APPLICATION OF VARIATIONAL BAYESIAN ESTIMATION AND CLUSTERING TO ACOUSTIC MODEL ADAPTATION

Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura and Naonori Ueda

Speech Open Laboratory, 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

In this paper, we apply *Variational Bayesian Estimation and Clustering for speech recognition* (VBEC) to an acoustic model adaptation. VBEC can estimate parameter posteriors even when a model includes hidden variables, by using Variational Bayesian approach. In addition, VBEC can select an appropriate model structure in clustering triphone states, according to the amount of available adaptation data. Unlike a conventional Bayesian method such as Maximum A Posteriori (MAP), VBEC is useful even in the case of small amounts of data, because the amount of data per one Gaussian increases due to the model structure selection, and over-training is suppressed. We conduct an off-line supervised adaptation experiment on isolated word recognition, and show the advantage of the proposed method over the conventional method, especially when dealing with small amounts of adaptation data.

1. INTRODUCTION

Acoustic model adaptation techniques have been widely studied and used for practical problems in speech recognition [1] [2]. The adaptation can reduce modeling mismatches in speakers, speaking style, speaking environment and so on, by only using a limited amount of adaptation data. One of the most popular approaches of the adaptation is based on Maximum A Posteriori (MAP) estimation [1]. MAP is a Bayesian approach, and estimates acoustic model parameters while taking into account the effect of prior statistics. Acoustic model parameters estimated by MAP theoretically stay close to the prior statistics with small amounts of data, and approach the statistics estimated by Maximum Likelihood (ML) with large amounts of data. For instance, in the case that we adapt an acoustic model to a particular speaker, we often regard Speaker Independent (SI) HMM statistics (such as mean and covariance), that are computed with sufficient amount of data in advance, as priors. The MAP adaptation suppresses over-training problems even with small amounts of data because the resulting performance is guaranteed by the SI HMM statistics.

The MAP adaptation, however, suppresses only the move of parameter values in the case of small amounts of data, with the model structure left as it is, and provides only limited effectiveness. If the model size becomes small in the case of small amounts of data, we can increase the amount of data per model parameter. The effectiveness of the adaptation is expected to improve by selecting a model structure appropriately according to the amount of adaptation data.

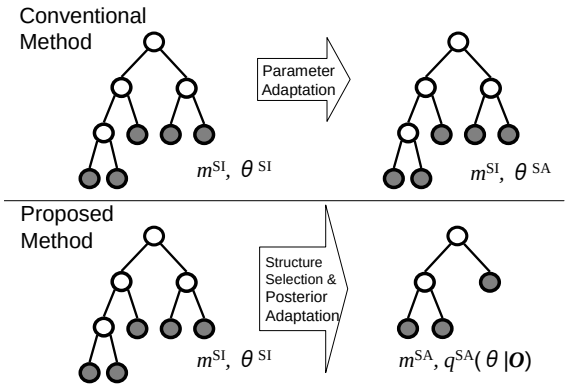

Fig. 1. Basic idea of a model structure selection and parameter (posterior) adaptation to Speaker Adapted (SA) model. Here, m is an index of a model structure, θ is a set of model parameters and $q(\theta|\boldsymbol{O})$ is a posterior for θ. In this figure, the model structure is expressed as a tree.

To achieve the function, we propose a new acoustic model adaptation based on *Variational Bayesian Estimation and Clustering for speech recognition* (VBEC) [3]. VBEC can estimate parameter posteriors even when a model includes hidden variables, by using Variational Bayesian (VB) approach [4]. Also, VBEC can select a model structure automatically, according to the amount of available adaptation data [5] [6]. Therefore, in the proposed adaptation, not only parameter posteriors, but also the model structure, are changed as the amount of adaptation data increases. Figure 1 gives a graphical representation.

2. MAP APPROACH

In sections 2 and 3, we explain two Bayesian adaptations based on MAP and VBEC. Let $\boldsymbol{O} = \{\boldsymbol{o}^t \in \mathcal{R}^D : t = 1, ..., T\}$ be a set of D dimensional feature vectors for a phoneme unit, and $p(\boldsymbol{O}|\theta_j, m)$ be a probability distribu-

tion with a set of parameters θ_j for an HMM state j. In this paper, we formalize a single Gaussian model per one HMM state. Therefore, the index of a Gaussian mixture component is not required. Here, m denotes an index of a model structure and is regarded as a random variable. In the Bayesian approach, we can calculate acoustic scores and output the optimal classification results if we obtain $p(\theta_j|\boldsymbol{O}, m)$. Therefore, obtaining parameter posteriors $p(\theta_j|\boldsymbol{O}, m)$ is important. However, if the acoustic model includes hidden variables such as sequences of hidden states and Gaussian mixture components, it is difficult to obtain posteriors analytically because the calculation includes complicated multivative integrals.

The MAP approach avoids them by estimating a parameter $\hat{\theta}_j$ from $p(\boldsymbol{O}|\theta_j, m)$ and a prior $p(\theta_j|\theta_j^0, m)$, instead of calculating the posteriors directly. This equation is shown as follows:

$$\hat{\theta}_j = \underset{\theta_j}{\operatorname{argmax}}\, p(\boldsymbol{O}|\theta_j, m)p(\theta_j|\theta_j^0, m), \tag{1}$$

where θ_j^0 is a set of hyper-parameters. This procedure is known as a *point estimation*, and can be carried out by iterative calculations based on the Baum-Welch or Viterbi algorithm, even if an acoustic model includes hidden variables. Although an ML estimation is also a point estimation, the difference between MAP and ML estimations is whether priors are included. The MAP approach can carry out adaptation training so that the prior statistics approach ML statistics as the amount of adaptation data increases. If we estimate $\hat{\theta}_j$, the optimal classification can be performed by

$$\hat{j} = \underset{j}{\operatorname{argmax}}\, p(\boldsymbol{x}|\hat{\theta}_j, m). \tag{2}$$

3. VBEC APPROACH

The VB approach avoids the integrals by using a variational approximation. Let $q(\theta_j|\boldsymbol{O}, \theta_j^0, m)$ be an arbitrary distribution over a parameter θ_j conditioned by $\boldsymbol{O}, \theta_j^0, m$, and consider the Kullback-Leibler distance between $q(\theta_j|\boldsymbol{O}, \theta_j^0, m)$ and $p(\theta_j|\boldsymbol{O}, \theta_j^0, m)$. Then, we can obtain an inequality as follows:

$$\mathrm{KL}[q|p] \leq \log p(\boldsymbol{O}|\theta_j^0, m) - \mathcal{F}^m[q], \tag{3}$$

where $\mathcal{F}^m$ is a variational objective function defined as follows:

$$\mathcal{F}^m[q] = \left\langle \log \frac{p(\boldsymbol{O}|\theta_j, m)p(\theta_j|\theta_j^0, m)}{q(\theta_j|\boldsymbol{O}, \theta_j^0, m)} \right\rangle_{q(\theta_j|\boldsymbol{O}, \theta_j^0, m)}. \tag{4}$$

Here, $\langle u(y)\rangle_{p(y)}$ represents the expectation of $u(y)$ with respect to the distribution $p(y)$. From equation (3), if the right hand side is small, q approaches p, meaning the appropriate posteriors $\tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, m)$ for a fixed m can be estimated by maximizing $\mathcal{F}^m$ with respect to $q(\theta_j|\boldsymbol{O}, \theta_j^0, m)$ based on a variational method. This procedure is known as a *distribution estimation* because distributions, not parameters, are estimated. The distribution estimation can be carried out by iterative calculations, even if an acoustic model includes hidden variables [4].

Assuming that priors are conjugate distributions, the appropriate posteriors for the mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix Σ can be estimated as follows:

$$\begin{aligned} \tilde{q}(\boldsymbol{\mu}_j|\boldsymbol{O}, \theta_j^0, m) &= \mathcal{N}(\boldsymbol{\mu}_j|\tilde{\boldsymbol{\nu}}_j, \tilde{\xi}_j^{-1}\Sigma_j^2) \\ \tilde{q}((\Sigma_j^d)^{-2}|\boldsymbol{O}, \theta_j^0, m) &= \mathcal{G}((\Sigma_j^d)^{-2}|\tilde{\eta}_j, \tilde{R}_j^d), \end{aligned} \tag{5}$$

where d denotes a component of a D dimensional vector, $\mathcal{N}$ denotes a Gaussian distribution and $\mathcal{G}$ denotes a gamma distribution. In this paper, we omit results of posteriors over state transition parameters and weight factor parameters of Gaussian mixtures because we only adapt the mean and covariance. Here, $\tilde{\theta}_j \equiv \{\tilde{\boldsymbol{\nu}}_j, \tilde{\xi}, \tilde{\eta}, \tilde{R}_j\}$ is a set of posterior parameters defined as:

$$\begin{aligned} &\tilde{\xi}_j = \xi^0 + \textstyle\sum_{t=1}\tilde{\zeta}_j^t, \quad \tilde{\eta}_j = \eta^0 + \textstyle\sum_{t=1}\tilde{\zeta}_j^t, \\ &\tilde{\boldsymbol{\nu}}_j = \left(\xi^0\boldsymbol{\nu}_j^0 + \textstyle\sum_{t=1}^T \tilde{\zeta}_j^t\boldsymbol{o}^t\right)/(\xi^0 + \textstyle\sum_{t=1}\tilde{\zeta}_j^t), \\ &\tilde{R}_j^d = R_j^{0,d} + \xi^0(\nu_j^{0,d} - \tilde{\nu}_j^d)^2 + \textstyle\sum_{t=1}^T \tilde{\zeta}_j^t(o^{t,d} - \tilde{\nu}_j^d)^2, \end{aligned}$$

where $\tilde{\theta}_j$ is composed of θ_j^0 and an occupation probability defined as $\tilde{\zeta}_j^t \equiv \tilde{q}(s^t = j|\boldsymbol{O}, \theta_j^0, m)$. Furthermore, s^t denotes a state at time t and $\tilde{\zeta}_j^t$ denotes the probability of occupation in state j at time t. We focus on the $\tilde{\boldsymbol{\nu}}_j$. If the adaptation data are small, $\tilde{\boldsymbol{\nu}}_j$ stays close to $\boldsymbol{\nu}_j^0$, while on the other hand, if the adaptation data are large, $\tilde{\boldsymbol{\nu}}_j$ approaches the mean of the adaptation data, $\sum_{t=1}^T \tilde{\zeta}_j^t\boldsymbol{o}^t / \sum_{t=1}^T \tilde{\zeta}_j^t$. In other words, the peak of the mean posterior moves from the prior mean to the ML mean as the amount of adaptation data increases. Therefore, the VB approach can also be used to carry out adaptation training.

Moreover, using the VB approach, we can select an appropriate model structure. Similar to parameter posterior estimation, we can obtain an inequity between an arbitrary distribution $q(m|\theta_j^0, \boldsymbol{O})$ over a model structure m conditioned by $\theta_j^0, \boldsymbol{O}$, and model posterior $p(m|\theta_j^0, \boldsymbol{O})$ as follows:

$$\mathrm{KL}[q|p] \leq \log p(\boldsymbol{O}|\theta_j^0) + \left\langle \log \frac{q(m|\theta_j^0, \boldsymbol{O})}{p(m|\theta_j^0)} - \mathcal{F}^m \right\rangle_q \tag{6}$$

Assuming the model prior $p(m|\theta_j^0)$ as a constant, and maximizing $\mathcal{F}^m$ with respect to q based on a variational method, we can select an appropriate model structure $\tilde{m}$ as follows [5] [6]:

$$\tilde{m} = \underset{m}{\operatorname{argmax}}\, \tilde{q}(m|\theta_j^0, \boldsymbol{O}) = \underset{m}{\operatorname{argmax}}\, \mathcal{F}^m, \tag{7}$$

where $\tilde{q}(m|\theta_j^0, \boldsymbol{O})$ denotes the appropriate variational posterior over the model structure m. Equation (7) and the obtaining of $\tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, m)$ indicate that by maximizing $\mathcal{F}^m$

not only with respect to $q(\theta_j|\boldsymbol{O}, \theta_j^0, m)$, but also m, we can obtain the appropriate parameter posteriors, $\tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, \tilde{m})$ in the appropriate model structure.

In recognition, if $\tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, \tilde{m})$ is estimated, the optimal classification can be performed by

$$\tilde{j} = \underset{j}{\text{argmax}} \int d\theta_j p(\boldsymbol{x}|\theta_j, \tilde{m}) \tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, \tilde{m}) \qquad (8)$$

This equation is based on the Bayesian Prediction Classification (BPC). The effectiveness of BPC in an acoustic model adaptation has been proved in [7].

We name the speech recognition framework based on the Bayesian approach as VBEC, whose effectiveness has already been proved in speech recognition experiments [3].

Thus, VBEC features the mechanism of model structure selection and parameter posterior estimation, both of which are expected to further suppress the effect of over-training in acoustic model adaptation.

4. ACOUSTIC MODEL ADAPTATION SCHEME USING VBEC

Figure 2 shows the off-line supervised acoustic model adaptation scheme based on the VBEC. The VBEC-based adaptation consists of two phases: model structure selection and parameter posterior estimation.

In this paper, the model structure is the state-clustering structure in a set of triphone-HMMs. Hierarchical state-clustering structures can be obtained by successive clustering of HMM states using phonetic decision tree [8]. In VBEC, we use $\mathcal{F}^m$ as described in section 3, instead of likelihood or MDL, to measure the quality of a model structure. Consequently, due to the nature of the VB approach, the appropriate structure can be selected from the hierarchical structures in the tree without heuristic tuning to the amount of training data [9], unlike the conventional approach. Taking advantage of this, we can adapt the model structure by selecting a structure $\tilde{m}$ appropriately for a given amount of adaptation data. Once the model structure has been adapted, we carry out the posterior estimation by the iterative calculation until $\mathcal{F}^m$ converges, thereby obtaining the appropriate variational posterior $\tilde{q}(\theta_j|\boldsymbol{O}, \theta_j^0, \tilde{m})$. Thus, a VBEC-adapted model is produced.

In speech recognition using the VBEC-adapted model, we calculate acoustic scores based on the BPC approach, so that Bayesian approaches are consistently employed in both stages of adaptation and recognition.

5. EXPERIMENT

We conducted an experiment to evaluate the effectiveness of the proposed VBEC adaptation. The experiment compares how VBEC and MAP work with variable amounts of adaptation data. We performed the experiments under the conditions shown in Tables 1 and 2, and kept a single Gaussian per state, mainly to evaluate the effect of the model structure selection.

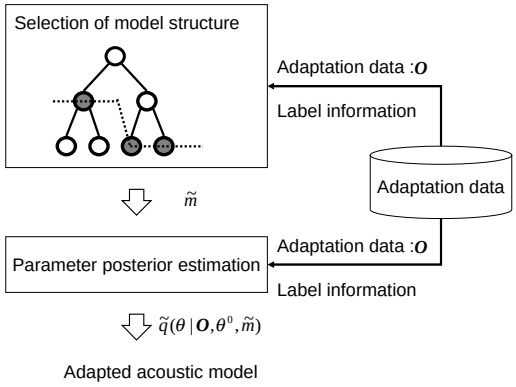

Fig. 2. Acoustic model adaptation scheme using VBEC

The prior-training, adaptation and recognition data used in these experiments are shown in Table 3. The total training data for priors consisted of about 3,000 Japanese sentences spoken by 30 males; these sentences were designed so that the phonemic balance was maintained. The total adaptation and recognition data consisted of 1,200 Japanese words spoken by one male who is not included in the prior-training data. The experiment was, thus, designed to evaluate the VBEC and MAP adaptation approaches by reducing modeling mismatches in speaker and speaking style, i.e., adaptation of the speaker-independent model for sentence utterance to the speaker-specified model for word utterance. We divided a set of isolated word data into two data sets: adaptation and recognition data. The total adaptation data consisted of 1,000 words, while other words were assigned to the recognition data. There were 5,000 word candidates in the task. Several subsets were randomly extracted from the adaptation data set, and each of these subsets was used to construct a set of adapted acoustic models. As a result, about 20 sets of adapted acoustic models for several amounts of adaptation data were prepared.

For prior training, we constructed hierarchical state-clustering structures of speaker-independent triphone HMMs by phonetic decision tree [8] using conventional ML criterion. All required prior statistics for the MAP and VBEC adaptations were kept in these structures. Figure 3 compares the recognition results obtained by the MAP and VBEC methods for several amounts of adaptation data. The recognition result obtained by the SI model is also shown in the figure. Both the SI and MAP model structures were the

Table 1. Acoustic Conditions

Sampling Rate	16 kHz (Quantization 16 bit)
Feature Vector	12 - order MFCC with Δ MFCC
Window	Hamming
Frame Size/Shift	25/10 ms

Table 2. Prepared HMM

# of States	3 (Left to Right)
# of Phoneme Categories	27
Output Distribution	Single Gaussian

same, and there were 2,000 clustered states. Although we determined the model structure so that the SI model achieved the best recognition rate, the SI model's performance was rather low because there were modeling mismatches in both speaker and speaking style. Figure 4 shows the number of clustered states (i.e. model structure) obtained in the structure selection. From 10 words to 30 words, the performance of MAP was worse than that of SI because the small amount of adaptation data strongly affected over-training on the model, and the MAP that leaves the model structure as it is could not suppress them. On the other hand, in the same data size area, VBEC method's performance was better than that of SI excluding the 10 words result, and was better than MAP by 10 percents from 10 words to 50 words, due to its suppression of the over-training effect due to the model selection and parameter posterior estimation. In fact, as shown in Figure 4, the number of clustered states selected by the VBEC method was much smaller than 2,000, and the amount of adaptation data per Gaussian was much larger than that of the MAP method. Moreover, the resulting performances of the VBEC method are almost identical to those of the MAP method for data of more than 100 words, although the selected number of clustered states was still much smaller than 2,000. In summary, the proposed VBEC adaptation performed similarly to or better than the MAP adaptation under all conditions of adaptation data size, with a much more compact model structure. In particular, when the adaptation data size was small, VBEC adaptation outperformed MAP adaptation.

In addition, the VBEC method is expected to take less decoding time than the MAP method due to its very compact model structure. In fact, the VBEC method took about half as much time as the MAP method when processing 100 words of data, although it includes VB and BPC calculations that are more complicated than likelihood calculations.

6. SUMMARY

We proposed a new acoustic model adaptation based on VBEC. The method can select an appropriate model structure and adapt parameter posteriors within a VB framework.

Table 3. Prior-Training, Adaptation and Recognition Data

Prior-Training	Continuous Sentences (ASJ)
Adaptation/Recognition	Isolated Words (ATR)

ASJ: Acoustical Society of Japan

ATR: Advanced Telecommunications Research Institute International

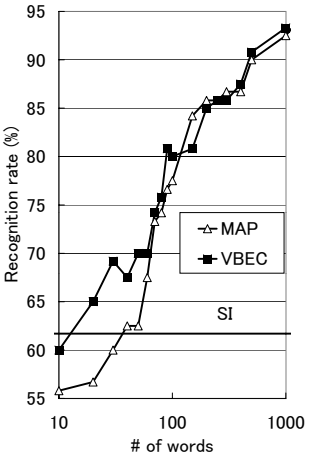

Fig. 3. Recognition results with varying adaptation data.

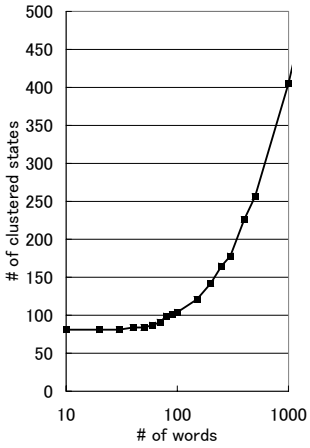

Fig. 4. Number of clustered states in VBEC with varying adaptation data.

The experimental result shows that our proposed method is superior to the MAP method. We are going to extend our method to a Gaussian mixture model in the near future.

7. REFERENCES

[1] J-L. Gauvain and C-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, (1994).

[2] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171-185, (1995).

[3] S. Watanabe, et al. "Application of Variational Bayesian Approach to Speech Recognition," *NIPS'02, MIT Press*. (to appear)

[4] S. Waterhouse et al., "Bayesian Methods for Mixture of Experts," *NIPS'95, MIT Press*, pp. 351-357, (1995).

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

[6] N. Ueda and Z. Ghahramani, "Optimal Model Inference for Bayesian Mixture of Experts," *Proc. NNSP'00*, pp. 145-154, (2000).

[7] Q. Huo and C-H. Lee, "Combined On-Line Model Adaptation and Bayesian Predictive Classification for Robust Speech Recognition," *Proc. Eurospeech'97*, pp. 1847-1850, (1997).

[8] J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD thesis, Cambridge University, (1995).

[9] S. Watanabe, et al., "Constructing Shared-State Hidden Markov Models Based on a Bayesian Approach," *Proc. ICSLP'02*, vol. 4, pp. 2669-2672, (2002).