successively splitting nodes by selecting the pair of a question and node that maximizes the gain of the objective function at each step.

In the Bayesian approach, an optimal model structure

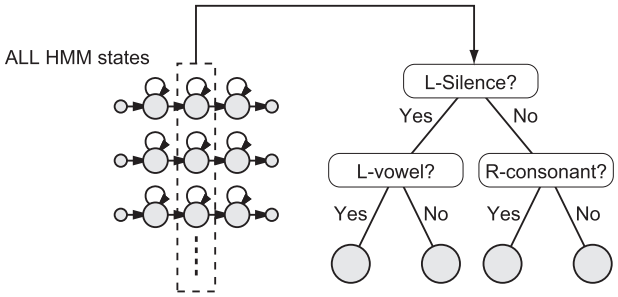


Fig. 1 Overview of decision tree based context clustering.

can be selected by maximizing the objective function $\mathcal{F}$. When a node is split into two nodes by the question q, the gain $\Delta\mathcal{F}_q$ is defined as the difference of $\mathcal{F}$ before and after splitting.

$$\Delta\mathcal{F}_q = \mathcal{F}_q^y + \mathcal{F}_q^n - \mathcal{F}_q^p \qquad (36)$$

where $\mathcal{F}_q^y$ and $\mathcal{F}_q^n$ are the value of objective function $\mathcal{F}$ of split nodes by a question q, and $\mathcal{F}_q^p$ is the value before a splitting. The question $\hat{q}$ for splitting a node is chosen from the question set as follows.

$$\hat{q} = \arg\max_q \Delta\mathcal{F}_q \qquad (37)$$

By splitting nodes until $\Delta\mathcal{F}_{\hat{q}} \leq 0$, the decision tree that maximizes the objective function $\mathcal{F}$ is obtained.

In the decision tree based context clustering, it is typically assumed that the state occupancies are not changed by the split nodes. Then, the objective function $\mathcal{F}$ can be computed as follows.

$$\begin{aligned}
\mathcal{F} &= -\log C_{\boldsymbol{\Lambda}} - \left\langle \log Q(\mathbf{Z}) \right\rangle_{Q(\mathbf{Z})} \\
&= -\sum_{i=1}^{N} \log C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i} + \text{Const} \qquad (38)
\end{aligned}$$

From Eq. (38), the gain of the objective function $\Delta\mathcal{F}_q$ can be computed by the normalization term of the posterior distribution $C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i}$. The normalization term $C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i}$ is defined as follows.

$$\log C_{\boldsymbol{\mu}_i, \boldsymbol{S}_i} = \log \frac{\bar{C}_{\mathcal{N}_i} \bar{C}_{\mathcal{W}_i}}{C_{\mathcal{N}_i} C_{\mathcal{W}_i}} (2\pi)^{\frac{\bar{\eta} D}{2}} \qquad (39)$$

where $C_{\mathcal{N}_i}$ and $C_{\mathcal{W}_i}$ denote the normalization terms of prior Gauss-Wishart distribution.

$$C_{\mathcal{N}_i} = (2\pi)^{-\frac{D}{2}} \xi_i^{\frac{D}{2}} \qquad (40)$$

$$C_{\mathcal{W}_i} = \frac{|B_i|^{\frac{\eta_i}{2}}}{2^{\frac{\eta_i D}{2}} \pi^{\frac{D(D-1)}{4}} \prod_{j=1}^{D} \Gamma(\frac{\eta+1-j}{2})} \qquad (41)$$

where $\Gamma(\cdot)$ is the Gamma function. The normalization terms of posterior Gauss-Wishart distribution are also denoted by $\bar{C}_{\mathcal{N}_i}$ and $\bar{C}_{\mathcal{W}_i}$, and they are represented by using posterior hyper-parameters $\bar{\xi}_i$, $\bar{\eta}_i$, and $\bar{\boldsymbol{B}}_i$ instead of prior hyper-parameters ξ_i, η_i, and $\boldsymbol{B}_i$ in Eqs. (40) and (41), respectively.

The posterior hyper-parameters $\bar{\xi}_i$, $\bar{\eta}_i$, and $\bar{\boldsymbol{B}}_i$ can be calculated by using equations described in Sect. 2.4. From Eqs. (38)–(41), $\mathcal{F}$ can be computed by using the prior and posterior hyper-parameters. Since it is assumed that the state occupancies are not changed in the context clustering and the posterior hyper-parameters can be represented by using sufficient statistics and the prior hyper-parameters, the prior hyper-parameters are important parameters for the Bayesian context clustering.

If we have prior data $\tilde{\boldsymbol{O}}$ which is obtained from similar conditions (e.g., speaker, domain, recording condition) as the training data, the prior distribution can be constructed as $P(\boldsymbol{\Lambda}) = P(\boldsymbol{\Lambda} \mid \tilde{\boldsymbol{O}})$. When the prior data is given, the prior distribution is obtained by using the same approximation techniques as the variational Bayesian method described in Sect. 2.2.

$$\begin{aligned} P(\boldsymbol{\Lambda} \mid \tilde{\boldsymbol{O}}) &\approx \tilde{Q}(\boldsymbol{\Lambda}) \\ &= \tilde{C}_{\boldsymbol{\Lambda}} \tilde{P}(\boldsymbol{\Lambda}) \exp\left\{\left\langle \log P(\tilde{\boldsymbol{O}}, \tilde{\boldsymbol{Z}} \mid \boldsymbol{\Lambda})\right\rangle_{\tilde{Q}(\tilde{\boldsymbol{Z}})}\right\} \end{aligned} \quad (42)$$

where $\tilde{\boldsymbol{Z}}$ is a sequence of latent variables, and $\tilde{Q}(\tilde{\boldsymbol{Z}})$ is an approximate distribution of $P(\tilde{\boldsymbol{Z}} \mid \tilde{\boldsymbol{O}}, \boldsymbol{\Lambda})$. Although Eq. (42) still includes prior of prior distribution $\tilde{P}(\boldsymbol{\Lambda})$, we assumed that the prior of prior distribution $\tilde{P}(\boldsymbol{\Lambda})$ is a uniform distribution before the prior data is given. Then, prior distribution $P(\boldsymbol{\Lambda} \mid \tilde{\boldsymbol{O}})$ can be obtained as follows.

$$\begin{aligned} P(\boldsymbol{\Lambda} \mid \tilde{\boldsymbol{O}}) &\approx \tilde{C}_{\boldsymbol{\Lambda}} \exp\left\{\left\langle \log P(\tilde{\boldsymbol{O}}, \tilde{\boldsymbol{Z}} \mid \boldsymbol{\Lambda})\right\rangle_{\tilde{Q}(\tilde{\boldsymbol{Z}})}\right\} \\ &= \mathcal{D}(\{\pi_i\}_{i=1}^N \mid \{\tilde{T}_{0i} + 1\}_{i=1}^N) \\ &\quad \times \prod_{i=1}^N \mathcal{D}(\{a_{ij}\}_{j=1}^N \mid \{\tilde{T}_{ij} + 1\}_{j=1}^N) \\ &\quad \times \prod_{i=1}^N \{\mathcal{N}(\boldsymbol{\mu}_i \mid \tilde{\boldsymbol{o}}_i, (\tilde{T}_i \boldsymbol{S}_i)^{-1}) \\ &\quad \times \mathcal{W}(\boldsymbol{S}_i \mid \tilde{T}_i + D, (\tilde{T}_i \tilde{\boldsymbol{C}}_i))\} \end{aligned} \quad (43)$$

The distribution $\tilde{Q}(\tilde{\boldsymbol{Z}})$ can be estimated via the EM algorithm using prior data $\tilde{\boldsymbol{O}}$. Statistics $\tilde{T}_{0i}$, $\tilde{T}_{ij}$ and $\tilde{T}_i$ denote the occupancy probabilities of initial state i, state transition from i to j, and state i with respect to the prior data, respectively. Moreover, $\tilde{\boldsymbol{o}}_i$ and $\tilde{\boldsymbol{C}}_i$ denote the mean vector and the covariance matrix of prior data in the i-th state, respectively. Thus, the prior distribution can be determined by sufficient statistics of the prior data. However, prior distributions are heuristically determined in many cases, because the prior data is not usually given in HMM-based speech recognition. Hyper-parameters affect the model selection as tuning parameters, therefore a determination technique of prior distributions is required to automatically select an appropriate model structure. One possible approach is to optimize the hyper-parameters so as to maximize the marginal likelihood of training data, as like the empirical Bayesian method [14]. However, it still needs tuning parameters which control influences of prior distributions, and often leads to the overfitting problem as the ML criterion. In this paper, we propose the prior distribution determination technique using cross validation and apply it to the context clustering.

3.2 Bayesian Approach Using Cross Validation

The cross validation method is a popular strategy for model selection [15], [16]. The main idea behind cross validation is to split data for estimating the risk of each model. Part of data is used for training each model, and the remaining part is used for estimating the risk of the model. Then, the cross validation method selects the model with the smallest estimated risk. The basic form of cross validation is K-fold cross validation. In the K-fold cross validation method, the training data is randomly divided into K different groups. Then, a model is trained using $K-1$ groups of data, and the objective function is computed for the group excluded in the training. This process is repeated for K times with different combinations of $K-1$ groups. The value of objective function is accumulated and the accumulated value is used for evaluation of model structure.

In the Bayesian approach using K-fold cross validation, the training data $\boldsymbol{O}$ is divided at random into K subsets of training data $\left\{\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \ldots, \boldsymbol{O}^{(K)}\right\}$. For the k-th evaluation, $\boldsymbol{O}^{(\bar{k})} = \left\{\boldsymbol{O}^{(j)} \mid j \neq k\right\}$ and $\boldsymbol{O}^{(k)}$ are respectively used for the determination of prior distributions and the estimation of posterior distributions, i.e., $\boldsymbol{O}^{(\bar{k})}$ and $\boldsymbol{O}^{(k)}$ are used as prior data and training data. The Bayesian approach using cross validation considers the log marginal likelihood $\log P(\boldsymbol{O}^{(k)} \mid \boldsymbol{O}^{(\bar{k})})$. Using Jensen's inequality, the lower bound of log marginal likelihood $\mathcal{F}^{(k)}$ is defined as Eq. (6).

$$\log P(\boldsymbol{O}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \geq \mathcal{F}^{(k)} \quad (44)$$

The optimal VB posterior distributions of model parameters are given by maximizing $\mathcal{F}^{(k)}$ with the variational method as Eq. (9).

$$\begin{aligned} Q(\boldsymbol{\Lambda}^{(k)}) &= C_{\boldsymbol{\Lambda}^{(k)}} P(\boldsymbol{\Lambda}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \\ &\quad \times \exp\left\{\left\langle \log P(\boldsymbol{O}^{(k)}, \boldsymbol{Z}^{(k)} \mid \boldsymbol{\Lambda}^{(k)})\right\rangle_{Q(\boldsymbol{Z}^{(k)})}\right\} \end{aligned} \quad (45)$$

where $C_{\boldsymbol{\Lambda}^{(k)}}$ is a normalization term of $Q(\boldsymbol{\Lambda}^{(k)})$ and $P(\boldsymbol{\Lambda}^{(k)} \mid \boldsymbol{O}^{(\bar{k})})$ is a prior distribution of the k-th cross validation which represents prior data $\boldsymbol{O}^{(\bar{k})}$. Figure 2 is an overview of the Bayesian approach using cross validation.

The cross validation method can select robust model structures because the objective value is calculated by evaluating open data. The Bayesian approach obtains robust predictive distributions and selects robust model structures while taking account of the amount of training data because posterior distributions of model parameters are used. Consequently, the Bayesian approach using cross validation can select model structures while taking account of the uncertainty of the data variables and model parameters, and the robustness can be improved from the standard Bayesian approach.

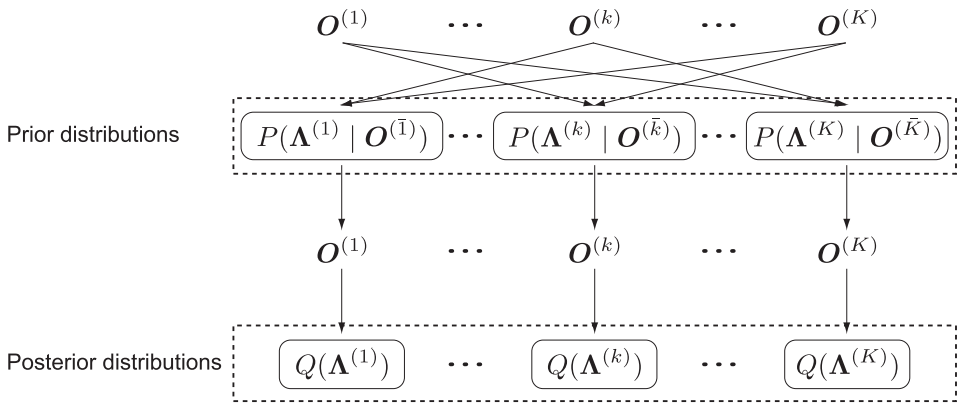


Fig. 2 Overview of Bayesian approach using cross validation.

3.3 Bayesian Context Clustering Using Cross Validation

The objective function $\mathcal{F}^{(\mathrm{CV})}$ is used in the Bayesian context clustering using cross validation. It is obtained by summing $\mathcal{F}^{(k)}$ for each fold.

$$\mathcal{F}^{(\mathrm{CV})} = \sum_{k=1}^{K} \mathcal{F}^{(k)} \tag{46}$$

In the proposed method, an optimal model structure can be selected by maximizing the objective function $\mathcal{F}^{(\mathrm{CV})}$. The question $\tilde{q}$ for splitting a node is chosen from the question set as Eq. (37).

$$\tilde{q} = \arg\max_{q} \Delta\mathcal{F}_{q}^{(\mathrm{CV})} \tag{47}$$

where $\Delta\mathcal{F}_{q}^{(\mathrm{CV})}$ is the gain in the value of the objective function $\mathcal{F}^{(\mathrm{CV})}$ when a node is split by the question q. The gain $\Delta\mathcal{F}_{q}^{(\mathrm{CV})}$ is obtained by

$$\Delta\mathcal{F}_{q}^{(\mathrm{CV})} = \mathcal{F}_{q}^{(\mathrm{CV})^{y}} + \mathcal{F}_{q}^{(\mathrm{CV})^{n}} - \mathcal{F}_{q}^{(\mathrm{CV})^{p}} \tag{48}$$

By splitting nodes until $\Delta\mathcal{F}_{\tilde{q}}^{(\mathrm{CV})} \leq 0$, the decision tree that maximizes the objective function $\mathcal{F}^{(\mathrm{CV})}$ is obtained.

The prior distribution of the k-th cross validation $P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} \mid \boldsymbol{O}^{(\bar{k})})$ is obtained from Eq. (43).

$$\begin{aligned} &P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \\ &\quad = \mathcal{N}(\boldsymbol{\mu}^{(k)} \mid \bar{\boldsymbol{o}}^{(\bar{k})}, (\bar{T}^{(\bar{k})}\boldsymbol{S}^{(k)})^{-1}) \\ &\qquad \times \mathcal{W}(\boldsymbol{S}^{(k)} \mid \bar{T}^{(\bar{k})} + D, (\bar{T}^{(\bar{k})}\bar{\boldsymbol{C}}^{(\bar{k})})) \end{aligned} \tag{49}$$

where $\bar{T}^{(\bar{k})}$, $\bar{\boldsymbol{o}}^{(\bar{k})}$ and $\bar{\boldsymbol{C}}^{(\bar{k})}$ respectively denote the occupancy probability, the mean vector and the covariance matrix of a subset of training data $\boldsymbol{O}^{(\bar{k})}$. These parameters are efficiently computed in context clustering because it is assumed that the state occupancies are not changed by splitting nodes. Moreover, the posterior distributions $Q(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)})$ can be estimated by Eqs. (23)–(28). Here, since the assumption the state occupancies are not changed by splitting nodes are used, the posterior distributions of all folds are represented by the same parameters. Therefore, although the Bayesian approach using cross validation increases the computational cost, the prior and posterior distributions are efficiently calculated in context clustering.

4. Experiments

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

4.1 Experimental Conditions

The 20,000 and 1,000 Japanese sentences uttered by male speakers from Japanese Newspaper Article Sentences (JNAS) [22] were used for model training. The 100 Japanese sentences uttered by male speakers, which were not included in the training data, from JNAS were used for evaluation. The average lengths of the training 20,000 utterances, training 1,000 utterances and test 100 utterances were 6.16 seconds, 6.42 seconds, and 5.83 seconds, respectively. Speech signals were sampled at a rate of 16 kHz and widowed at a 10 ms frame rate using a 25 ms Hamming window. The feature vectors consisted of the 0th through 13th mel-frequency cepstral coefficients (MFCCs), their delta and delta-delta coefficients. A three-state, left-to-right and no skip structure HMMs were used as triphone HMMs, and 204 questions were prepared in decision tree context clustering. In these experiments, we used a phoneme network imposing the constraints of Japanese phoneme transitions. However, phoneme N-gram probabilities and the language model weight were not used. The insertion penalty was adjusted for each experiment so that the number of insertion and deletion errors become almost equal. The experimental conditions are summarized in Table 1.

In recent HMM-based speech recognition systems, a multi-mixture Gaussian is typically used as a state output probability distribution. Although the VB method has been applied to multi-mixture HMMs [8], to evaluate the effect of only the proposed context clustering algorithm, each state output probability distribution was assumed to be modeled

Table 1 Experimental conditions.

Training data	JNAS 20,000 utterances JNAS 1,000 utterances
Test data	JNAS 100 utterances
Sampling rate	16 kHz
Feature vector	13-order MFCC + ΔMFCC + ΔΔMFCC
Window	Hamming
Frame size	25 ms
Frame shift	10 ms
Number of HMM states	3 (left-to-right)
Number of phoneme categories	43

Table 2 K-fold cross validation (20,000 utterances).

	K				
	5	10	20	100	200
Number of states	14,072	14,360	14,474	14,575	14,610
Phoneme accuracy (%)	80.4	80.3	80.3	80.3	80.4

Table 3 K-fold cross validation (1,000 utterances).

	K				
	5	10	20	100	200
Number of states	3,919	4,065	4,101	4,141	4,156
Phoneme accuracy (%)	78.7	78.7	79.4	78.9	79.0

by a single Gaussian distribution with a diagonal covariance matrix in these experiments. Then, since the likelihood of each dimension is computed independently, the Gauss-Wishart distribution is equal to the Gauss-Gamma distribution.

4.2 Number of Folds in Cross Validation

In these experiments, the several number of folds in Bayesian context clustering using cross validation were compared. Tables 2 and 3 show the number of states and phoneme accuracies with the acoustic models trained by 20,000 and 1,000 utterances, respectively, when the number of folds for cross validation were varied. As the number of folds increased, the computational cost was also proportionally increased and the resultant model structure became stable. Results show that the phoneme accuracy did not improve much with acoustic models trained by 20,000 utterances when the number of K was changed. However, in 1,000 utterances training condition, the phoneme accuracies were not stable. So, the large number of folds are required when the training data is small.

4.3 Comparison of Conventional Approaches

In these experiments, the following three approaches were compared.

- **ML-MDL**: Acoustic models were trained by the ML criterion and model structures were selected by the MDL criterion.
- **ML-CVML**: Acoustic models were trained by the ML criterion and model structures were selected by cross validation with the ML criterion.
- **Bayes-CVBayes**: Acoustic models were trained by the Bayesian criterion and model structures were selected by cross validation with the Bayesian criterion.

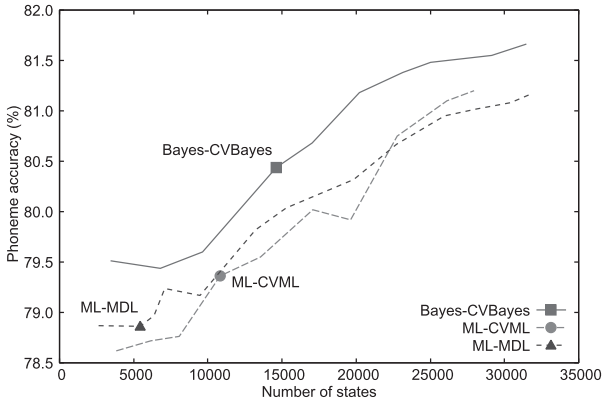


Fig. 3 Phoneme accuracies of **ML-MDL**, **ML-CVML** and **Bayes-CVBayes** trained by 20,000 utterances versus the number of states.

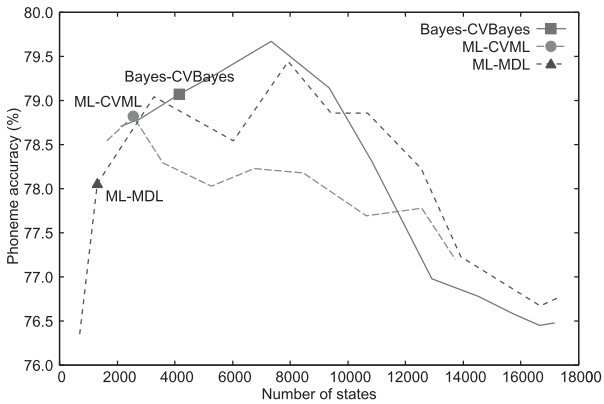


Fig. 4 Phoneme accuracies of **ML-MDL**, **ML-CVML** and **Bayes-CVBayes** trained by 1,000 utterances versus the number of states.

Figures 3 and 4 show the phoneme accuracies of acoustic models trained by 20,000 and 1,000 utterances, respectively. For **ML-CVML** and **Bayes-CVBayes**, 200-fold cross validation was used. To evaluate the performance of model selection, the phoneme accuracies with varying the size of decision trees are also shown. The decision trees were generated by changing a threshold of the stopping criterion $\Delta\mathcal{F} \leq \mathit{threshold}$ in the context clustering. In these figures, the lines represent the phoneme accuracies for each model structure and the points represent the phoneme accuracies of the model structure selected automatically by each method. These figures show that the proposed method **Bayes-CVBayes** selected the largest model structure, and the conventional method **ML-MDL** selected the smallest model structure in both training conditions. The model structure selected by **Bayes-CVBayes** was closer to that performed the highest accuracy than **ML-MDL**. Consequently, the proposed method **Bayes-CVBayes** outper-

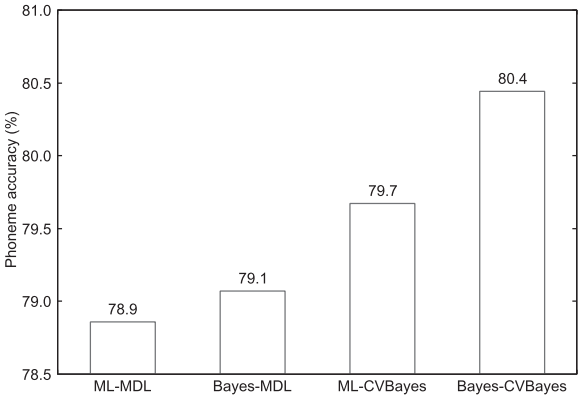


Fig. 5 Phoneme accuracies when the acoustic models were trained by 20,000 utterances with the swapped decision tree.

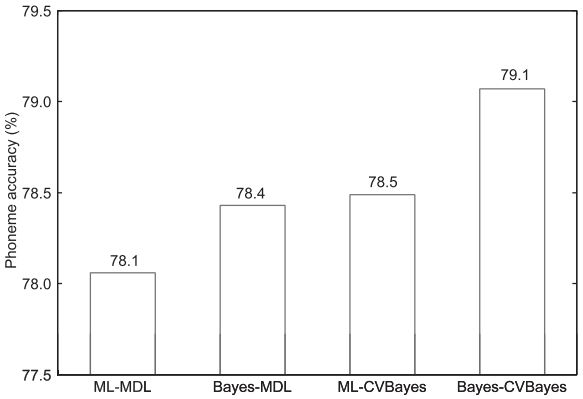


Fig. 6 Phoneme accuracies when the acoustic models were trained by 1,000 utterances with the swapped decision tree.

forms the conventional method,**ML-MDL** and **ML-CVML**. In Fig. 3, **Bayes-CVBayes** achieved a 8.08% relative error reductions over **ML-MDL**.

It can be considered that the improvement of the proposed method caused by two factors, marginalization of model parameters and model selection. To discuss the impact of these two factors, an additional experiment was performed by swapping the model structures of **ML-MDL** and **Bayes-CVBayes**. The following two approaches were compared to **ML-MDL** and **Bayes-CVBayes**.

- **Bayes-MDL**: Acoustic models were trained by the Bayesian criterion and model structures selected by **ML-MDL** were used.
- **ML-CVBayes**: Acoustic models were trained by ML criterion and model structures selected by **Bayes-CVBayes** were used.

Figures 5 and 6 show the phoneme accuracies of acoustic models trained by 20,000 and 1,000 utterances, respectively. Although the difference between **Bayes-MDL** and **ML-MDL** was the marginalization by the Bayesian approach, the phoneme accuracies of **Bayes-MDL** were improved from **ML-MDL** on both training conditions. Furthermore, the phoneme accuracies of **ML-CVBayes** were

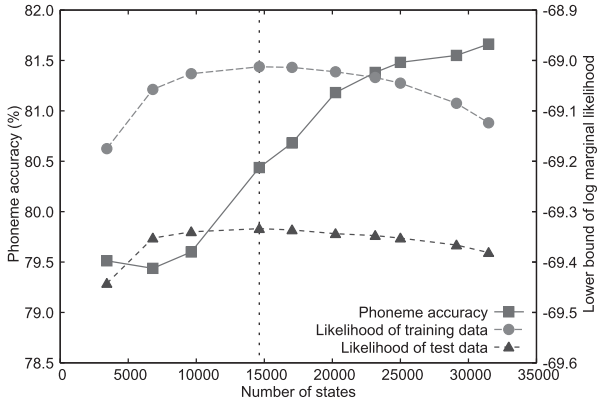


Fig. 7 Log marginal likelihoods on both training and test data versus the number of states when the acoustic models were trained by 20,000 utterances.

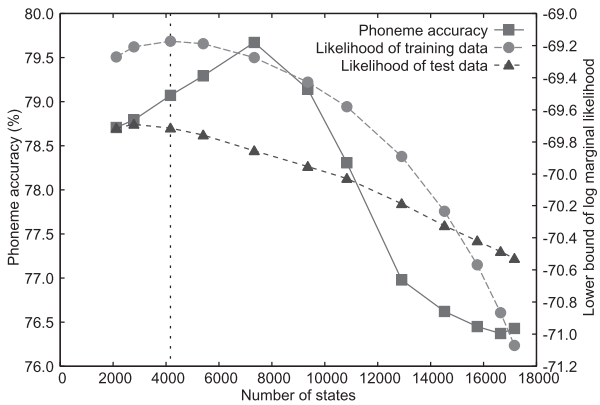


Fig. 8 Log marginal likelihoods on both training and test data versus the number of states when the acoustic models were trained by 1,000 utterances.

also improved when compared with **ML-MDL** on both training conditions, due to the model selection based on the Bayesian criterion with cross validations. Therefore, these results clearly showed that the Bayesian approach was effective for both the model training and the model selection. However, **Bayes-MDL** and **ML-CVBayes** were worse than **Bayes-CVBayes**. This means that training criterion and model selection were strongly related, and these should be consistently performed based on the Bayesian criterion.

4.4 Marginal Likelihood of the Training and Test Data

Figures 7 and 8 show the relation among the lower bound $\mathcal{F}^{(\mathrm{CV})}$ for the training data, $\mathcal{F}$ for the test data with the correct phoneme sequences, and the phoneme accuracies. In these figures, a similar tendency between $\mathcal{F}^{(\mathrm{CV})}$ and $\mathcal{F}$ was observed, and the model structure which gave the highest $\mathcal{F}^{(\mathrm{CV})}$ also achieved the highest $\mathcal{F}$. However, the phoneme accuracy was not proportional to $\mathcal{F}$, and the proposed method could not select the model structure which achieved the highest phoneme accuracy. This means that although the proposed method could select the model struc-

ture which can accurately predict acoustic features for each HMM state, it is not identical to the performance in the classification problem. This is because the likelihood of incorrect phoneme sequences including insertion and deletion errors were not considered in the proposed method. This result suggests that a Bayesian criterion which can represent the classification performance directly is required.

5. Conclusion

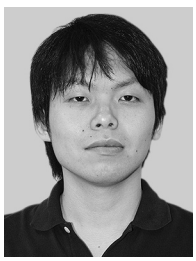
This paper proposed the Bayesian context clustering using cross validation for speech recognition based on the variational Bayesian framework. In the proposed method, the prior distributions are determined by using cross validation, and the determined prior distribution is applied to the context clustering. The results on continuous phoneme recognition experiments demonstrated that the proposed method outperformed the context clustering based on the MDL criterion and cross validation with ML estimates. The proposed method could determine prior distributions without any tuning parameters, and select the model structure which can accurately predict acoustic features for each HMM state. As future work, we will apply a Bayesian criterion using cross validation for selecting the number of mixtures, and apply a Bayesian criterion which represents the classification performance directly to the context clustering.

Acknowledgements

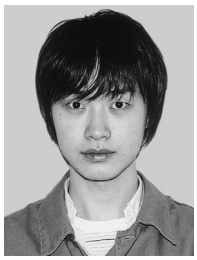
The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org). A part of this research was supported by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists 22-10062.

References

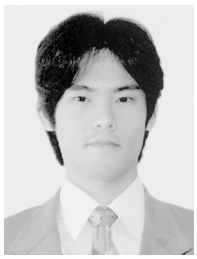
[1] B.H. Juang and L.R. Rabiner, "Hidden Markov models for speech recognition," Technometrics, vol.33, no.3, pp.251–272, 1991.
[2] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, Bayesian data analysis, Chapman & Hall, 1995.
[3] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on a Bayesian prediction approach," IEEE Trans. Speech Audio Process., vol.7, no.4, pp.426–440, 1999.
[4] Q. Huo and C.H. Lee, "A Bayesian predictive classification approach to robust speech recognition," IEEE Trans. Speech Audio Process., vol.8, no.2, pp.200–204, 2000.
[5] S. Watanabe and A. Nakamura, "Effects of Bayesian predictive classification using variational Bayesian posteriors for sparse training data in speech recognition," Proc. Interspeech, pp.1105–1108, 2005.
[6] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. UAI 15, 1999.
[7] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. Speech Audio Process., vol.12, no.4, pp.365–381, 2004.
[8] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering," Proc. ICASSP 2004, vol.1, pp.813–816, 2004.
[9] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.38, no.4, pp.599–609, 1990.
[10] J.J. Odell, The use of context in large vocabulary speech recognition, PhD dissertation, Cambridge University, 1995.
[11] S. Young, J.J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," Proc. ARPA Workshop on Human Language Technology, pp.307–312, 1994.
[12] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proc. Eurospeech, pp.99–102, 1997.
[13] G. Schwarz, "Estimating the dimension of a model," The Annals of Staistics, vol.6, no.2, pp.461–464, 1978.
[14] H. Robbins, "An empirical Bayes approach to statistics," Proc. 3rd Berkeley Symp. Math. Statist. Probab., vol.1, pp.157–163, 1956.
[15] R. Kohavi, "A study of cross-valication and bootstrap for accuracy estimation and model selection," Proc. International Joint Conference on AI, pp.1137–1145, 1995.
[16] S. Arlot and A. Celisse, "A survey of cross-valication procedures for model selection," Statistics Surveys, vol.4, pp.40–79, 2010.
[17] T. Shinozaki, "HMM state clustering based on efficient cross-validation," Proc. ICASSP, vol.1, pp.1157–1160, 2006.
[18] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for HMM-based speech recognition," Proc. Interspeech, pp.936–939, 2008.
[19] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," J. Royal Statist. Coc. Ser. B (methodological), vol.39, pp.1–38, 1977.
[20] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[21] S. Kullback and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol.22, pp.79–86, 1951.
[22] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, 1999.



Kei Hashimoto was born in Saitama, Japan, on March 1, 1984. He received the B.E. and M.E. degrees in computer science, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan in 2006 and 2008, respectively. He is currently a Doctor's candidate at Nagoya Institute of technology. From October 2008 to January 2009, he was an intern researcher at National Institute of Information and Communications Technology (NICT), Kyoto, Japan. From April 2010, he was a Research Fellow of the Japan Society for the Promotion of Science (JSPS) in the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan. His research interests include speech recognition, speech synthesis and machine translation. He is a student member of the Acoustical Society of Japan.



Heiga Zen was born in Osaka, Japan, on March 4, 1979. He received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr.Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively. During 2003, he was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an intern/co-op researcher in the Human Language Technology group at IBM T.J. Watson Research Center, Yorktown Heights, NY. He is currently a postdoctoral fellow at Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis. He received the Awaya and Itakura Awards from the Acoustical Society of Japan (ASJ) in 2006 and 2008, respectively, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2008. He is a member of ASJ and ISCA.



Yoshihiko Nankaku received the B.E. degree in Computer Science, and the M.E. and Ph.D. degrees in the Department of Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004 respectively. After a year as a postdoctoral fellow at the Nagoya Institute of Technology, he is currently an Assistant Professor at the same Institute. His research interests include statistical machine learning, speech recognition, speech synthesis, image recognition, and multi-modal interface. He is a member of the Acoustical Society of Japan (ASJ).

Akinobu Lee was born in Kyoto, Japan, on December 19, 1972. He received the B.E. and M.E. degrees in information science, and the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 1996, 1998, and 2000, respectively. During 2000–2005, he was an Assistant Professor in Nara Institute of Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan. He is now an Associate Professor of Nagoya Institute of Technology. His research interests include speech recognition and spoken language understanding. He is a member of the IEEE, ISCA, IPSJ and ASJ.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He published over 60 journal papers and over 150 conference papers, and received 5 paper awards. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 2000 to 2003. Currently he is a member of ISCA Advisory Council and an associate editor of IEEE Transactions on Audio, Speech & Language Processing, and acts as organizer and reviewer for many major speech conferences, workshops and journals. His research interests include speech speech coding, speech synthesis and recognition, and statistical machine learning.