AUTOMATIC DETERMINATION OF ACOUSTIC MODEL TOPOLOGY USING VARIATIONAL BAYESIAN ESTIMATION AND CLUSTERING

†*Shinji Watanabe,* ‡*Atsushi Sako and* †*Atsushi Nakamura*

† Speech Open Laboratory, NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
‡ Department of Science and Technology, Ryukoku University
1-5, Yokotani, Oe-cho, Seta, Otsu-shi, Shiga, 520-2194, Japan

ABSTRACT

We describe the automatic determination of an acoustic model for speech recognition, which is very complicated and includes latent variables, using VBEC: Variational Bayesian Estimation and Clustering for speech recognition. We propose an efficient Gaussian Mixture Model (GMM) based phonetic decision tree construction within the VBEC framework. The proposed method features a novel approach to reduce the unrealistically large number of computations needed for iterative calculations in the GMM-based decision tree method to a practical level by assuming that each Gaussian per state has the same occupancy and is represented by the same posterior distribution for the covariance parameter. The experimental results confirmed that VBEC *automatically* provided a *optimum* model topology with the highest performance level.

1. INTRODUCTION

Precise acoustic modeling is important in terms of speech recognition. The acoustic model has a very complicated structure: a category is expressed by a clustered-state triphone Hidden Markov Model (HMM) that possesses an output distribution represented by a Gaussian Mixture Model (GMM). Certain algorithms have been proposed to deal with the complicated model structure (model topology) [1, 2]. However, some heuristic tuning is required since these algorithms are based on the Maximum Likelihood (ML) criterion, which cannot determine the model topology because ML increases monotonically as the number of model parameters increases. If we are to eliminate the need for heuristic tuning we must find a way to determine the acoustic model topology automatically.

Some partially successful approaches to the automatic determination of the acoustic model topology have been reported that used a Minimum Description Length or Bayesian Information Criterion (MDL/BIC). These approaches determine the clustered-state triphone HMM structure with a single Gaussian and the total number of single Gaussians in a model on the assumption that the acoustic model has no latent variables [3, 4]. However, the MDL/BIC criterion cannot theoretically determine the total acoustic model topology since it includes latent variables.

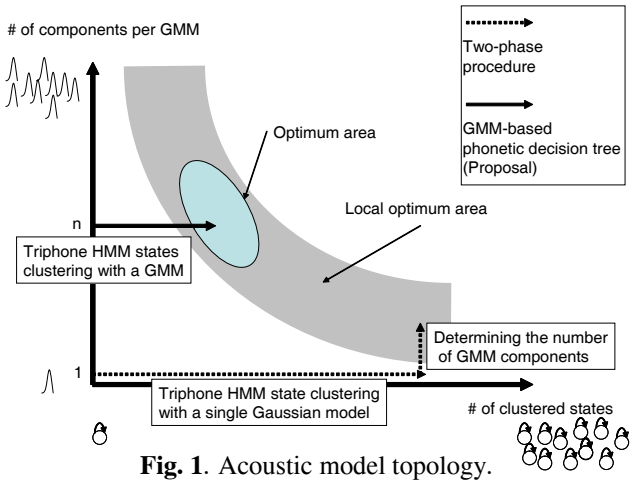

Fig. 1. Acoustic model topology.

Recently, a new framework has been proposed for the automatic determination of the acoustic model topology, namely Variational Bayesian Estimation and Clustering for speech recognition (VBEC) [5]). VBEC is a total Bayesian framework using Variational Bayes (VB) [6], and can theoretically determine a complicated model structure by using the VB objective function even when latent variables are included. In previous work, automatic determination using VBEC was confirmed based on a two-phase procedure for determining the model topology [7], i.e., clustering triphone HMM states with a single Gaussian model, and then determining the number of components per state, as shown in Figure 1. Although this procedure could determine the model topology within a practical computation time, the obtained topology is only locally optimized at each phase and the obtained performance was not the best.

In this study, our goal is to obtain the optimum topology using VBEC. If clustered-state triphone HMMs with a multiple-component GMM are constructed by using a GMM-based phonetic decision tree [8], the model topology is automatically determined by selecting the most appropriate of the several topologies, which maximizes the VB objec-

tive function. Although the construction of the GMM-based phonetic decision tree is also automatically determined within a VBEC framework, the construction requires an unrealistic number of computations because the VB objective function is calculated by an iterative algorithm for each node and for each phonetic question. To reduce the number of computations to a practical level, we propose a new approach for realizing the phonetic decision tree method within a VBEC framework by assuming that each Gaussian in GMM has the same statistics.

Finally, we undertake experiments to determine the acoustic model topology using GMM-based decision tree constructions and achieve the optimum topology of an acoustic model with the highest performance within a practical computation time.

2. VARIATIONAL BAYESIAN ESTIMATION AND CLUSTERING FOR SPEECH RECOGNITION

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 can classify categories robustly using a predictive posterior [5]. In this paper, we focus on the ability of VBEC, which can determine a model topology that includes latent variables. In this section, we briefly review VBEC (see [5, 6] for details).

Let $\boldsymbol{O} = \{\boldsymbol{O}^t \in \mathcal{R}^D | t = 1, ..., T\}$ be a set of training data sequences of D dimensional feature vectors for a phoneme category. In the acoustic modeling of speech recognition, the output distribution is parameterized by HMM and GMM as $p(\boldsymbol{O}, S, V|\Theta, m)$ where S and V are sets of HMM state and GMM component sequences, respectively. Θ is a set of distribution parameters, e.g., the state transition, mixture weight, and mean and covariance of Gaussian parameters. m denotes the model topology index. In the Bayes approach, Θ, S, V and m are regarded as probabilistic variables. In VB, VB posterior distributions $q(\Theta|\boldsymbol{O}, m)$, $q(S, V|\boldsymbol{O}, m)$, and $q(m|\boldsymbol{O})$ are introduced to approximate the true corresponding posterior distributions. The optimal VB posterior distributions over Θ and S, V, and the appropriate model topology that maximizes the optimal $q(m|\boldsymbol{O})$ can be obtained by maximizing the following objective function:

$$\mathcal{F}^m = \left\langle \log \frac{p(\boldsymbol{O}, S, V|\Theta, m)p(\Theta|m)}{q(S, V|\boldsymbol{O}, m)q(\Theta|\boldsymbol{O}, m)} \right\rangle_{\substack{q(S,V|\boldsymbol{O},m) \\ q(\Theta|\boldsymbol{O},m)}}, \quad (1)$$

w.r.t. $q(\Theta|\boldsymbol{O}, m), q(S, V|\boldsymbol{O}, m)$, and m. Here $\langle f(y) \rangle_{p(y)}$ denotes the expectation of $f(y)$ w.r.t. $p(y)$. $p(\Theta|m)$ is a prior distribution and is set as a conjugate prior distribution.

$\mathcal{F}^m$ is calculated by the VB posterior distributions that are parameterized by $\widetilde{\Phi} \equiv \{\widetilde{\phi}, \widetilde{\varphi}, \widetilde{\boldsymbol{\nu}}, \widetilde{\xi}, \widetilde{\eta}, \widetilde{R}\}$ defined as:

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

$\widetilde{\Phi}$ is composed of $\widetilde{\gamma}^t_{ij}$, $\widetilde{\zeta}^t_{jk}$ and a set of hyper-parameters $\Phi^0 \equiv \{\phi^0, \varphi^0, \boldsymbol{\nu}^0_{jk}, \xi^0, \eta^0, R^0_{jk}\}$. $\widetilde{\gamma}^t_{ij}$ and $\widetilde{\zeta}^t_{jk}$ are obtained by $q(S, V|\boldsymbol{O}, m)$ and denote the posterior transition probability from state i to state j at time t, and the posterior occupation probability on mixture component k in state j at time t, respectively.

By substituting the obtained $\widetilde{\Phi}$ of VB posterior distributions into Eq. (1), we obtain VB objective function $\mathcal{F}^m$. Therefore, the optimal model topology $\widetilde{m}$ can be selected by $\widetilde{m} = \arg\max_m p(m|\boldsymbol{O}) \approx \arg\max_m \mathcal{F}^m$. In an acoustic model that includes latent variables, the calculation of $\mathcal{F}^m$ requires an iterative algorithm similar to the expectation maximization algorithm.

3. DETERMINATION OF ACOUSTIC MODEL TOPOLOGY USING VBEC

In this section, we explain how to realize the automatic determine of the acoustic model topology using VBEC. To obtain the optimum topology, we adopt a method using the GMM-based phonetic decision tree. In the VBEC framework, an appropriate phonetic question at each split is chosen to increase the VB objective function $\mathcal{F}^m$, unlike the conventional approach using likelihood as the objective function. When a node n is split into a yes node (n^Q_Y) and a no node (n^Q_N) by a question Q, the appropriate question $\widetilde{Q}(n)$ is chosen to maximize the gain of $\mathcal{F}^m$ from the question set $\{Q\}$, i.e., $\widetilde{Q}(n) = \arg\max_{\{Q\}} \Delta\mathcal{F}^{Q(n)}$, where $\Delta\mathcal{F}^{Q(n)} \equiv \mathcal{F}^{\Omega(n^Q_Y),\Omega(n^Q_N)} - \mathcal{F}^{\Omega(n)}$ is the gain in total objective function when a node n is split by Q. By stopping splitting when $\Delta\mathcal{F}^{Q(n)} < 0$, an appropriate model topology is selected witout using manual tuning.

There are generally two conditions for phonetic decision tree construction [1]:

1. Frame-to-state assignments during splitting are fixed.
2. A single Gaussian for one state is used.

The assumptions are used in order to avoid the unrealistic number of objective function computations generated by the iterative algorithm by eliminating latent variables from the model. Although condition 1 is also available in the GMM-based phonetic decision tree, condition 2 is not. Therefore, GMM-based phonetic decision tree construction takes an unrealistic amount of computation time.

To avoid the impractical computation, we propose a new approach that approximates the GMM-based $\mathcal{F}^m$ calculation as a non-iterative algorithm by assuming that each Gaussian per state has the same occupancy and is represented by the same VB posterior distribution for the covariance. Sufficient statistics of a clustered state for a node n (occupancy $\zeta(n)$, mean vector $\boldsymbol{\mu}(n)$ and covariance matrix $\Sigma(n)$) are obtained with a non-iterative calculation. By assuming that a k-th component Gaussian in L-component GMM has the occupancy $\zeta(n)/L$, mean vector $\boldsymbol{\mu}(n)$ and covariance matrix $\Sigma(n)$ by utilizing the sufficient statistics of a state, the VB posterior parameters in Eq. (2) are obtained as follows:

$$\begin{cases} \widetilde{\varphi}_k(n) &= \varphi^0 + \zeta(n)/L \\ \widetilde{\xi}_k(n) &= \xi^0 + \zeta(n)/L \\ \widetilde{\boldsymbol{\nu}}_k(n) &= \big(\xi^0\boldsymbol{\mu}^0 + \zeta(n)\boldsymbol{\mu}(n)/L\big)/\widetilde{\xi}_k(n) \\ \widetilde{\eta}_k(n) &= \eta^0 + \zeta(n)/L \\ \widetilde{R}_{k,d}(n) &= R_d^0 + \widetilde{\xi}_k(n)(\nu_d^0 - \widetilde{\nu}_{k,d}(n))^2 + \\ & \quad \zeta(n)(\Sigma_d(n))^2/L \end{cases} \tag{3}$$

Here, we assume that the other Gaussians of the state n have the same occupancy and are represented by the same VB posterior distribution for the covariance. Then, $\mathcal{F}^m$ is obtained with a non-iterative calculation by substituting Eq. (3) into Eq. (1). Consequently, clustered-state triphone HMMs with a GMM are obtained with non-iterative algorithms. Then, conditions 1 and 2 is dropped and VB training is used to estimate the VB posteriors for a fixed model topology with a given number of components. This procedure is repeated for a range of setting of the number of components per state. The final model is determined by selecting the clustered-state triphone HMM with the best number of components per state, maximizing the VB objective function $\mathcal{F}^m$ in Eq. (1).

4. PRELIMINARY EXPERIMENTS

Before proving the effectiveness of the proposed method, we conducted preliminary experiments to examine the recognition performance of conventional ML-based acoustic models with manually varied model topologies, as baselines with which to compare the performance of the automatically determined model topology. Also, through this examination, we could see how the performance was distributed over the numbers of states and components per state. Several topologies were produced with manually varied conditions for the number of states, i.e., the sizes of the phonetic decision trees, and the number of components per state. We obtained a total of 216 acoustic models. The experimental conditions are summarized in Table 1. The training data consisted of about 3,000 Japanese sentences (4.1 hours) spoken by 30 males. The recognition data consisted of 100 Japanese city names spoken by 25 males (a total of 2,400 words).

Figure 2 shows the results for the examined recognition rate on a contour map. We can see a high performance area (96 %) along an inversely proportional curve in the map. The curve satisfied the relationship whereby the product of the numbers of states and components per state was about 20,000. Therefore, the results suggested that high performance levels were obtained when the total number of Gaussians was about 20,000. Moreover, we can see the top score (98.0 %) in the high performance area where the numbers of states and components per state were 1,000 and 15, respectively. We can also see that the other highest scores are distributed across the region around the top rate (white area, i.e., more than 97.0 %), which is regarded as the optimum area. Thus, although there were points showing high levels of performance for either the arbitrary number of states or the arbitrary number of components per state, these points do not necessarily indicate the optimum performance. Namely, the optimum performance cannot be found by the two-phase procedure, namely first determining the number of states and then determining the number of components per state, but by a procedure which determines the numbers of states and components per state in one phase.

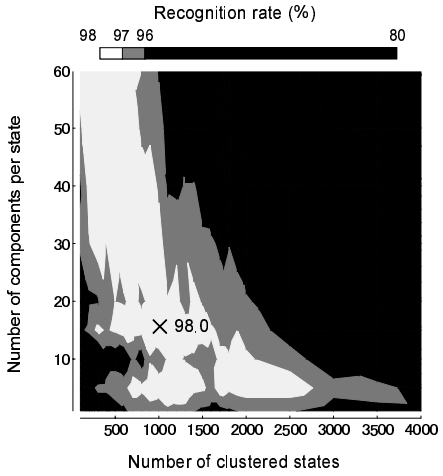

Fig. 2. Recognition rates for the number of total clustered states and components per state.

5. EXPERIMENTS

The proposed non-iterative algorithm based on VBEC using the approximate values of $\mathcal{F}^m$ described in Section 3 enables the automatic determination of both the numbers of states and components. We conducted experiments to prove the effectiveness of the proposed algorithm. The experimental conditions were the same as those described in Section 4. Here, we employed conventional ML decoding

Table 1. Experimental conditions

Sampling rate/Quantization	16 kHz / 16 bit
Feature vector	12 - order MFCC with Δ MFCC
Window	Hamming
Frame size/shift	25/10 ms
# of states	3 (Left to Right)
# of phoneme categories	27
# of phonetic questions	44

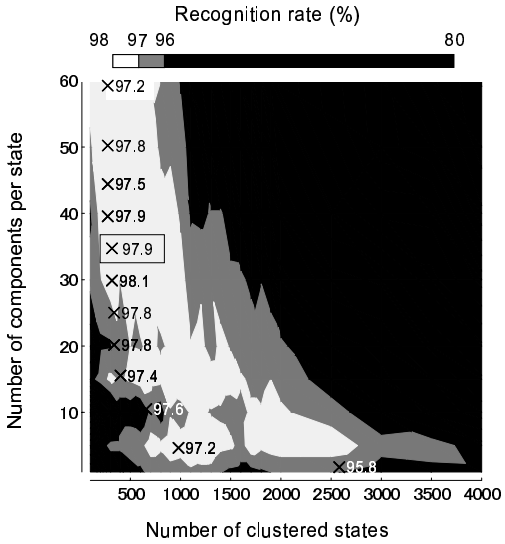

Fig. 3. Determined model topologies and their recognition rates.

in recognition instead of the Bayesian Predictive Classification (BPC) based decoding of VBEC. This was to allow us to evaluate the pure effect of the automatic determination of the model topology using the proposed algorithm.

The proposed non-iterative algorithm was used to produce a set of clustered-state triphone HMMs, which made 11 sets of clustered-state HMMs in total (1, 5, 10, 15, 20, 25, 30, 35, 40, 50 and 60 components per state). The selected model topologies and obtained recognition rates are plotted in Figure 3, and are overlaid on the contour map of Figure 2. Almost all the models were located in the white area and almost all the recognition rates were more than 97 %. Therefore, it is confirmed that each of model topologies was selected appropriately and a high recognition rate was obtained.

The proposed algorithm was finalized by selecting the set of the clustered-state triphone HMMs with the highest $\mathcal{F}^m$ value as the optimum acoustic model *without seeing the recognition rate*. Figure 4 shows the $\mathcal{F}^m$ values and the recognition rates along the line connecting the points of the selected topologies in Figure 3, where the horizontal axis is the number of components per state. Figure 4 suggests that the proposed algorithm could work well since the recognition rate and $\mathcal{F}^m$ behaved similarly. Although the recognition rate (**97.9** %) at the highest $\mathcal{F}^m$ value fell short of the top score (98.1 %) on the jagged line in Figure 4, the rates were close enough. Moreover, the rate of 97.9 % was comparable to the top ML rate of 98.0 % that we obtained manually in the preliminary experiments (Section 4). The resultant numbers of states and components per state were 254 and 35, respectively. This combination of numbers was substantially included in the optimum area in Figure 2. Thus, we confirmed that our proposed method can *automatically* determine the *optimum* acoustic model topology.

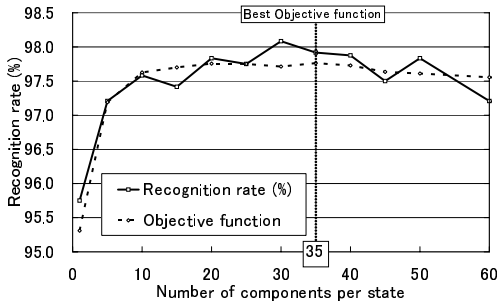

Fig. 4. Recognition rates and objective functions for non-iterative VBEC construction.

6. SUMMARY

In this paper, we realized the automatic determination of the optimum topology in an acoustic model by constructing a Gaussian Mixture Model (GMM)-based phonetic decision tree within a Variational Bayesian Estimation and Clustering for speech recognition (VBEC) framework. Our proposed new approach in the tree construction can calculate the Variational Bayes objective function with a non-iterative algorithm by assuming that each Gaussian per state has the same occupancy and is represented by the same posterior distribution for the covariance. Experiments showed that the obtained method could determine the optimum topology within a practical computation time, and the performance was comparable to the best recognition rate obtained by the conventional maximum likelihood approach with manual tuning. Thus, by using the proposed method, *VBEC can automatically determine an acoustic model topology with the highest performance levels*, enabling us to dispense with manual tuning procedures when constructing acoustic models.

7. 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] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 21, pp. 79–86, 2000.

[4] S. Chen and R. Gopinath, "Model selection in acoustic modeling," in *Proc. Eurospeech*, 1999, vol. 3, pp. 1087–1090.

[5] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, *Application of variational Bayesian approach to speech recognition*, NIPS 15, MIT Press, 2003.

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

[7] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Bayesian acoustic modeling for spontaneous speech recognition," in *Proc. SSPR*, 2003, pp. 47–50.

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