BAYESIAN ACOUSTIC MODELING FOR SPONTANEOUS SPEECH RECOGNITION

Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura and Naonori Ueda

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
watanabe, minami, ats, ueda @cslab.kecl.ntt.co.jp

ABSTRACT

Accurate acoustic model construction for spontaneous speech recognition requires that various speech fluctuation factors such as speaking variations and speaker variances are dealt with. The Bayesian approach has advantages for the speech fluctuation modeling because it enables an appropriate model selection for given speech data, unlike the maximum likelihood approach. However, the Bayesian approach includes complicated integrals that have prevented it from being realized in a large-scale task such as spontaneous speech recognition. In this paper, we apply a practical Bayesian framework: *Variational Bayesian Estimation and Clustering for speech recognition* (VBEC), to spontaneous speech recognition. In particular, we focus on the selection of an appropriate acoustic model structure. The effectiveness of the VBEC is shown through recognition experiments using real spontaneous speech data.

1. INTRODUCTION

Spontaneous speech includes various fluctuation factors, which derive from speaking variations, speaker variances, and so on. To model these fluctuation factors statistically for spontaneous speech recognition, it is important to prepare an appropriate acoustic model structure for given speech data. Figure 1 shows the general view of speech recognition performance for a model structure. If the amount of training data is infinite, recognition performance improves as the model becomes more complex. However, since the amount of training data is actually finite, an overly **complex** model causes over-training and results in a decrease in performance. On the other hand, a model that is **simple** also decreases the performance because the model cannot accurately represent speech fluctuations. Therefore, maintaining the balance between model complexity and training data size is quite important for speech recognition performance.

The Bayesian approach is advantageous in maintaining such balance because it enables not only model posterior estimation, but also appropriate model selection for given training data. However, the Bayesian approach requires complicated integral and expectation computations to obtain posteriors, and has not been applied to large-scale speech recognition.

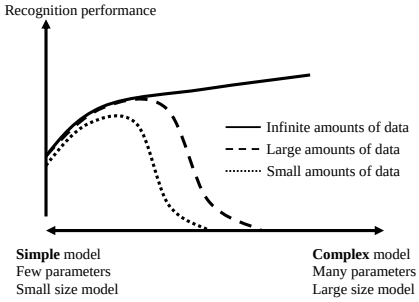


Fig. 1. Model selection and recognition performance according to the amounts of training data.

Recently, a *Variational Bayesian* (VB) approach was proposed, which aims to avoid the complicated computations by using the variational approximation technique [1, 2]. In VB, approximate posteriors (VB posteriors) can be obtained effectively by using EM-like iterative calculations while the advantages of the Bayesian approaches are still retained, unlike the Maximum a posterior (MAP) or Bayesian information criterion (BIC) approaches. In this paper, we apply a practical Bayesian framework: *Variational Bayesian Estimation and Clustering for speech recognition* (VBEC) [3] to spontaneous speech recognition.

The VBEC framework includes estimation of acoustic model posteriors, selection of the appropriate acoustic model and acoustic classification using a Bayesian prediction, each of which is based on the VB approach. In this paper, we focus on the acoustic model selection: clustering triphone HMM states and determining the number of Gaussians per state. The effectiveness of the VBEC is examined through recognition experiments using Japanese spontaneous speech data.

2. VARIATIONAL BAYESIAN APPROACH

In this section, we briefly explain the VB approach. Let $\boldsymbol{O}$ be a given data set. In the Bayesian approach, we are interested in posterior distributions over model parameters, $p(\Theta|\boldsymbol{O}, m)$, and model structure, $p(m|\boldsymbol{O})$. Here, Θ is a set

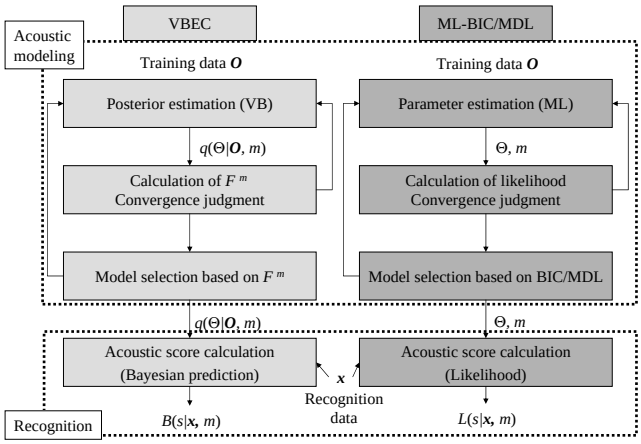


Fig. 2. Total frameworks of speech recognition based on VBEC and ML-BIC/MDL. $B(s|\boldsymbol{x}, m)$ represents a Bayesian prediction score and $L(s|\boldsymbol{x}, m)$ represents a likelihood score of a phoneme category s for recognition data $\boldsymbol{x}$.

of model parameters and m is an index of the model structure. Let us consider a general acoustic modeling with hidden variables. Let S be a set of hidden state sequences and K be a set of Gaussian mixture component sequences. The model with a fixed model structure m can then be defined by the joint distribution $p(\boldsymbol{O}, S, K|\Theta, m)$.

In VB, the variational posteriors $q(\Theta|\boldsymbol{O}, m)$, $q(S, K|\boldsymbol{O}, m)$, and $q(m|\boldsymbol{O})$ are introduced to approximate the true corresponding posteriors. The optimal variational posteriors over Θ, S and K, and the optimal model structure m that maximizes the optimal $q(m|\boldsymbol{O})$ can be obtained by maximizing the following objective function:

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

w.r.t. $q(\Theta|\boldsymbol{O}, m), q(S, K|\boldsymbol{O}, m)$, and m. Here, $\langle f(x)\rangle_{p(x)}$ denotes the expectation of $f(x)$ w.r.t. $p(x)$, and $p(\Theta|m)$ is a prior distribution. This optimization can be effectively performed by an EM-like iterative algorithm (see [1, 2] for the details).

3. TOTAL FRAMEWORK OF SPEECH RECOGNITION BASED ON VBEC

In this section, we clarify the total framework of the VBEC in contrast with a conventional ML-BIC/MDL framework (model parameter estimation and acoustic score calculations are based on maximum likelihood (ML), and model selection is based on the BIC or minimum description length (MDL) criterion). Figure 2 shows a comparative sketch of the two frameworks. Although global flows between them are the same, in VBEC, the objective functions for model

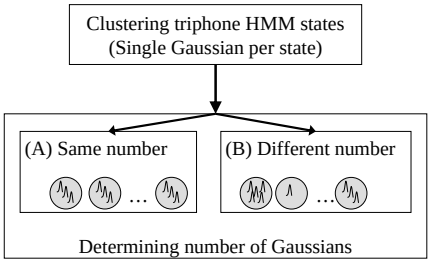


Fig. 3. Acoustic model selection of VBEC

posterior estimation and model selection are the same ($\mathcal{F}^m$), whereas in ML-BIC/MDL, the objective function for model parameter estimation is based on likelihood, and the objective function for model selection is based on BIC/MDL. Therefore, VBEC is consistent in an acoustic model construction because it uses the same objective function $\mathcal{F}^m$.

Moreover, VBEC is based on posterior estimation (VB posterior estimation), whereas ML-BIC/MDL is based on parameter estimation. Therefore, in recognition, the unknown data $\boldsymbol{x}$ are classified to a particular phoneme unit s using the Bayesian predictive posterior $p(s|\boldsymbol{x}, \boldsymbol{O}, m)$, which is obtained by approximating the true posterior with the estimated VB posteriors $q(\theta_s|\boldsymbol{O}, m)$. The optimal classification can be performed by:

$$\begin{aligned} s &= \arg\max_{s'} p(s'|\boldsymbol{x}, \boldsymbol{O}, m) \\ &\approx \arg\max_{s'} \int p(\boldsymbol{x}|s', \theta_{s'}, m) q(\theta_{s'}|\boldsymbol{O}, m) d\theta_{s'} \end{aligned} \quad (2)$$

The integral for a frame can be solved analytically to be Student distributions. Therefore, we can compute a Bayesian prediction score for a frame, and can compute a sequence score by summing up each frame score based on the Viterbi algorithm.

Thus, we can construct a VBEC framework by estimating posteriors, selecting an appropriate model structure and obtaining recognition results using Bayesian prediction classification (BPC), each of which is based on the VB approach.

4. ACOUSTIC MODEL SELECTION OF VBEC

In this section, we focus on an acoustic model selection of VBEC. In acoustic modeling, model selection mainly consists of two processes: clustering triphone HMM states and determining the number of Gaussian per state; the number of combinations of all possible model structures is huge. Moreover, both model selections include hidden variables, and the objective function computations require iterative calculation. Therefore, it is unrealistic to examine $\mathcal{F}^m$ for all possible model structures. An efficient procedure to search the optimal structure is required. One of the solutions is a two-phase procedure that we employ, i.e., firstly clustering

Table 1. Experimental conditions

Sampling rate	16 kHz (Quantization 16 bit)
Feature vector (26 dim.)	12 - order MFCC + Δ MFCC + Energy + Δ Energy
Window	Hamming
Frame size/shift	25/10 ms
# of HMM states	3 (Left to right)
# of phoneme categories	43
Vocabulary size	10,000
Perplexity	152.8

triphone HMM states and then determining the number of Gaussians per state (Figure 3).

The appropriate clustered-state triphone HMM is determined by using the phonetic decision tree algorithm [4]. In VBEC, $\mathcal{F}^m$ is used as the splitting and stop criteria for the tree [3]. To reduce computations, we assume that the frame assignments for states while splitting are fixed, similar to the conventional approach. By using that condition, hidden variables are removed, and all variational posteriors and $\mathcal{F}^m$ can be obtained as closed forms without using an iterative procedure.

Once we have obtained the clustered-state model structure, we determine the number of Gaussians per state and complete the acoustic model selection. In general, there are two approaches to determine the number of Gaussians, as shown in Figure 3. In the first approach (A), the number of Gaussians per state is the same for all clustered states. The objective function $\mathcal{F}^m$ is calculated for each number of Gaussians, and the number of Gaussians, which maximizes $\mathcal{F}^m$, is determined as the appropriate one. In the second approach (B), the number of Gaussians per state is not the same for all clustered states; here, we split and merge the Gaussians to increase $\mathcal{F}^m$ and determine the number of Gaussians in each state.

5. EXPERIMENTS

We conducted four experiments to prove the effectiveness of VBEC for spontaneous speech recognition. The first experiment was designed to compare VBEC with ML-BIC/MDL for constructing clustered-state HMMs. The second experiment was designed to verify how appropriately VBEC model selection could determine the number of Gaussians per state. The third experiment was designed to examine the effectiveness of VBEC for the combination of both acoustic model selections, i.e., the state cluster and number of Gaussians. The above three experiments did not include the BPC-based recognition, enable us to evaluate the sole effectiveness of the VBEC model selection. The fourth experiment was designed to verify the effectiveness of BPC-based recognition. We performed all of the experiments under the conditions shown in Tables 1 and set priors, which is required in the Bayesian approach, using monophone HMM state statistics (e.g., mean and covariance). The total training data for acoustic model consisted of 88 (10 hours) and the test data consisted of 4 (1.5 hours) lectures (Table 2), spoken by males in the corpus of spontaneous Japanese (CSJ) [5]. We used the standard trigram language model provided by the CSJ monitor version (trained by 186 lectures).

5.1. Model selection for clustered-state HMMs

In this experiment, we compared VBEC with ML-BIC/MDL for constructing clustered-state HMMs. We kept a single Gaussian per state in both frameworks to evaluate the model structure obtained in the triphone HMM state clustering.

Table 2. Recognition results of VBEC and ML-BIC/MDL

	# of clustered states	Word accuracy
VBEC	4,281	52.3
ML-BIC/MDL	5,386	51.7

Table 2 compares the number of clustered states in the triphone HMM state clustering and the recognition results obtained by VBEC with ML-BIC/MDL. VBEC selected a smaller model structure and performed better than ML-BIC/MDL. In this case, the model structure obtained by ML-BIC/MDL was large, which caused over-training and reduced the performance. From the result, we confirmed that VBEC is effective for triphone HMM state clustering in spontaneous speech recognition.

5.2. Model selection for Gaussian mixtures

In this experiment, from the single Gaussian model obtained by the previous experiment, we increased the number of Gaussians per state with the two approaches mentioned in Section 4: the number of Gaussians per state is (A) the same, or (B) different.

Figure 4 shows the total value of the objective function $\mathcal{F}^m$ and the word accuracy for each number of Gaussians obtained by (A). As the number of Gaussians increased, the word accuracy also increased at first, then decreased gradually due to the over-training effect, similar to the general performance for the model structure given in Figure 1. The total value of $\mathcal{F}^m$, then, behaved similarly to the word accuracy. This shows that VBEC can determine the number of Gaussians appropriately. The selected model structure had 4,281 states and **12,843** Gaussians (3 Gaussians per state) and the word accuracy was **55.4** in approach (A). Moreover, the selected model structure had **18,210** Gaussians and the word accuracy was **56.0** in approach (B). Although both fell short of the maximum value (56.4) in Figure 4, they were close enough for practical purposes.

5.3. Acoustic model selection over clustered states and Gaussian mixtures

In this experiment, we first examined the performance of the acoustic model structure on the numbers of clustered states

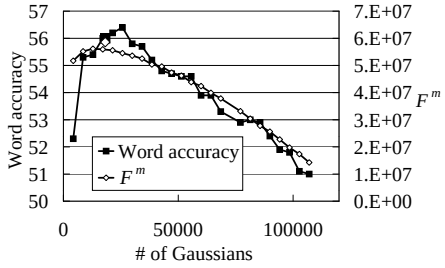
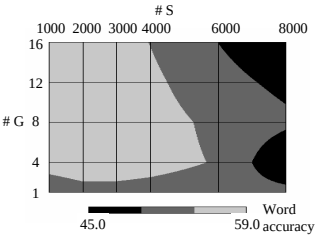


Fig. 4. Number of total Gaussians vs. word accuracy and objective function $\mathcal{F}^m$.

Table 3. Word accuracies for the numbers of clustered states (# S) and Gaussians per state (# G). The contour graph on the right is obtained by these results.

# G	# S					
	1000	2000	3000	4000	6000	8000
16	58.3	58.3	56.3	54.2	49.6	45.5
12	58.3	58.4	56.6	55.3	51.4	48.4
8	57.7	58.2	57.0	56.1	53.3	50.7
4	55.9	56.7	56.6	55.9	54.0	45.2
1	51.9	53.0	53.0	52.7	51.8	51.2



and Gaussians per state by using the ML approach, i.e., controlling the number of clustered states by using the likelihood difference threshold, and then increasing the number of Gaussians in each state. The performance was compared with the VBEC results in Section 5.2.

Table 3 shows the results of word accuracies over the number of clustered states and Gaussians per state. These results were used to construct the contour graph on the right hand side. From Table 3, good results (≥ 55.0) were obtained for 1,000 to 4,000 clustered states and more than 4 Gaussians per state. On the other hand, the model structure obtained by approach (A) in Experiment 5.2 had 4,281 states and 3 Gaussians per state, which is outside of this range. The main reason for this presumably comes from the two-phase procedure of the model selection, i.e., firstly clustering triphone HMM states and then determining the number of Gaussians per state, where model structure is only locally optimized at each phase. As for the recognition performance, although both approaches (A) and (B) figured good word accuracies (55.4 and 56.0, respectively), they could not reach the best performance given in Table 3. The local optimization of the model structure certainly affected the performance. Therefore, VBEC performance is expected to rise by improving the selection procedure so as to optimize model structure globally.

5.4. Effects of Bayesian prediction classification

We examined the effectiveness of total VBEC framework by carrying out the BPC-based recognition using acoustic models obtained in Experiment 5.2. Table 4 shows the word accuracy for varying numbers of Gaussians per state. The BPC outperformed the ML-based classification, especially in the case of the large numbers of Gaussians per state. It is clear that introducing the Bayesian approach into both modeling and recognition makes the whole framework consistent and leads to improvement in recognition performance.

Table 4. The effects of BPC for the numbers of Gaussians per state (# G)

# G	1	2	3	4	5	6	7	8
BPC	52.3	55.2	55.8	56.4	56.4	56.9	56.7	56.8
No BPC	52.3	55.3	55.4	56.0	56.2	56.4	55.8	55.7
# G	9	10	11	12	13	14	15	16
BPC	56.8	56.9	57.1	56.7	56.7	56.6	56.7	56.5
No BPC	55.2	54.8	54.7	54.6	54.6	53.9	53.9	53.3

6. SUMMARY

We applied VBEC to spontaneous speech recognition. The experiments show the effectiveness of VBEC model selection for clustering triphone HMM states and for determining the number of Gaussians per state, respectively. The combination of both model selections enabled good recognition performance, and the performance is expected to rise by improving the selection procedure so as to optimize the state clusters and the numbers of Gaussians simultaneously. The effectiveness of the BPC is also confirmed in the experiment, especially when dealing with large numbers of Gaussians per state. From these results, we can conclude that VBEC is effective for spontaneous speech recognition.

7. ACKNOWLEDGEMENT

We thank the Japanese Science and Technology Agency Priority Program, “Spontaneous Speech: Corpus and Processing Technology,” for providing speech data and transcriptions.

8. REFERENCES

[1] H. Attias, “Learning Parameters and Structure of Latent Variable Models by Variational Bayes,” *Proc. Uncertainty in Artificial Intelligence*, (1999).

[2] N. Ueda and Z. Ghahramani, “Bayesian Model Search for Mixture Models Based on Optimizing Variational Bounds,” *Neural Networks*, vol. 15, pp. 1223-1241, (2002).

[3] S. Watanabe, et al., “Application of Variational Bayesian Approach to Speech Recognition,” *NIPS’02, MIT Press*.

[4] J. Odell, “The Use of Context in Large Vocabulary Speech Recognition,” PhD thesis, *Cambridge University*, (1995).

[5] S. Furui, et. al., “Toward the Realization of Spontaneous Speech Recognition –Introduction of a Japanese Priority Program and Preliminary Results–,” *Proc. ICSLP’00*, vol. 3, pp. 518-521, (2000).