FEATURE SPACE VARIATIONAL BAYESIAN LINEAR REGRESSION AND ITS COMBINATION WITH MODEL SPACE VBLR

Seong-Jun Hahm, Atsunori Ogawa, Marc Delcroix, Masakiyo Fujimoto, Takaaki Hori, and Atsushi Nakamura*

NTT Communication Science Laboratories, NTT Corporation, Kyoto, 619-0237, Japan

{seongjun.hahm, ogawa.atsunori, marc.delcroix, fujimoto.masakiyo, hori.t, nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

In this paper, we propose a *tuning-free Bayesian linear regression* approach for speaker adaptation. We first formulate feature space variational Bayesian linear regression (fVBLR). Using a lower bound as the objective function, we can optimize a binary tree structure and control parameters for prior density scaling. We experimentally verified the proposed fVBLR could achieve performance comparable to that of the conventional fine-tuned fSMAPLR and SMAPLR. For further performance improvement regardless of the amount of adaptation data, we combine fVBLR with model space VBLR (fVBLR+VBLR). Therefore, feature space normalization and model space adaptation are *consistently* performed based on a variational Bayesian approach without any tuning parameters. In the experiment, the proposed fVBLR+VBLR showed performance improvement compared with both fVBLR and VBLR.

Index Terms— speaker adaptation, fSMAPLR, SMAPLR, fVBLR, VBLR

1. INTRODUCTION

Linear regression approaches, such as maximum likelihood linear regression (MLLR), have been widely used for speaker adaptation [1–8]. To achieve effective linear regression, a binary tree structure of Gaussian clusters was proposed [1]. Using the binary tree structure that is optimized with the appropriate occupancy threshold, estimation of the transformation matrix is effectively performed regardless of the amount of adaptation data. However, for small amount of adaptation data, the performance of MLLR drops severely. Maximum a posteriori linear regression (MAPLR) was proposed to stabilize estimation of transformation matrix especially for small amount of adaptation data by incorporating prior distribution [5].

The prior distributions can be obtained from the parent nodes in a binary tree structure [6–9]. These methods include structural maximum a posteriori linear regression (SMAPLR) [6] and feature space SMAPLR (fSMAPLR, also known as Constrained SMAPLR;CSMAPLR) [8]. The above mentioned MAP-based linear regression approaches set the occupancy threshold for the tree structure and the control parameter that controls the contribution of the prior density (i.e., transformation matrix of the parent node in a binary tree). These two parameters are usually determined empirically using the development set [6, 10]. Furthermore, once the control parameter is set, every node uses the same control parameter for scaling prior distribution (see Fig. 1(a) in Section 4.2). Using this control parameter, we cannot appropriately control the contribution of the prior distribution of each node because the amount of adaptation data assigned to each node is different from node to node.

To solve these problems in parameter tuning, VBLR approach has recently been proposed as a *tuning-free* SMAPLR approach [11]. VBLR is a *fully Bayesian* treatment of linear regression for hidden Markov models (HMMs). VBLR analytically derives the variational lower bound of the marginalized log-likelihood (evidence). By using the variational lower bound as an objective function, we can optimize the tree structure and control parameter of the linear regression without controlling them as tuning parameters. The tree structure is automatically determined according to the amount of adaptation data[1] and we can determine the appropriate control parameters for scaling the prior of each node separately.

In this work, we expand the VBLR approach to feature space. We formulate a *feature space VBLR* approach (fVBLR; Section 2) that has all the advantages of model space VBLR (Section 3). We also use the obtained lower bound as the objective function (Section 2.3). Using the lower bound, we can optimize the tree structure and control parameters for prior density scaling (Section 2.4). For further performance improvement regardless of the amount of adaptation data, we combine fVBLR and model space VBLR (fVBLR+VBLR). This approach is an extension of our previous approach that employs MAP-based adaptation [13]. We also demonstrated the effectiveness of combination of fSMAPLR and model space VBLR [14]. In this paper, we evaluate the effectiveness of the proposed methods (fVBLR, fVBLR+VBLR) for speaker adaptation for a large vocabulary speech recognition (Section 4).

2. FEATURE SPACE VARIATIONAL BAYESIAN LINEAR REGRESSION

Variational Bayesian approaches estimate the entire posterior distribution of the parameters and latent variables than a single most probable value (point estimation) for generalizing MAP-based approaches. In this section, we formulate feature space variational Bayesian linear regression (fVBLR). In the model space VBLR [11], the variational lower bound is analytically derived by using conjugate distributions as prior distributions, and by assuming the conditional independence on the posterior distributions. The marginalized log-likelihood with a set of hyperparameters $\mathbf{\Psi}$ and a model (tree) structure m is represented by

$$\begin{aligned}&\ln p(\boldsymbol{O}|\mathbf{\Psi}, m)\\ &= \ln \int q(\boldsymbol{W}, \boldsymbol{S}) \frac{p(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{W}|\mathbf{\Psi}, m)}{q(\boldsymbol{W}, \boldsymbol{S})} d\boldsymbol{W} d\boldsymbol{S},\end{aligned} \tag{1}$$

*He is now at Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A. (seongjun.hahm@utdallas.edu).

[1]Using VB approach, model structure itself can also be estimated [12]. In this paper, we only determined the depth of the tree based on lower bound.

where $\boldsymbol{O}$ is the feature vector set, $\boldsymbol{W}$ is the extended feature space transformation matrix, and $\boldsymbol{S}$ represents the sequences of HMM states and mixture components of Gaussian mixture models. Instead of direct simplification of the above marginal log-likelihood, we constrain the posterior to be a simpler, factorized approximation $q(\boldsymbol{W}, \boldsymbol{S}) \approx q(\boldsymbol{W})q(\boldsymbol{S})$ [15].

The lower bound of the marginalized likelihood is represented by using Jensens inequality

$$\begin{aligned}
&\ln p(\boldsymbol{O}|\boldsymbol{\Psi}, m) \\
&= \ln \int q(\boldsymbol{W})q(\boldsymbol{S}) \frac{p(\boldsymbol{O}, \boldsymbol{S}|\boldsymbol{W})p(\boldsymbol{W})}{q(\boldsymbol{W})q(\boldsymbol{S})} d\boldsymbol{W} d\boldsymbol{S} \\
&\geq \underbrace{\left\langle \ln \frac{p(\boldsymbol{O}, \boldsymbol{S}|\boldsymbol{W})p(\boldsymbol{W})}{q(\boldsymbol{W})q(\boldsymbol{S})} \right\rangle_{\substack{q(\boldsymbol{W}) \\ q(\boldsymbol{S})}}}_{\triangleq \mathcal{G}(\boldsymbol{\Phi}, m)},
\end{aligned} \tag{2}$$

where $p(\boldsymbol{W})$ is a prior distribution of $\boldsymbol{W}$, and $q(\boldsymbol{W})$ and $q(\boldsymbol{S})$ are arbitrary distributions. In the above equations (1) and (2), we omitted $\boldsymbol{\Psi}, m$ in $p(\boldsymbol{W}|\boldsymbol{\Psi}, m)$, $q(\boldsymbol{W}|\boldsymbol{\Psi}, m)$, and $q(\boldsymbol{S}|\boldsymbol{\Psi}, m)$ for simplicity.

The variational lower bound defined in Eq. (2) can be decomposed as follows:

$$\mathcal{G}(\boldsymbol{\Psi}, m) = \underbrace{\left\langle \ln \frac{p(\boldsymbol{O}, \boldsymbol{S}|\boldsymbol{W})p(\boldsymbol{W})}{q(\boldsymbol{W})} \right\rangle_{\substack{q(\boldsymbol{W}) \\ q(\boldsymbol{S})}}}_{\triangleq \mathcal{L}(\boldsymbol{\Psi}, m)} - \langle \ln q(\boldsymbol{S}) \rangle_q (\boldsymbol{S}). \tag{3}$$

From Eq. (3), we can take only the first logarithmic evidence term for $\boldsymbol{\Psi}$ and m because the second term does not depend on the transformation matrix $\boldsymbol{W}$. Considering the conditional independence assumption over cluster r, node (cluster) index of the binary tree, $\mathcal{L}(\boldsymbol{\Psi}, m)$ can be represented by

$$\mathcal{L}(\boldsymbol{\Psi}, m) = \sum_r \left\langle \ln \frac{p(\boldsymbol{O}, \boldsymbol{S}|\boldsymbol{W}_r)p(\boldsymbol{W}_r)}{q(\boldsymbol{W}_r)} \right\rangle_{\substack{q(\boldsymbol{W}_r) \\ q(\boldsymbol{S})}}. \tag{4}$$

2.1. Conjugate distribution for prior distribution

Similar to model space VBLR, fVBLR also adopts the conjugate distributions as prior distributions to obtain an analytical solution. The matrix normal distribution (prior distribution) is defined as follows:

$$\begin{aligned}
&p(\boldsymbol{W}) \propto \mathcal{N}(\boldsymbol{W}|\boldsymbol{C}, \boldsymbol{\Phi}, \boldsymbol{V}) \\
&= \prod_r \frac{\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{V}_r^{-1}(\boldsymbol{W}_r - \boldsymbol{C}_r)'\boldsymbol{\Phi}^{-1}(\boldsymbol{W}_r - \boldsymbol{C}_r)\right]\right\}}{(2\pi)^{n(n+1)/2}|\boldsymbol{V}_r|^{n/2}|\boldsymbol{\Phi}|^{(n+1)/2}},
\end{aligned} \tag{5}$$

where $\boldsymbol{C}$, $\boldsymbol{V}$, and $\boldsymbol{\Phi}$ are the hyperparameters for that distribution family. $\boldsymbol{C}$ is the $n \times (n+1)$ location matrix and $\boldsymbol{V}$ is the $(n+1) \times (n+1)$ scaling matrix. To obtain a simple solution for the final analytical solution, we set the following constraints on $\boldsymbol{\Phi}$ and $\boldsymbol{V}_r$ [5, 7, 11]:

$$\begin{aligned}
\boldsymbol{\Phi} &\approx \boldsymbol{I}_n \\
\boldsymbol{V}_r &\approx \rho_r^{-1}\boldsymbol{I}_{n+1}
\end{aligned}, \tag{6}$$

where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and ρ_r indicates a feature space control parameter.

By substituting Eq. (6) into Eq. (5), we obtain the following:

$$\begin{aligned}
&\mathcal{N}(\boldsymbol{W}_r|\boldsymbol{C}_r, \boldsymbol{I}_n, \rho_r^{-1}\boldsymbol{I}_{n+1}) \\
&= \left(\frac{\rho_r}{2\pi}\right)^{\frac{n(n+1)}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\rho_r(\boldsymbol{W}_r - \boldsymbol{C}_r)'(\boldsymbol{W}_r - \boldsymbol{C}_r)\right]\right).
\end{aligned} \tag{7}$$

2.2. Posterior distribution of transformation matrix

From the variational calculation for $\mathcal{G}(\boldsymbol{\Psi}, m)$ with respect to $q(\boldsymbol{W}_r)$, we obtain the following posterior distribution [11]:

$$\tilde{q}(\boldsymbol{W}_r) \propto p(\boldsymbol{W}_r) \exp\left(\langle \boldsymbol{O}, \boldsymbol{S}|\boldsymbol{W}_r \rangle_{q(\boldsymbol{S})}\right). \tag{8}$$

After expectation with respect to $q(\boldsymbol{S})$, we can obtain the following expression:

$$\begin{aligned}
&\tilde{q}(\boldsymbol{W}_r) \propto p(\boldsymbol{W}_r|\boldsymbol{C}_r, \boldsymbol{V}_r) \\
&\exp\left\{\sum_{t,u \in r} \gamma_u(t) \left(\ln \mathcal{N}(\boldsymbol{W}_r\boldsymbol{\xi}(t))|\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u\right) + \ln|\boldsymbol{A}_r|\right\},
\end{aligned} \tag{9}$$

where $\boldsymbol{W}_r = [\,\boldsymbol{b}_r \;\; \boldsymbol{A}_r\,]$ is the $n \times (n+1)$ extended transformation matrix, which is composed of the $n \times 1$ bias term $\boldsymbol{b}_r$ and the $n \times n$ transformation matrix $\boldsymbol{A}_r$, $\gamma_u(t)$ is the posterior probability of being in the u-th Gaussian mixture component at frame t, $\boldsymbol{\xi}(t) = [\,1 \;\; \boldsymbol{o}(t)'\,]$ is the $(n+1) \times 1$ extended observation vector, and $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ are the mean vector and covariance matrix for Gaussian component u, respectively. Here we use Gaussian component u only in the specific node r ($u \in r$).

The above equation has the same form as the auxiliary Q-function of fSMAPLR [8][2]. In fSMAPLR, the transformed feature vector $\hat{\boldsymbol{o}}(t)$ is represented by

$$\hat{\boldsymbol{o}}(t) = \boldsymbol{A}_r\boldsymbol{o}(t) + \boldsymbol{b}_r = \boldsymbol{W}_r\boldsymbol{\xi}(t), \tag{10}$$

where $\boldsymbol{o}(t)$ represents an $n \times 1$ speech feature vector at frame t.

Since the VBLR approach is based on hierarchical prior setting, we need to define the hyperparameters for a specific node, namely

$$\begin{aligned}
\boldsymbol{C}_r &= \begin{cases} [\boldsymbol{0}_n' \;\; \boldsymbol{I}_n] & \text{if } r \text{ is root node} \\ \boldsymbol{W}_{r(p)} & \text{otherwise} \end{cases}, \\
\boldsymbol{V}_r &= \rho_r^{-1}\boldsymbol{I}_{n+1}.
\end{aligned} \tag{11}$$

where $r(p)$ denotes the parent node of the r-th node. Generally, an identity transformation matrix is used as the initial prior for the root node [6, 8, 11]. Here ρ_r^{-1} depends on each node different from ρ of fSMAPLR. Because in the fSMAPLR, only one ρ is used for the entire tree structure. We note that one of the advantages of the proposed method is that the control parameters ρ_r can be optimized *automatically* for each node independently without using the heuristic rule [9].

By substituting Eq. (5) into Eq. (9) and taking the logarithm of Eq. (9), Eq. (9) can be rewritten as

$$\begin{aligned}
&\ln \tilde{q}(\boldsymbol{W}_r) \\
&\propto -\frac{1}{2}\mathrm{tr}\left[\rho_r\boldsymbol{W}_r'\boldsymbol{W}_r + \boldsymbol{W}_r'\boldsymbol{W}_r\boldsymbol{G}_r^{(i)} - 2\rho_r\boldsymbol{W}_r'\boldsymbol{C}_r \right. \\
&\quad \left. -2\boldsymbol{W}_r'\boldsymbol{k}_r^{(i)}\right] + \beta_r \ln|\boldsymbol{A}_r| \\
&= -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{W}_r'\boldsymbol{W}_r(\rho_r\boldsymbol{I}_{n+1} + \boldsymbol{G}_r^{(i)}) - 2\boldsymbol{W}_r'(\rho_r\boldsymbol{C}_r + \boldsymbol{k}_r^{(i)})\right] \\
&\quad + \beta_r \ln|\boldsymbol{A}_r|,
\end{aligned} \tag{12}$$

where we disregarded the terms that do not depend on $\boldsymbol{W}_r$. And the 2nd and 1st order statistics of $\boldsymbol{G}_r^{(i)}$ and $\boldsymbol{k}_r^{(i)}$ are calculated by

[2] We can use the predictive distribution to estimate the transformation matrix. In this paper, we use point estimation by fSMAPLR approach because we verified that the obtained results are the same as fSMAPLR.

$$\boldsymbol{G}_r^{(i)} = \sum_{u \in r} \frac{1}{\sigma_u^{(i)2}} \sum_{t=1}^{T} \gamma_u(t) \boldsymbol{\xi}(t) \boldsymbol{\xi}(t)', \tag{13}$$

$$\boldsymbol{k}_r^{(i)} = \sum_{u \in r} \frac{1}{\sigma_u^{(i)2}} \mu_u^{(i)} \sum_{t=1}^{T} \gamma_u(t) \boldsymbol{\xi}(t)', \tag{14}$$

where $\mu_u^{(i)}$ is the i-th mean of $\boldsymbol{\mu}_u$ and $\sigma_u^{(i)}$ is the i-th diagonal component of $\boldsymbol{\Sigma}_u$.

The optimization of Eq. (9) can be performed by using row-by-row iterative estimation [4]. The i-th row of the transformation matrix $\boldsymbol{W}_r$ is calculated by

$$\boldsymbol{w}_{ri} = \left(\alpha \boldsymbol{p}_i + \tilde{\boldsymbol{k}}_r^{(i)}\right) \left(\tilde{\boldsymbol{G}}_r^{(i)}\right)^{-1}, \tag{15}$$

where α is the solution of a simple quadratic equation that maximizes Eq. (9) and $\boldsymbol{p}_i$ is the extended cofactor row vector, $[\,0 \;\; \mathrm{cof}(\boldsymbol{A}_{i1}) \;\; \ldots \;\; \mathrm{cof}(\boldsymbol{A}_{in})\,]$. And the statistics $\tilde{\boldsymbol{G}}_r^{(i)}$ and $\tilde{\boldsymbol{k}}_r^{(i)}$ smoothed by the prior distribution are represented by

$$\begin{aligned} \tilde{\boldsymbol{G}}_r^{(i)} &= \rho_r \boldsymbol{I}_{n+1} + \boldsymbol{G}_r^{(i)}, \\ \tilde{\boldsymbol{k}}_r^{(i)} &= \rho_r \boldsymbol{c}_{ri} + \boldsymbol{k}_r^{(i)}, \end{aligned} \tag{16}$$

where $\boldsymbol{c}_{ri}$ is the i-th row of the location matrix $\boldsymbol{C}_r$.

Finally, by defining the following matrix variables,

$$\tilde{\boldsymbol{V}}_r = \left(\rho_r \boldsymbol{I}_{n+1} + \boldsymbol{G}_r^{(i)}\right)^{-1}, \tag{17}$$

$$\tilde{\boldsymbol{c}}_{ri} = \left(\rho_r \boldsymbol{c}_{ri} + \boldsymbol{k}_r^{(i)}\right) \tilde{\boldsymbol{V}}_r, \tag{18}$$

we can obtain posterior distribution of $\boldsymbol{W}_r$ analytically as follows:

$$\begin{aligned} &\tilde{q}(\boldsymbol{W}_r) = \mathcal{N}(\boldsymbol{W}_r | \tilde{\boldsymbol{C}}_r, \boldsymbol{I}_n, \tilde{\boldsymbol{V}}_r) \\ &= (2\pi)^{-\frac{n(n+1)}{2}} |\tilde{\boldsymbol{V}}_r|^{-\frac{n}{2}} \\ &\exp\left(-\frac{1}{2}\mathrm{tr}\left[(\boldsymbol{W}_r - \tilde{\boldsymbol{C}}_r)'(\boldsymbol{W}_r - \tilde{\boldsymbol{C}}_r)\tilde{\boldsymbol{V}}_r^{-1}\right]\right). \end{aligned} \tag{19}$$

As already explained, the posterior distribution also becomes a matrix normal distribution.

2.3. Variational lower bound

Finally, after taking the expectation of Eq. (4) with respect to $q(\boldsymbol{S})$, we can obtain the following equation, which provides an analytical result for the lower bound [11]:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Psi}, m) \\ &= \sum_r \left\langle \ln \frac{p(\boldsymbol{O}, \boldsymbol{S} | \boldsymbol{W}_r)^{\gamma_u(t)} p(\boldsymbol{W}_r | \boldsymbol{C}_r, \boldsymbol{V}_r)}{q(\boldsymbol{W}_r)} \right\rangle_{q(\boldsymbol{S})} \\ &\propto \frac{n}{2} \ln |\boldsymbol{V}_r^{-1}| + \frac{n}{2} \ln |\tilde{\boldsymbol{V}}_r| - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}_r' \boldsymbol{C}_r \boldsymbol{V}_r - \tilde{\boldsymbol{C}}_r' \tilde{\boldsymbol{C}}_r \tilde{\boldsymbol{V}}_r^{-1}\right]. \end{aligned} \tag{20}$$

In this equation, the term $\ln |\boldsymbol{A}_r|$ is cancelled out because both $p(\boldsymbol{O}, \boldsymbol{S} | \boldsymbol{W}_r)^{\gamma_u(t)}$ and $\ln \tilde{q}(\boldsymbol{W}_r)$ have the same terms with opposite signs. Finally, fVBLR and VBLR have the similar form and the only difference between the feature space and model space is the transformation matrix and the 2nd and 1st order statistics. fVBLR also considers the variational lower bound as an objective function for control parameter and model structure optimization.

2.4. Control parameter and model structure optimization

Using Eq. (20) , we first optimize the control parameter ρ_r by using $\mathcal{L}(\boldsymbol{\Psi}, m)_r$, namely

$$\tilde{\rho}_r = \underset{\rho_r}{\mathrm{argmax}} \mathcal{L}(\boldsymbol{\Psi}, m)_r. \tag{21}$$

We use line search for optimizing the control parameter $\tilde{\rho}_r$. Then, we decide the model (tree) structure m without using the empirical occupancy threshold. Using the obtained $\tilde{\rho}_r$, we calculate $\mathcal{L}(\boldsymbol{\Psi}, m)_r$. If we focus on node r in the tree, and if node r is not a leaf node, we compute the following difference of the logarithmic evidences between the current node r and two child nodes $r(c1)$ and $r(c2)$

$$\Delta\mathcal{L}(\boldsymbol{\Psi}, m)_r \triangleq \mathcal{L}(\boldsymbol{\Psi}, m)_r - \mathcal{L}(\boldsymbol{\Psi}, m)_{r(c1)} - \mathcal{L}(\boldsymbol{\Psi}, m)_{r(c2)}. \tag{22}$$

If the sign of $\Delta\mathcal{L}(\boldsymbol{\Psi}, m)_r$ is positive, we continue splitting the node r to $r(c1)$ and $r(c2)$, and if the sign is negative, we stop splitting at node r. This optimization is efficiently accomplished by using a depth-first search.

3. MODEL SPACE VARIATIONAL BAYESIAN LINEAR REGRESSION

In this section, we explain model space variational Bayesian linear regression (VBLR). In the model space VBLR [11], the variational lower bound was analytically derived by using conjugate distributions (Eq. (5)) as prior distributions[3], and assuming the conditional independence on the posterior distributions. Here we only show the final lower bound in the model space.

$$\begin{aligned} \mathcal{L}^{\mathcal{M}}(\boldsymbol{\Psi}, m) \propto \frac{n}{2} \ln |(\boldsymbol{V}_r^{\mathcal{M}})^{-1}| + \frac{n}{2} \ln |\tilde{\boldsymbol{V}}_r^{\mathcal{M}}| \\ - \frac{1}{2}\mathrm{tr}\left[\left(\boldsymbol{C}_r^{\mathcal{M}}\right)' \boldsymbol{C}_r^{\mathcal{M}} \boldsymbol{V}_r^{\mathcal{M}} - \left(\tilde{\boldsymbol{C}}_r^{\mathcal{M}}\right)' \tilde{\boldsymbol{C}}_r^{\mathcal{M}} (\tilde{\boldsymbol{V}}_r^{\mathcal{M}})^{-1}\right]. \end{aligned} \tag{23}$$

For the model space variables, we use $(\cdot)^{\mathcal{M}}$ notation. In the model space, optimization of the control parameter and model structure is the same as explained in Section 2.4. A detailed explanation of the method can be found in [11].

4. EXPERIMENTS

4.1. Experimental conditions

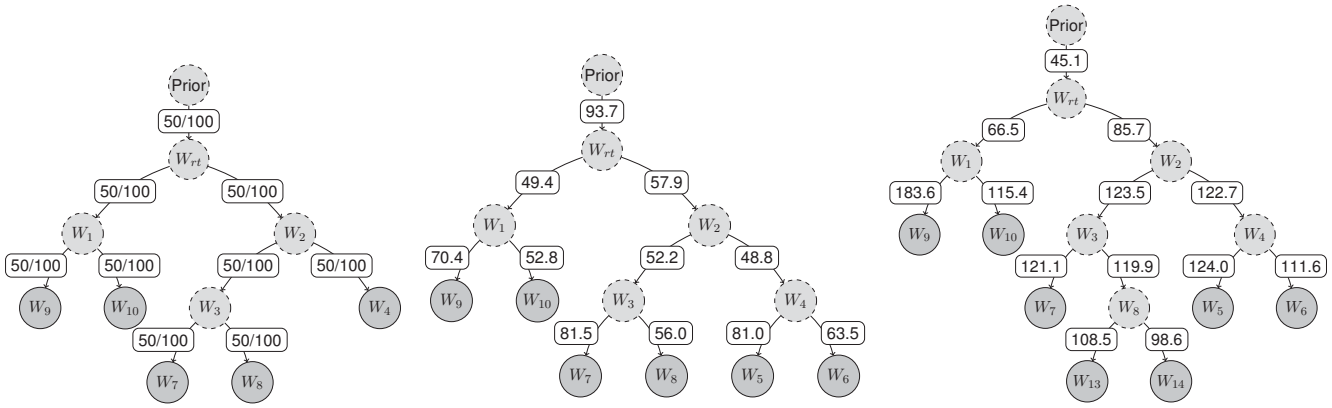
The training data consisted of 967 talks from the Corpus of Spontaneous Japanese (CSJ) [16] conference presentations of 558 speakers (234 hours of speech data). The test set consisted of 30 talks (6.4 hours, 70,369 words) from 30 speakers (20 males and 10 females). Table 1 shows the experimental setup.

The acoustic model training, decoding, and the following acoustic model adaptation procedures were performed with the NTT speech recognition platform SOLON [17]. For fSMAPLR and SMAPLR, the occupancy threshold was empirically set to 500. The control parameter for the feature space (ρ=50) and model space ($\rho^{\mathcal{M}}$=100) were also set empirically. In all our experiments, we only considered block-diagonal forms of the transformation matrix regardless of the adaptation methods.

4.2. Unsupervised adaptation experiment

We performed experiments using conventional fMAPLR, SMAPLR, VBLR and fSMAPLR+SMAPLR methods, and the proposed

[3]The prior distribution for a feature and model space has the same form. In Eq. (5), the transformation matrix $\boldsymbol{W}$ has to be replaced with $\boldsymbol{W}^{\mathcal{M}}$.



(a) (f)SMAPLR (ρ=50, $\rho^{\mathcal{M}}$=100, 5 leaf nodes) (b) fVBLR (6 leaf nodes) (c) VBLR (7 leaf nodes)

Fig. 1. Comparison of optimized model structure and control parameter sets (Speaker ID = "A01M0097", 5 adaptation utterances≈44.68 seconds). In the figure, each gray circle means a leaf node and each number between two nodes represents the estimated control parameter.

Table 2. Word error rate comparison of unsupervised adaptation (%).

Adaptation Method	Baseline	Number of adaptation utterances (average length in seconds)								
		1 (4.9)	2 (12.6)	3 (19.7)	4 (27.4)	5 (34.8)	10 (69.1)	20 (138.9)	50 (340.9)	all (678.3)
fSMAPLR	22.4	21.6	20.7	**20.2**	**20.1**	20.2	19.8	19.1	18.8	**18.4**
fVBLR	22.4	21.6	20.7	20.3	20.2	**20.1**	**19.6**	**19.0**	**18.7**	18.5
SMAPLR	22.4	21.8	20.9	20.6	20.2	20.2	19.8	19.4	**19.1**	**18.7**
VBLR	22.4	**21.7**	**20.8**	**20.5**	20.2	**20.1**	19.8	**19.3**	19.3	18.9
fSMAPLR+SMAPLR	22.4	21.8	**20.5**	20.0	**19.5**	**19.4**	**18.9**	**18.3**	**18.0**	**17.7**
fVBLR+VBLR	22.4	**21.4**	20.7	20.0	19.8	19.7	19.2	18.6	18.3	18.1

Table 1. Experimental setup

Sampling rate	16 kHz
Feature vector	MFCC + Energy + Δ + $\Delta\Delta$ (39 dims.)
Frame length	25 ms
Frame shift	10 ms
Window type	Hamming
CMN	Applied
No. of categories	43 phonemes
HMM topology	Context-dependent 2,000 states, 16 mixtures 3-state left-to-right HMM
Training method	ML Baum-Welch
Language model	3-gram (Kneser-Ney smoothing)
Vocabulary size	100,808
Perplexity	119.8
OOV rate	1.3

fVBLR and fVBLR+VBLR methods. The experimental results are shown in Table 2.

At first, we compare the results of fSMAPLR and fVBLR. Comparing these two results, we found that the fVBLR had the similar results with the fSMAPLR. We could find the similar trends between SMAPLR and VBLR, and between fSMAPLR+SMAPLR and fVBLR+VBLR. From these results, we can say that variational Bayesian based tuning-free approach can equivalently perform with the conventional fine-tuned approaches.

Furthermore, fVBLR+VBLR showed performance improvement compared with the performance obtained by each of fVBLR and VBLR. We had the best performance when only 1 utterance is used for the adaptation. This result also shows that the variational lower bound can achieve a better approximation of the marginalized log likelihood especially for the small amount of data [18, 19].

Figure 1 shows the estimated control parameter of each node and the tree structures for the speaker A01M0097 using 5 adaptation utterances. Figure 1(a) shows the tree structure and control parameter obtained with the conventional (f)SMAPLR. The tree structure is just determined by empirical occupancy threshold. And the control parameter for all nodes is fixed to an estimated value from the preliminary experiment using development set. Figure 1(b) and (c) show the obtained control parameters and tree structures by fVBLR and VBLR. fVBLR and VBLR shows the control parameters vary from node to node. Because the variational Bayesian approach is based on the lower bound, the optimized tree structures are different from each other (5 leaf nodes for (f)SMAPLR, 6 leaf nodes for fVBLR, and 7 leaf nodes for VBLR).

Through these experiments, we found the effectiveness of the proposed fVBLR and fVBLR+VBLR adaptation without tuning the hyperparameters empirically.

5. CONCLUSIONS

In this paper, we have proposed the *feature space* VBLR (fVBLR) and combined it with the model space VBLR (fVBLR+VBLR). These are *fully Bayesian* and *parameter tuning-free* approaches. In the experiments, we confirmed that the proposed method can equivalently perform with the conventional fine-tuned approaches. Furthermore, fVBLR+VBLR showed better performance than that obtained by each of fVBLR and VBLR. Future work will include the selection of the type of transformation matrix based on the lower bound.

6. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Speaker adaptation of HMMs using linear regression," *Cambridge University, Cambridge, UK, Tech. Rep. CUED/F-INFENG/TR*, vol. 181, 1994.

[2] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, pp. 171–185, 1995.

[3] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[4] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[5] O. Siohan, C. Chesta, and C.-H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 417–428, 2001.

[6] O. Siohan, T.A. Myrvoll, and C.-H. Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, 2002.

[7] X. Lei, J. Hamaker, and X. He, "Robust feature space adaptation for telephony speech recognition," in *Proc. of INTERSPPECH*, 2006, pp. 773–776.

[8] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. of INTERSPEECH*, 2006, pp. 2286–2289.

[9] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 276–287, 2001.

[10] P.C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001.

[11] S. Watanabe, A. Nakamura, and B.-H. Juang, "Bayesian linear regression for Hidden Markov Model based on optimizing variational bounds," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2011, pp. 1–6.

[12] 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, no. 4, pp. 365–381, 2004.

[13] S.-J. Hahm, S. Watanabe, M. Fujimoto, A. Ogawa, T. Hori, and A. Nakamura, "Normalization and adaptation by consistently employing MAP estimation," in *Proc. of International Workshop on Statistical Machine Learning for Speech Processing (IWSML)*, 2012.

[14] S.-J. Hahm, A. Ogawa, M. Fujimoto, T. Hori, and A. Nakamura, "Speaker adaptation using variational Bayesian linear regression in normalized feature space," in *Proc. of INTERSPEECH*, 2012, p. Tue.O4a.05.

[15] M.J. Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University of London, 2003.

[16] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. of LREC2000 (Second International Conference on Language Resources and Evaluation)*, 2000, vol. 2, pp. 947–952.

[17] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[18] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. of Uncertainty in artificial intelligence (UAI)*, 1999, vol. 2, pp. 21–30.

[19] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, vol. 15, no. 10, pp. 1223–1242, 2002.