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

2004－SLP－54（10）
2004／12／20

ベイズ的音声認識VBECを用いたモデル構造自動構築法の多様な音声データに対する頑健性

渡部　晋治†　　中村　篤†

† 日本電信電話(株) NTT コミュニケーション科学基礎研究所　〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{watanabe,ats}@cslab.kecl.ntt.co.jp

あらまし　現在の音声認識システムが自然発話の認識・理解において十分な性能を示すことができない理由に頑健性の欠落が挙げられる．頑健性の欠落の一因としては，統計的モデル学習における，パラメータ推定に基づく最尤法の汎化能力の低さが考えられる．既存の隠れマルコフモデルのみならずそれを超える音響モデリング技術においても，統計的モデル学習は引き続き根幹技術の一つであると考えられ，その汎化能力を高めることは音声認識における普遍的課題といえる．事後確率分布推定にもとづくベイズ法は，モデルパラメータの周辺化操作による高い汎化能力ゆえに，最尤法に比べて頑健なモデル学習を可能にする．ベイズ的音声認識VBECは，変分ベイズ法を用いた事後確率分布推定にもとづくため，ベイズ法の長所である汎化能力の高い頑健な推定を実現する本格的なベイズ音声認識の枠組みである．また，VBECはモデル構造を確率変数とみなすことにより，モデル構造事後確率分布の事後確率最大化をもとにして，人手無しで音響モデルの自動構築を高い精度で実現できる．本稿では発話スタイル・使用言語の異なる学習・評価セット(孤立単語音声・読み上げ音声・講演音声・英語読み上げ音声)を用い，データによらずVBECの音響モデル自動構築が有効であることを示す．また，VBECで自動構築された音響モデルが評価データによらず十分な性能を示すことを先のJNASで作られた音響モデルを質問応答音声で認識することにより検証する．
キーワード　音声認識，VBEC，音響モデルトポロジーの自動決定，発話様式・言語・評価データに対しての頑健性

Robustness of acoustic model topology determined by VBEC for different speech data sets

Shinji WATANABE† and Atsushi NAKAMURA†

† Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
E-mail: †{watanabe,ats}@cslab.kecl.ntt.co.jp

Abstract　A lack of robustness with acoustic modeling often degrades the performance of spontaneous speech recognition and understanding. One reason for this shortcoming is that the Maximum Likelihood (ML) approach based on model parameter estimation has a poor generalization ability. This makes it important to improve the generalization ability of robust training of models including HMM and future techniques beyond HMM. The Bayesian approach is based on posterior distribution estimation, and has a better generalization ability than the ML approach due to the marginalization effect of model parameters. Variational Bayesian Estimation and Clustering for speech recognition (VBEC) is a total Bayesian framework in the sense that all speech recognition procedures are based on posterior distribution estimation within the Variational Bayes method, which includes the Bayesian advantage of highly generalized model training. In addition, a VBEC specification of the posterior distribution estimation enables automatic determination of an acoustic model topology without heuristics, by regarding model complexity as a probabilistic variable, and by selecting the appropriate model that scores the maximum probability value. In this paper, we describe experiments for different speaking-style (isolated word, continuous speech and spontaneous lecture speech) and language sets (Japanese and English) of training data, and show the effectiveness of VBEC, which automatically determines the model topology robustly according to the speech types of the training data. We also examine the robustness of the determined models for a mismatched condition between training and test data tasks.
Key words　Speech recognition, VBEC, Automatic determination of acoustic model topology, Robustness for speaking style, language and mismatched condition between training and test data.

1. Introduction

Speech recognition performance depends strongly on the preciseness of the acoustic modeling of speech. An acoustic model has a very complicated structure: a category is a set of clustered-state triphone Hidden Markov Models (HMMs) each of which possesses an output distribution described by a Gaussian Mixture Model (GMM). Although certain algorithms have been proposed for dealing with this complicated model structure (model topology) [1-3], they require heuristic tuning since they are based on the Maximum Likelihood (ML) criterion. That is to say, since the likelihood value increases monotonically as the number of model parameters increases, ML always leads to the selection of the model structure with the largest number of parameters, and this ap-

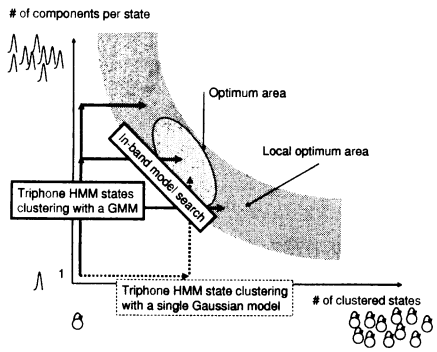

Figure 1 Optimum model search for an acoustic model.

proach cannot determine the model topology appropriately without using heuristic tuning. Therefore, only *experts* who well understand the acoustic model topology can design the models by setting the heuristic tuning, i.e., by setting the number of clustered states and the number of GMM components empirically. If we are to eliminate the need for heuristic tuning in ML for non-experts, we must find a way to determine the acoustic model topology automatically.

Recently, we and others have proposed a new framework for the automatic determination of the acoustic model topology, namely Variational Bayesian Estimation and Clustering for speech recognition (VBEC) [4, 5]. VBEC is a total Bayesian framework in the sense that all its training and classification procedures are based on approximated posteriors by using Variational Bayes (VB) [6–8]. VBEC can theoretically determine a complicated model topology by selecting the best VB objective function model, which corresponds to selecting the maximum probability of the VB posterior for the model complexity, even when latent variables are included. The VB method is a powerful algorithm for practical posterior computation and has been applied to other forms of speech processing [9,10]. Additionally, we achieved the optimum topology area without falling into the local optimum area using VBEC based on an efficient model search algorithm and GMM-based decision tree clustering utilizing the acoustic model characteristics [11]. However, in [11] the proposed automatic determination method was only examined for such a small task as isolated word recognition, and its effectiveness should be examined by recognizing various types of speech data sets. In this paper, we describe experiments for different speaking-style and language sets (100 Japanese city names, Japanese read speech, English read speech and Japanese lecture speech) of training data, and show the effectiveness with which VBEC automatically determined the model topology robustly according to the speech types of the training data. We also examine the robustness of the determined models for a mismatched condition between training and test data tasks, by recognizing question utterances for a question answering system by using the acoustic model trained by the Japanese read speech data set

2. VBEC

VBEC is a total Bayesian framework: it includes two major Bayesian abilities that are superior to the ML approach, in that it can determine an appropriate model topology and classify categories robustly using a predictive posterior [5] (Bayesian Predictive Classification using VB posterior distributions, VB-BPC). In this paper, we focus on model topology determination, which is a VBEC capability. In this section, we briefly review the VBEC framework and show the VB objective function used for determining a model topology (see [5, 7, 8] for details).

2.1 Variational Bayes

Let $\boldsymbol{O}$ be a given data set. Then, in the Bayesian approach we are interested in posterior distributions over the model parameters, $p(\Theta|\boldsymbol{O}, m)$, and in the model topology, $p(m|\boldsymbol{O})$. Here, Θ is a set of model parameters and m is an index of the model topology. Here, we derive VB posterior distributions for a model topology and discuss model selection by employing the posterior distribution for a model with latent variable Z. We introduce an arbitrary posterior distribution $q(m|\boldsymbol{O})$ and consider the KL divergence between $q(m|\boldsymbol{O})$ and the true posterior distribution $p(m|\boldsymbol{O})$. Then, the lower bound of $\mathrm{KL}[q(m|\boldsymbol{O})|p(m|\boldsymbol{O})]$ is obtained by using Jensen's inequality, and the following objective function is defined as follows:

$$\mathcal{F}^m[q(\Theta|\boldsymbol{O}, m), q(Z|\boldsymbol{O}, m)] \equiv \left\langle \log \frac{p(\boldsymbol{O}, Z|\Theta, m)p(\Theta|m)}{q(\Theta|\boldsymbol{O}, m)q(Z|\boldsymbol{O}, m)} \right\rangle_{q(\Theta|\boldsymbol{O},m),q(Z|\boldsymbol{O},m)}. \quad (1)$$

Here, the brackets $\langle\rangle_q$ denote the expectation with respect to q. Therefore, the optimal posterior distribution for a model topology can be obtained as follows by a variational method with respect to $q(m|\boldsymbol{O})$ that minimizes the lower bound:

$$\widetilde{q}(m|\boldsymbol{O}) \propto p(m) \exp\left(\mathcal{F}^m[q(\Theta|\boldsymbol{O}, m), q(Z|\boldsymbol{O}, m)]\right). \quad (2)$$

Assuming that $p(m)$ is a uniform distribution, the proportion relation between $\widetilde{q}(m|\boldsymbol{O})$ and $\mathcal{F}^m$ is obtained by utilizing the monotonical behavior of the exponential function. Therefore, the optimal model topology in the sense of maximum posterior probability can be selected as follows:

$$\widetilde{m} = \arg\max_m \widetilde{q}(m|\boldsymbol{O}) = \arg\max_m \mathcal{F}^m. \quad (3)$$

This indicates that by maximizing objective function $\mathcal{F}^m$ with respect to m, we can select the optimal model topology [7, 8] even if a model includes latent variables. $\mathcal{F}^m$ can be calculated by using VB posterior distributions for model parameters $q(\Theta|\boldsymbol{O}, m)$ and $q(Z|\boldsymbol{O}, m)$. $q(\Theta|\boldsymbol{O}, m)$ and $q(Z|\boldsymbol{O}, m)$ are calculated by an iterative calculation such as the Expectation-Maximization (EM) algorithm (VB-EM algorithm [6])

2.2 VB posterior distribution for acoustic model parameters

In this section, we introduce concrete forms of the VB posterior distributions for model parameters $q(\Theta|\boldsymbol{O}, m)$. First, we set output and prior distributions. Let $\boldsymbol{O} = \{\boldsymbol{o}^t \in \mathcal{R}^D : t = 1, ..., T\}$ be a sequential speech data set of a phoneme category. We use D to denote the dimension number of the feature vector and T to denote the frame number. The output distribution, which represents a phoneme acoustic model, is expressed by

$$p(\boldsymbol{O}, S, V|\Theta, m) = \prod_{t=1}^{T} a_{s^{t-1}s^t} w_{s^t v^t} b_{s^t v^t}(\boldsymbol{o}^t), \quad (4)$$

where S is a set of sequences of HMM states, V is a set of sequences of Gaussian mixture components, and s^t and v^t denote the state and mixture components at a frame t. Here, S and V are sets of discrete hidden variables, which are the concrete forms of Z in Section 2.1. The parameter a_{ij} denotes the state transition probability from state i to state j, and w_{jk} is the k-th weight factor of the Gaussian mixture for

state j. In addition, $b_{jk}(\boldsymbol{O}^t)(=\mathcal{N}(\boldsymbol{O}^t|\boldsymbol{\mu}_{jk}, \Sigma_{jk}))$ denotes the Gaussian with the mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix Σ_{jk}. $\Theta = \{a_{ij}, w_{jk}, \boldsymbol{\mu}_{jk}, \Sigma_{jk}^{-1} | i, j = 1, ..., J, k = 1, ..., L\}$ is a set of output distribution parameters. Here, J denotes the number of states in an HMM sequence and L denotes the number of Gaussian components in a state.

The prior distribution is assumed to be a conjugate distribution and is expressed as follows:

$$
\begin{aligned}
&p(\Theta|m) \\
&= \prod_{i=1}^{J}\prod_{j=1}^{J}\prod_{k=1}^{L} p(\{a_{ij'}\}_{j'=1}^{J}|m)p(\{w_{jk'}\}_{k'=1}^{L}|m)p(b_{jk}|m) \\
&= \prod_{i,j,k} \mathcal{D}(\{a_{ij'}\}_{j'=1}^{J}|\phi^0)\mathcal{D}(\{w_{jk'}\}_{k'=1}^{L}|\varphi^0) \\
&\mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\nu}_{jk}^0, (\xi^0)^{-1}\Sigma_{jk}) \prod_{d=1}^{D} \mathcal{G}(\Sigma_{jk,d}^{-1}|\eta^0/2, R_{jk,d}^0/2), \qquad (5)
\end{aligned}
$$

where $b_{jk} = \{\boldsymbol{\mu}_{jk}, \Sigma_{jk}^{-1}\}$. Here, $\Phi^0 = \{\phi^0, \varphi^0, \xi^0, \boldsymbol{\nu}_{jk}^0, \eta^0, R_{jk}^0\}$ is a set of prior parameters. We set $\phi^0, \varphi^0, \xi^0, \eta^0$ so that they are independent of i, j, k. $\mathcal{D}$ denotes a Dirichlet distribution and $\mathcal{G}$ denotes a gamma distribution. The prior distributions of a_{ij} and w_{jk} are represented by the Dirichlet distributions, and the prior distribution of $\boldsymbol{\mu}_{jk}$ and Σ_{jk} is represented by the normal-gamma distribution. If the covariance matrix elements are off the diagonal, a normal-Wishart distribution is set as the prior distribution of $\boldsymbol{\mu}_{jk}$ and Σ_{jk}.

From the output and prior distributions, we can obtain the optimal VB posterior distribution for the output distribution parameters $\widetilde{q}(\Theta|\boldsymbol{O}, m)$ (see [5] for details).

$$
\begin{aligned}
&\widetilde{q}(\Theta|\boldsymbol{O}, m) \\
&= \prod_{i,j,k} \widetilde{q}(\{a_{ij'}\}_{j'=1}^{J}|\boldsymbol{O}, m)\widetilde{q}(\{w_{jk'}\}_{k'=1}^{L}|\boldsymbol{O}, m)\widetilde{q}(b_{jk}|\boldsymbol{O}, m) \\
&= \prod_{i,j,k} \mathcal{D}(\{a_{ij'}\}_{j'=1}^{J}|\{\widetilde{\phi}_{ij'}\}_{j'=1}^{J})\mathcal{D}(\{w_{jk'}\}_{k'=1}^{L}|\{\widetilde{\varphi}_{jk'}\}_{k'=1}^{L}) \\
&\mathcal{N}(\boldsymbol{\mu}_{jk}|\widetilde{\boldsymbol{\nu}}_{jk}, (\widetilde{\xi}_{jk})^{-1}\Sigma_{jk}) \prod_{d} \mathcal{G}(\Sigma_{jk,d}^{-1}|\widetilde{\eta}_{jk}/2, \widetilde{R}_{jk,d}/2) \qquad (6)
\end{aligned}
$$

Note that Eqs. (5) and (6) belong to the same function family, and the only difference is that the set of prior parameters Φ^0 in Eq. (5) is replaced with a set of posterior distribution parameters $\widetilde{\Phi} \equiv \{\widetilde{\phi}_{ij}, \widetilde{\varphi}_{jk}, \widetilde{\xi}_{jk}, \widetilde{\boldsymbol{\nu}}_{jk}, \widetilde{\eta}_{jk}, \widetilde{R}_{jk}\}$ in Eq. (6). Here, $\widetilde{\Phi}$ are defined as:

$$
\begin{cases}
\widetilde{\phi}_{ij} &= \phi^0 + \sum_t \widetilde{\gamma}_{ij}^t \\
\widetilde{\varphi}_{jk} &= \varphi^0 + \sum_t \widetilde{\zeta}_{jk}^t \\
\widetilde{\xi}_{jk} &= \xi^0 + \sum_t \widetilde{\zeta}_{jk}^t \\
\widetilde{\boldsymbol{\nu}}_{jk} &= \left(\xi^0 \boldsymbol{\nu}_{jk}^0 + \sum_t \widetilde{\zeta}_{jk}^t \boldsymbol{O}^t\right)/\widetilde{\xi}_{jk} \\
\widetilde{\eta}_{jk} &= \eta^0 + \sum_t \widetilde{\zeta}_{jk}^t \\
\widetilde{R}_{jk,d} &= R_{jk,d}^0 + \xi^0(\nu_{jk,d}^0 - \widetilde{\nu}_{jk,d})^2 + \sum_t \widetilde{\zeta}_{jk}^t (O_d^t - \widetilde{\nu}_{jk,d})^2
\end{cases} \qquad (7)
$$

where

$$
\begin{cases}
\widetilde{\gamma}_{ij}^t &\equiv \widetilde{q}(s^{t-1} = i, s^t = j|\boldsymbol{O}, m) \\
\widetilde{\zeta}_{jk}^t &\equiv \widetilde{q}(s^t = j, v^t = k|\boldsymbol{O}, m)
\end{cases} \qquad (8)
$$

Here, $\widetilde{\gamma}_{ij}^t$ is a VB transition posterior distribution, which denotes the transition probability from a state i to a state j at a frame t, and $\widetilde{\zeta}_{jk}^t$ is a VB occupation posterior distribution, which denotes the occupation probability of a mixture component k in a state j at a frame t, in the VB approach. Therefore, $\widetilde{\Phi}$ can be calculated from Φ^0, $\widetilde{\gamma}_{ij}^t$ and $\widetilde{\zeta}_{jk}^t$, enabling $\widetilde{q}(\Theta|\boldsymbol{O}, m)$ to be obtained. $\widetilde{\gamma}_{ij}^t$ and $\widetilde{\zeta}_{jk}^t$ can be obtained by the VB forward - backward algorithm or Viterbi algorithm. Thus, VB posteriors can be calculated iteratively like the Baum-Welch algorithm even for complicated latent variable models such as the acoustic models within the VB-EM algorithm. Therefore, we refer to these calculations designed to obtain VB posteriors as a VB Baum-Welch algorithm, which is proposed in [4,5]. VBEC is based on the VB Baum-Welch algorithm.

2.3 VB objective function

In this section, we provide the concrete form of the VB objective function $\mathcal{F}^m$, which is a criterion for both posterior distribution estimation and model topology optimization. By substituting the VB posterior distribution obtained by the VB Baum-Welch algorithm in Section 2.2 into Eq. (1), we obtain analytical results for $\mathcal{F}^m$, and therefore, this calculation also requires the VB-EM algorithm used in the VB posterior calculation. We can separate $\mathcal{F}^m$ into two components: one is composed solely of $\widetilde{q}(S, V|\boldsymbol{O}, m)$, whereas the other is mainly composed of $\widetilde{q}(\Theta|\boldsymbol{O}, m)$. Therefore, we define $\mathcal{F}_{\Theta}^m$ and $\mathcal{F}_{S,V}^m$, and represent $\mathcal{F}^m$ as follows:

$$
\begin{aligned}
\mathcal{F}^m &= -\sum_{S,V} \widetilde{q}(S, V|\boldsymbol{O}, m) \log\left(\widetilde{q}(S, V|\boldsymbol{O}, m)\right) \\
&+ \left\langle \sum_{S,V} \widetilde{q}(S, V|\boldsymbol{O}, m) \log\left(\frac{p(\boldsymbol{O}, S, V|\Theta, m)p(\Theta|m)}{\widetilde{q}(\Theta|\boldsymbol{O}, m)}\right)\right\rangle_{\widetilde{q}(\Theta|\boldsymbol{x}, m)} \\
&\equiv -\mathcal{F}_{S,V}^m + \mathcal{F}_{\Theta}^m. \qquad (9)
\end{aligned}
$$

We provide the $\mathcal{F}_{\Theta}^m$ result, which is used in Section 3., as:

$$
\begin{aligned}
\mathcal{F}_{\Theta}^m &= \sum_i \log\left(\frac{\Gamma(\sum_j \phi^0)}{\Gamma(\sum_{j'} \widetilde{\phi}_{ij})} \frac{\prod_j \Gamma(\widetilde{\phi}_{ij})}{\prod_j \Gamma(\phi^0)}\right) \\
&+ \sum_j \log\left(\frac{\Gamma(\sum_k \varphi^0)}{\Gamma(\sum_k \widetilde{\varphi}_{jk})} \frac{\prod_k \Gamma(\widetilde{\varphi}_{jk})}{\prod_k \Gamma(\varphi^0)}\right) \\
&+ \sum_{j,k} \log\left((\pi)^{-\frac{\widetilde{\zeta}_{jk} D}{2}} \left(\frac{\xi^0}{\widetilde{\xi}_{jk}}\right)^{\frac{D}{2}} \frac{\left(\Gamma\left(\frac{\widetilde{\eta}_{jk}}{2}\right)\right)^D |R_{jk}^0|^{\frac{\eta^0}{2}}}{\left(\Gamma\left(\frac{\eta^0}{2}\right)\right)^D |\widetilde{R}_{jk}|^{\frac{\widetilde{\eta}_{jk}}{2}}}\right), \qquad (10)
\end{aligned}
$$

where $\Gamma(\cdot)$ denotes a gamma function. From Eq. (10), $\mathcal{F}_{\Theta}^m$ can be calculated by using the statistics of the posterior distribution parameters $\widetilde{\Phi}$ given in Eq. (7).

3. Determination of acoustic model topology using VBEC

In this section, we briefly describe how to determine the acoustic model topology automatically by using VBEC. In acoustic modeling, the specifications of the model topology are often represented by the number of clustered states and the number of GMM components per state, as shown in Figure 1. Then, the good models that provide good performance would be distributed in the inverse-proportion band where the total number of distribution parameters (approximately equal to the total number of Gaussians) exists in a restricted range where over-fitting and under-fitting are avoided. Moreover, there would be a unimodal optimum area in the band where the model topologies are represented by an appropriate number of pairs of clustered states and components, as shown in Figure 1. In order to realize the optimum model topology, we utilize the two characteristics of the acoustic model: the inverse-proportion band and the unimodality.

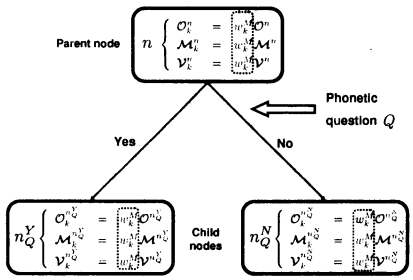

Figure 2 Estimation of inheritable GMM statistics while splitting.

Therefore, we first prepare a number of acoustic models in the band (in-band models) by using GMM-based decision tree clustering, and then choose the model that has the best VB objective function score from the in-band models (in-band model search). Thus we can determine the optimum model topology, as shown in Figure 1.

3.1 GMM-based decision tree clustering

In this section, we describe GMM-based decision tree clustering within the VBEC framework. We employ phonetic question based state clustering in order to reduce the state cluster combination. An appropriate choice of phonetic question at each node split leads a decision tree to proper growth to maximize an objective function, and appropriate state clusters become represented in its leaf nodes. We use the VB objective function $\mathcal{F}^m$ in Section 2. 3 as an objective function for the state clustering. Although the conventional method uses likelihood and requires a manually tuned threshold, $\mathcal{F}^m$ does not require such a threshold, and can determine clustering topology appropriately [5]. When a node n is split into yes (n_Y^Q) and no (n_N^Q) nodes according to a question Q, an appropriate question $\widetilde{Q}(n)$ is chosen from a set of questions so that the split gives the largest gain in an arbitrary objective function $\mathcal{F}^m$, i.e., $\widetilde{Q}(n) = \arg\max_Q \Delta\mathcal{F}^{Q(n)}$, where $\Delta\mathcal{F}^{Q(n)} \equiv \mathcal{F}^{n_Y^Q, n_N^Q} - \mathcal{F}^n$ is the overall gain in objective function when node n is split by Q. The automatic determination of the final state clustering topology is achieved by stopping the split when $\Delta\mathcal{F}^{n(Q)} < 0$ for any node n and any question Q. This corresponds to finding the appropriate model topology of a clustered state structure where the total $\mathcal{F}^m$ is maximized for all possible clustered state topologies.

Although $\Delta\mathcal{F}^{Q(n)}$ can be calculated by the VB Baum-Welch algorithm for each cluster combination, this requires an extremely long computation time. Therefore, we set the following two constraints to eliminate the latent variables involved in an acoustic model and thus avoid the VB Baum-Welch algorithm. The first is that the frame-to-state alignments are fixed throughout the process of state splitting. The second is that the ratio of GMM statistics for component k is conserved when splitting. Then, for example, the ratio of 0-th order statistics $\mathcal{O}_k^n$ for a node n is related to the ratios of the 0-th order statistics of its yes-node n_Q^Y and no-node n_Q^N for a question Q. Employing this relation for the upper node in a phonetic tree successively, the assumption yields the fact that the ratio at each node is equivalent to the ratio $\mathcal{O}_k^M / \sum_{k'} \mathcal{O}_{k'}^M (\equiv w_k^M)$ at the root node statistics in the tree (i.e., the ratio of the monophone HMM state statistics) as follows:

$$\frac{\mathcal{O}_k^{n_Q^Y}}{\sum_{k'} \mathcal{O}_{k'}^{n_Q^Y}} = \frac{\mathcal{O}_k^{n_Q^N}}{\sum_{k'} \mathcal{O}_{k'}^{n_Q^N}} = \frac{\mathcal{O}_k^n}{\sum_{k'} \mathcal{O}_{k'}^n} = \cdots = \frac{\mathcal{O}_k^M}{\sum_{k'} \mathcal{O}_{k'}^M} \equiv w_k^M. \tag{11}$$

where the suffix M indicates a monophone HMM state. Therefore, utilizing a similar discussion, the 0-th, 1-st and 2nd order statistics $\mathcal{O}_k^n$, $\boldsymbol{\mathcal{M}}_k^n$ and $\mathcal{V}_k^n$ of the k component in the n node are estimated using w_k^M as follows:

$$\begin{cases} \mathcal{O}_k^n &= w_k^M \mathcal{O}^n. &= w_k^M \sum_{j\in n} \sum_t \widetilde{\zeta}_j^t. \\ \boldsymbol{\mathcal{M}}_k^n &= w_k^M \boldsymbol{\mathcal{M}}^n &= w_k^M \sum_{j\in n} \sum_t \widetilde{\zeta}_j^t \boldsymbol{O}^t \\ \mathcal{V}_{k,d}^n &= w_k^M \mathcal{V}_d^n. &= w_k^M \sum_{j\in n} \sum_t \widetilde{\zeta}_j^t (\boldsymbol{O}_d^t)^2. \end{cases} \tag{12}$$

where, j denotes the index of a non-clustered triphone HMM state. This approach is based on the hierarchical knowledge of the tree structure, which shows that child node statistics are well complemented by parent node statistics. Thus, we can estimate the GMM statistics of each node $\mathcal{O}_k^n$, $\boldsymbol{\mathcal{M}}_k^n$ and $\mathcal{V}_k^n$ without using the VB-EM algorithm, but using the k component ratio of the monophone statistics w_k^M. We call this the estimation of inheritable node statistics because the ratio w_k^M is passed from a parent node to child nodes, as shown in Figure 2. Consequently, VB posteriors and VB objective functions can also be calculated without using the VB-EM algorithm while splitting. We provide the concrete form of the parameters for VB posteriors $\widetilde{\Phi}(n)$ and $\Delta\mathcal{F}^{Q(n)}$ from Eqs. (7) and (10), as follows:

$$\Delta\mathcal{F}^{Q(n)} = f(n_Q^Y) + f(n_Q^N) - f(n) - \sum_k w_k^M \log w_k^M, \tag{13}$$

where

$$f(n) \equiv \log \frac{\Gamma(L\varphi^0)}{\Gamma\left(\sum_k \widetilde{\varphi}_k^n\right)} \frac{\prod_k \Gamma(\widetilde{\varphi}_k^n)}{(\Gamma(\varphi^0))^L}$$

$$+ \sum_k \log \left((\pi)^{-\frac{\widetilde{\xi}_k^n D}{2}} \left(\frac{\xi^0}{\widetilde{\xi}_k^n}\right)^{\frac{D}{2}} \frac{\left(\Gamma\left(\frac{\widetilde{\eta}_k^n}{2}\right)\right)^D |R_k^{n,0}|^{\frac{\eta^0}{2}}}{\left(\Gamma\left(\frac{\eta^0}{2}\right)\right)^D \left|\widetilde{R}_k^n\right|^{\frac{\widetilde{\eta}_k^n}{2}}} \right). \tag{14}$$

$$\begin{cases} \widetilde{\varphi}_k^n &= \varphi^0 + w_k^M \mathcal{O}^n \\ \widetilde{\xi}_k^n &= \xi^0 + w_k^M \mathcal{O}^n \\ \widetilde{\eta}_k^n &= \eta^0 + w_k^M \mathcal{O}^n \\ \widetilde{\boldsymbol{\nu}}_k^n &= \frac{\xi^0 \boldsymbol{\nu}_k^{n,0} + w_k^M \boldsymbol{\mathcal{M}}^n}{\xi^0 + w_k^M \mathcal{O}^n} \\ \widetilde{R}_{k,d}^n &= R_{k,d}^{n,0} + w_k^M \mathcal{V}_d^n - w_k^M \frac{(\boldsymbol{\mathcal{M}}_d^n)^2}{\mathcal{O}^n} \\ &\quad + \frac{\xi^0 w_k^M \mathcal{O}^n}{\xi^0 + w_k^M \mathcal{O}^n} \left(\frac{\boldsymbol{\mathcal{M}}_d^n}{\mathcal{O}^n} - \boldsymbol{\nu}_{k,d}^{n,0} \right)^2 \end{cases} \tag{15}$$

To calculate $\Delta\mathcal{F}^{Q(n)}$, we must estimate w_k^M and set the prior parameters Φ^0 appropriately.

3.2 Prior and monophone HMM statistics

In this paper, we introduce an approach for obtaining w_k^M, $\boldsymbol{\nu}_k^{n,0}$ and $R_k^{n,0}$, which was first proposed in [11]. This approach assumes that w_k^M is the same for all the components in an L-component GMM, and is represented by $w_k^M = 1/L$, instead of calculating the GMM statistics of monophone HMM. In addition, to set $\boldsymbol{\nu}_k^{n,0}$ and $R_k^{n,0}$, we employ single Gaussian statistics of monophone HMM, which are easily computed by combining sufficient statistics $\mathcal{O}_j$, $\boldsymbol{\mathcal{M}}_j$ and $\mathcal{V}_j$ for all triphone HMM states. Then, w_k^M, $\boldsymbol{\nu}_k^{n,0}$ and $R_k^{n,0}$ are represented as follows:

$$\begin{cases} w_k^M &= 1/L \\ \boldsymbol{\nu}_k^{n,0} &= \sum_j \mathcal{O}_j \boldsymbol{\mathcal{M}}_j / \sum_j \mathcal{O}_j \\ R_{k,d}^{n,0} &= \eta^0 \sum_j \mathcal{O}_j \mathcal{V}_{j,d} / \sum_j \mathcal{O}_j. \end{cases} \tag{16}$$

The gain of VB objective function $\Delta\mathcal{F}^{Q(n)}$ is calculated by substituting Eq. (16) into Eqs. (13) and (14). Thus, we can construct a number of in-band model topologies by using GMM-based decision tree clustering, i.e., we realize the solid arrows seen in Figure 1 within a practical computation time without using the VB-EM algorithm. However, the final determination of an appropriate model from in-band models (in-band model search) does not sustain the inheritable statistics assumption because the number of GMM components is different for each model. Therefore, we have to calculate the VB objective function by dropping the inheritable statistics assumption. In addition, we also drop the constraints of the frame-to-state alignments in Section 3.1 to calculate the VB objective function as exactly as possible. In this situation, the acoustic model includes latent variables, and we require the exact VB objective function as described in Eq. (9), which is obtained by the VB Baum-Welch algorithm, instead of using Eqs. (13) and (14).

4. Experiments

An automatic method for determining acoustic models would be a very promising technique for practical speech application fields if it is unaffected by speech variation. Therefore, the availability of VBEC automatic determination should be examined experimentally using various speech data. This experimental section provides three subsections to confirm the robustness with respect to different speaking styles and languages, which are representative of speech variations, and of different test data by recognizing question utterances for a question answering system using an acoustic model trained by the Japanese read speech data set, whose conditions are mismatched with those of test data set. All the experiments in this paper were performed using the SOLON speech recognition toolkit [12] developed by NTT Communication Science Laboratories.

4.1 Speaking style variation

First, we focused on speaking style variation of the training data set by preparing an isolated word speech (100 city names provided by JEIDA), LVCSR based on Japanese read speech (JNAS: Japanese Newspaper Article Sentences) and LVCSR based on Japanese lecture speech (CSJ: Corpus of Spontaneous Japanese). Speaking style greatly influences acoustic features, and acoustic models need to be constructed *manually* depending on the style when the ML method is used. However, VBEC determination could allow us to replace manual construction with automatic construction for various speaking styles. Therefore, we examined the robustness of VBEC determination for various speaking styles. The configuration of feature extraction was 12-order MFCC + Δ MFCC (24 dim.) for 100 city names, 12-order MFCC + Δ MFCC + Energy + Δ Energy (26 dim.) for JNAS and 12-order MFCC + Δ MFCC + Δ Energy (25 dim.) + CMN for CSJ. The sampling rate was 16 kHz, the frame size was 25 ms and the frame shift was 10 ms. For JNAS and CSJ, we used standard trigram models with vocabularies of 20,000 and 30,000, respectively. For the 100 city name task, the training data consisted of about 3,000 Japanese sentences (4.1 hours) spoken by 30 males and the recognition data consisted of 100 Japanese city names spoken by 25 males (a total of 2,400 words). For the JNAS task, the training data consisted of about 20,000 Japanese sentences (34 hours) spoken by 122 males and the recognition data consisted of 100 Japanese sentences spoken by 10 males (a total of about 2,000 words). For the CSJ task, the training data consisted of about 800 Japanese lectures (190 hours) spoken by 200 males and the recognition data consisted of 10 Japanese lectures spoken by 10 males (a total of about 27,000 words).

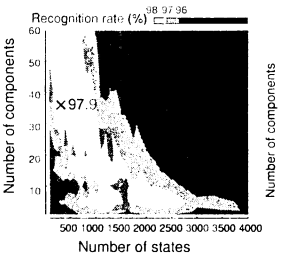

Figure 3 100 city name.

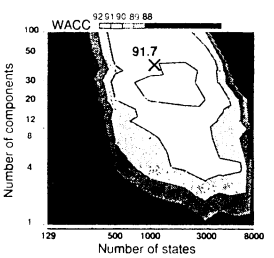

Figure 4 JNAS.

Determined model topologies and their word accuracies

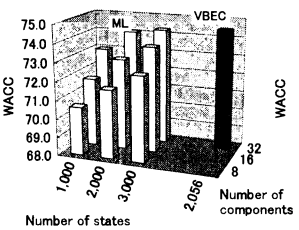

Figure 5 CSJ.

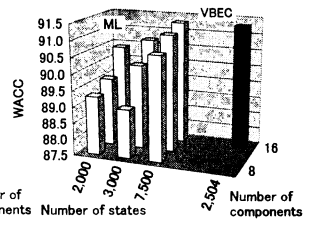

Figure 6 WSJ.

Determined model topologies and their word accuracies

First, we examined the recognition performance of conventional ML-based acoustic models with manually varied model topologies for a number of clustered states and GMM components per state, which we use as baselines with which to compare the performance of the automatically determined model topology. The contour maps in Figures 3 and 4, and the white bar in Figure 5 show the recognition performance obtained with the ML method. Then, we provided the model, whose topologies were determined by VBEC, with recognition performance. For all the tasks, the resultant combinations of the numbers of states and components per state, determined by VBEC, were included in the high performance area in Figures 3, 4 and 5. In addition, the recognition performance (97.9 %, 91.7 WACC and 74.5 WACC) of all the tasks reached the highest performance (98.0 %, 91.4 WACC and 74.2 WACC) obtained with ML methods. Consequently, we confirmed that VBEC determination is effective for various speaking styles, namely isolated word speech, continuous read speech and spontaneous lecture speech.

4.2 Language variation

In our second set of experiments, we focused on the effect on VBEC determination of language variation. The acoustic feature depends strongly on the languages, and the appropriate model topology will be changed depending on the language. Therefore, we must examine how VBEC determination works even for a different language task. We used English read speech (WSJ: Wall Street Journal) as a different language task from Japanese tasks. The feature extraction configuration is 12-order MFCC + Δ MFCC + ΔΔ MFCC + Energy + Δ Energy + ΔΔ Energy (39 dim.) + CMN. The other configuration was the same as that in Section 4.1. We used a standard trigram model that had a vocabulary of 20,000. The training data consisted of about 20,000 English sentences (36 hours) spoken by 143 males and the recognition

data consisted of 100 English sentences spoken by 5 males (a total of about 2,000 words).

As in Section 4.1, we prepared the recognition performance of conventional ML-based acoustic models with manually varied model topologies for a number of clustered states and GMM components per state. The white bar in Figure 6 shows the recognition performance obtained by the ML method and the black bar represents the VBEC determined model with the recognition performance. Although the determined model topology with 2,504 states and 32 components was far from the best ML results of 7,000 states and 32 components, its performance (91.3 WACC) matched the best ML performance (91.3 WACC), and we can say that VBEC determination is effective even for a different language task such as English rather than Japanese. In addition, the VBEC determined model exhibited the best ML performance with less than half the total number of Gaussians, which reduced the decoding time to less than half (8.29 RTF $\rightarrow$ 2.35RTF).

4.3 Mismatched condition with training and test data tasks

Finally, we examined how the model previously determined using VBEC worked for a test data set belonging to a different corpus from the training data corpus, i.e., the conditions, such as recording environments and speaking styles, between training and test data sets are mismatched. We used the determined acoustic model trained using JNAS in Section 4.1 and recognized spoken question utterances for a question answering system (QA) [12]. Since the question utterances include many proper nouns, we must prepare a language model with a very large vocabulary that includes the proper nouns. In this paper, we recognized speech by using very large vocabulary language models (0.2 million, 1.0 million and 1.8 million). The recognition data consisted of 250 Japanese question utterances spoken by 25 male (a total of about 3,000 words).

Figure 7 compares the recognition performance obtained using the VBEC determined acoustic model with 912 states and 40 components and the two ML based acoustic models with 1,000 states and 30 components and 2,000 states and 40 components in the mismatched (JNAS-QA) condition between training and test data. We also added the matched (JNAS-JNAS) condition result in Figure 7. The ML models were obtained by manually tuning the model topologies that maximized that recognition performance of two development data sets extracted from JNAS, which were not included in the training data. From Figure 7, the VBEC determined model was superior to the ML model topologies for every vocabulary size by more than 2 points in the mismatched condition, unlike the matched condition result. The reason of VBEC's superiority seems to be that the ML topologies were overly tuned for the JNAS development data sets that was mismatched with the question utterances, and could not accommodate mismatched data such as question answering speech, while the VBEC model was determined only by training data, and would be robust for the mismatched condition.

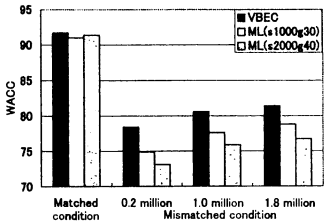

Figure 7 Word accuracies for question answering speech.

5. Experimental discussion and summary

In this paper, we introduced the *automatic* determination of the optimum topology for an acoustic model by using Gaussian Mixture Model (GMM)-based phonetic decision tree clustering and an efficient model search algorithm that utilized the acoustic model characteristics. This method was realized using the Variational Bayesian Estimation and Clustering for speech recognition (VBEC) framework. The robustness of the automatic determination in terms of speaking style and language was confirmed experimentally, and showed that VBEC determination is robust for various types of speaking style and language. Moreover, we also confirmed the robustness of the determined model for a mismatched condition between training and test data task, which shows the superiority of the VBEC determined model as regards open data, which cannot be prepared beforehand at the training stage. Thus, VBEC provides a consistent theoretical framework for total Bayesian speech recognition, and constitutes a very promising technique for practical speech recognition.

6. Acknowledgement

We thank Mr. Steven Dwyer and Dr. Takaaki Hori for providing us with their experimental conditions for WSJ and spoken question utterance tasks.

References

[1] J. Odell, *The use of context in large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 1995.

[2] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, pp. 17–41, 1997.

[3] T. Kato, S. Kuroiwa, T. Shimizu, and N. Higuchi, "Efficient mixture Gaussian synthesis for decision tree based state tying," in *Proc. ICASSP2001*, 2001, vol. 1, pp. 493–496.

[4] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, *Application of variational Bayesian approach to speech recognition*, NIPS 2002, MIT Press, 2002.

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

[6] S. Waterhouse, D. MacKay, and T. Robinson, *Bayesian methods for mixtures of experts*, NIPS 7, MIT Press, 1995.

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

[8] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, vol. 15, pp. 1223–1241, 2002.

[9] P. Somervuo, "Speech modeling using variational Bayesian mixture of Gaussians," in *Proc. ICSLP2002*, 2002, vol. 2, pp. 1245–1248.

[10] T. Jitsuhiro and S. Nakamura, "Automatic generation of non-uniform HMM structures based on variational Bayesian approach," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 805–808.

[11] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 813–816.

[12] T. Hori, "NTT Speech recognizer with OutLook On the Next generation: SOLON," in *Proc. NTT Workshop on Communication Scene Analysis*, 2004, vol. 1, SP-6.