Speaker Adaptation Using Variational Bayesian Linear Regression in Normalized Feature Space

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

NTT Communication Science Laboratories, NTT Corporation (Kyoto, Japan)

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

Abstract

In this paper, we propose a feature and model space adaptation approach capable of automatically optimizing the model topology and setting the hyper-parameters. In our proposed framework, we first perform feature space SMAPLR (fSMAPLR) to normalize the feature space. Then, variational Bayesian linear regression (VBBLR) is employed in the obtained normalized feature space. The purpose of this work is to develop a tuning-free SMAPLR technique in a normalized feature space. An important issue regarding the SMAPLR approach is how to decide the occupancy threshold for the tree structure and the control parameter for scaling prior information. Recently, the VBLR approach has been proposed as a tuning-free SMAPLR approach. By using a variational lower bound as an objective function, the model structure and contribution of the prior information can be decided instead of using empirical values. Large vocabulary continuous speech recognition experiments using the Corpus of Spontaneous Japanese (CSJ) confirm the effectiveness of the proposed method compared with the conventional SMAPLR approach.

Index Terms: speaker adaptation, SMAPLR, VBLR, normalized feature space

1. Introduction

The linear regression approach has been widely used for speaker adaptation. Maximum likelihood linear regression (MLLR) [1] was first proposed and then extended to maximum a posteriori linear regression (MAPLR) incorporating prior information [2]. MAPLR was further extended to structural MAPLR (SMAPLR) [3] using a hierarchical prior setting originally proposed by structural MAP (SMAP) [4]. In these approaches, setting the occupancy threshold for the tree structure according to the amount of adaptation data is a common and important problem in terms of obtaining good results. In addition, for SMAPLR, a control parameter for hierarchical prior setting can also be an important issue because the recognition performance is greatly affected by this parameter especially for a small amount of adaptation data. If the control parameter is not optimal, the SMAPLR result may be even worse than that of MAPLR.

Variational Bayesian linear regression (VBBLR) [5] was proposed for solving the problems of deciding the occupancy threshold and control parameter. VBBLR is a fully Bayesian treatment of linear regression for hidden Markov models (HMMs) that employs variational Bayesian techniques. VBBLR analytically derives the variational lower bound of the marginalized log-likelihood of the linear regression. By using the variational lower bound as an objective function, we can optimize the model topology and hyper-parameters of the linear regression without controlling them as tuning parameters.

In this work, we investigate tuning-free SMAPLR approaches in a normalized feature space. In our previous work [6], we reported the effectiveness of the feature and model space speaker adaptation which consistently uses MAP-based estimation. We first employ the feature space SMAPLR (fSMAPLR) [7] for the feature space normalization as in our previous work. In the model space, VBBLR is then performed in this normalized feature space. In conventional SMAPLR approaches, the control parameter for scaling the prior is determined empirically. Furthermore, once the control parameter set is given, no node uses the same control parameter for scaling prior information (the transformation matrix of the parent node of a given node). Using VBLR, we can determine the appropriate control parameter for scaling the prior of each node separately. Furthermore, based on the optimal control parameter, we can control the model structure of the binary tree. Therefore, we can control the hyper-parameter regardless of the amount of adaptation data. We evaluate the effectiveness of the proposed method by employing a speech recognition experiment.

2. Feature space normalization

In this section, we review the feature space normalization method. We first describe a prior distribution and then describe fSMAPLR using a prior distribution and hierarchical prior setting.

2.1. Prior distribution

In the MAPLR approach, the matrix version of a multivariate normal distribution is used for the prior distribution [2, 8]. We also adopt the matrix version of a multivariate normal distribution as a prior distribution. The prior distribution is defined as follows:

\[
p(W) \propto p(W | C, \Phi, V) = \frac{1}{n! |\Phi|} |\Phi|^{n/2} |\Phi + WW^T|^{-n/2} \exp \left\{-\frac{1}{2} \text{tr}(V^{-1}(W - C)\Phi^{-1}(W - C))\right\},
\]

where \(W\) is the transformation matrix, and \(C, \Phi, V\) are the hyper-parameters for that distribution family. \(C\) is the \(n\times(n+1)\) location matrix and \(V\) is the \((n+1) \times (n+1)\) scaling matrix. These hyper-parameters are used as prior information for estimating the transformation matrix.

2.2. fSMAPLR

fSMAPLR is an extended version of fMAPLR using a hierarchical prior setting [7]. The auxiliary Q-function with the prior distribution for estimating feature space transformation parameters is

\[
Q_{MAP}(W_r^F; \hat{W}_r^F) = -\frac{1}{2} \sum_{t, u} \gamma_u(t) \left\{ \log |\Sigma_u(t)| - \log |A_r^F|^2 + (W_r^F)\xi(t) - \mu_u(t)S_u(t) - (W_r^F)\xi(t) - \mu_u(t) \right\} \\
+ \log(p(W_r^F | \hat{C}_r, \hat{V}_r)).
\]

Prior Term

(2)
where $W^T_x = [b^T_x A^T_x]$ is the $n \times (n + 1)$ extended feature space transformation matrix, which is composed of the $n \times 1$ bias term $b^T_x$ and the $n \times n$ matrix $A^T_x$, $\gamma(t) = 1$ is the posterior probability of being in the $u$-th Gaussian mixture at frame $t$, $\xi(t) = [1 \; o(t)']'$ is the $(n + 1) \times 1$ extended observation vector, $\mu^{(u)}$ and $\Sigma^{(u)}$ are the mean vector and covariance matrix for Gaussian component $u$, respectively, and $r$ denotes a node (class) of the binary tree. The transformed feature vector $\hat{o}(t)$ is represented by

$$\hat{o}(t) = A^T_x o(t) + b^T_x = W^T_x \xi(t),$$

where $o(t)$ represents an $n \times 1$ speech feature vector at frame $t$. In this paper, we use the notations $W^T_x$, $A^T_x$, and $b^T_x$ to differentiate them from the matrices of a model space.

In fSMAPLR, a hierarchical prior setting is used to estimate the transformation matrix. The prior term has to be set for each node (class) $r$. In fSMAPLR, the hyper-parameters for a specific node are set by

$$C_r = \begin{cases} 0[r] & \text{if } r \text{ is root node } \\ W_{r(p)} & \text{otherwise} \end{cases},$$

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 [7].

The optimization of the above Q-function can be performed by using row-by-row iterative estimation [9]. The $i$-th row of the transform matrix $W^T_x$ is calculated by

$$w_{ri} = \left( \alpha p_n + k_{ri} \right)^{-1} G_{ri}^{-1},$$

where $\alpha$ is the solution of a simple quadratic equation that maximizes the Q-function and $p_n$ is the extended cofactor row vector, $[0 \; \text{cof}(A^T_1) \ldots \; \text{cof}(A^T_n)]$. And the statistics $G_{ri}$ and $k_{ri}$ smoothed by the prior distribution are represented by

$$G_{ri} = G_{ri}^{(i)} + \hat{V}_{ri}^{-1},$$

$$k_{ri} = k_{ri}^{(i)} + \hat{c}_r V_{ri}^{-1},$$

where $\hat{c}_r$ is the $i$-th row of the location matrix $\hat{C}_r$. And the sufficient statistics of $G_{ri}$ and $k_{ri}$ calculated from the observed adaptation data are calculated by

$$G_{ri} = \sum_{u \in \mathcal{C}} \frac{1}{\sigma_{ri}^{(u)}} \sum_{t=1}^T \gamma_u(t) \xi(t) \xi(t)',$n

$$k_{ri} = \sum_{u \in \mathcal{C}} \frac{1}{\sigma_{ri}^{(u)}} \mu_{ri}^{(u)} \sum_{t=1}^T \gamma_u(t) \xi(t),$$

where $\mu_{ri}^{(u)}$ is the $i$-th mean of $\mu^{(u)}$ and $\sigma_{ri}^{(u)}$ is the $i$-th diagonal component of $\Sigma^{(u)}$.

3. Variational Bayesian linear regression in normalized feature space

In this section, we describe variational Bayesian linear regression in a previously normalized feature space. In the VBLR approach [5], the variational lower bound is analytically derived by using conjugate distributions (Eq. 1) as prior distributions, and by assuming the conditional independence on the posterior distributions. In variational Bayesian approaches, the lower bound of the marginalized likelihood with a set of hyper-parameters $\Psi$ and a model structure $m$ is represented by

$$\log p(\hat{O}|m, \Psi) \geq \log \frac{p(\hat{O}|S|W^M)p(W^M; \hat{O}, m, \Psi)}{q(W^M; \hat{O}, m, \Psi)q(S|O, m, \Psi)} + \hat{\delta}_F(\Psi, m),$$

where $\hat{O}$ is the transformed feature vector set obtained with previously estimated feature-space transformation matrices, $W^M$ is the model-space transformation matrix, and $S$ represents the sequences of HMM states and mixture components of Gaussian mixture models. $p(W^M|O, m, \Psi)$ is a prior distribution of $W^M$, and $q(W^M|O, m, \Psi)$ and $q(S|O, m, \Psi)$ are arbitrary distributions. The variational Bayes considers the variational lower bound $F(\Psi, m)$ as an objective function.

The variational lower bound defined in Eq. (9) can be decomposed as follows:

$$F(\Psi, m) = \log \frac{\log p(\hat{O}|S|W^M)p(W^M; \hat{O}, m, \Psi)}{q(W^M; \hat{O}, m, \Psi)q(S|O, m, \Psi)} + \hat{\delta}_L(\Psi, m).$$

From Eq. (10), we can take only the first logarithmic evidence term for $m$ and $\Psi$ because the second term does not depend on the transformation matrix $W^M$. Considering the conditional independence assumption over cluster $r$, $L(\Psi, m)$ can be represented by

$$L(\Psi, m) = \sum_r \log \frac{p(\hat{O}|S|W^M)p(W^M; \hat{O}, m, \Psi)}{q(W^M; \hat{O}, m, \Psi)q(S|O, m, \Psi)}.$$

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

$$C_r = \begin{cases} 0[r] & \text{if } r \text{ is root node } \\ W_{r(p)} & \text{otherwise} \end{cases},$$

$$V_r = \rho^{-1} I_{n+1},$$

Finally, after taking the expectation of the Eq. (11), we can obtain the following equation, which provides an analytical result for the lower bound [5]:

$$L(\Psi, m) = \sum_r \left[ -\frac{n}{2} \log(2\pi) \sum_{u \in \mathcal{C}} \gamma_u - \frac{1}{2} \sum_{u \in \mathcal{C}} \gamma_u \log |\Sigma^{(u)}| \\
+ \frac{n(n+1)}{2} \log \rho_r + \frac{n}{2} \log |\hat{V}_r| - \frac{1}{2} \left( \rho_r \hat{C}_r \hat{C}_r - \hat{C}_r \hat{C}_r \hat{V}_r^{-1} + \sum_{u \in \mathcal{C}} \gamma_u(t) \Sigma^{(u)} \hat{O}(t) \hat{o}(t)' \right) \right].$$

where

$$\hat{V}_r = (\rho_r I_{n+1} + G_{r(i)})^{-1}. \tag{14}$$
\[ \bar{C}_r = (\rho_r C_r + \hat{k}^{(i)}) \tilde{V}_r. \] (15)

And the sufficient statistics from the observed adaptation data transformed by feature-space normalization are calculated by

\[ G^{(u)}_r = \frac{1}{\sigma_{L}^r} \gamma_u(t) \xi^{(u)} \sum_{i=1}^{T} \gamma_u(t), \] (16)

\[ \hat{k}^{(i)}_r = \sum_{u \in r} \gamma_u(t) \frac{1}{\sigma_{L}^u} \phi_u(t) \xi^{(u)'}, \] (17)

where \( \xi^{(u)} \) is the extended mean vector.

Using Eq. (13), we first optimize the hyper-parameter \( \rho_r \) by using \( L(\Psi, m) \), namely

\[ \hat{\rho}_r = \arg \max \rho L(\Psi, m). \] (18)

Then, we decide the model structure \( m \) without using the empirical occupancy threshold \( \tau \). Using the obtained \( \hat{\rho}_r \), we calculate \( L(\Psi, m) \). If we focus on node \( r \) in the tree, and if node \( r \) is not a leaf node, we compute the following difference between the logarithmic evidence of the parent and child nodes

\[ \Delta L(\Psi, m) \triangleq L(\Psi, m) - L(\Psi, m)_{v(1)} - L(\Psi, m)_{v(2)}. \] (19)

If the sign of \( \Delta L(\Psi, m) \) is positive, we continue splitting the node \( r \) to \( v(1) \) and \( v(2) \), and if the sign is negative, we stop splitting at node \( r \). This optimization is efficiently accomplished by using a depth-first search.

Table 1 shows the advantage of the VBLR technique compared with conventional SMAPLR.

| Table 1: Parameter setting method of SMAPLR and VBLR |
|-----------------|-----------------|
| SMAPLR          | empirical       |
| VBLR            | based on lower bound |
| VBLR            | based on optimal \( \rho \) |

### 4. Experiments

The speaker adaptation performance of the proposed method was tested using a Corpus of Spontaneous Japanese (CSJ) task [10] in an unsupervised fashion.

#### 4.1. Experimental conditions

The training data consisted of 967 talks taken from the CSJ conference presentations of 799 speakers (234 hours of speech data), and the training data for the language model construction consisted of 2,672 talks from the complete CSJ speech data (6.8M word transcriptions). The test set consisted of 30 talks (6.4 hours, 70,369 words) from 30 speakers (20 males and 10 females). Table 2 shows pre-processing, acoustic and language model information. The acoustic model training, decoding, and the following acoustic model adaptation procedures were performed with the NTT speech recognition platform SOLON [11]. For ISMAPLR and SMAPLR, the occupancy threshold \( \tau \) was set empirically at 500. In all our experiments, we only considered the block-diagonal forms of the transformation matrix regardless of the adaptation methods.

#### 4.2. Preliminary experiments

In this section, we show the results of preliminary experiments designed to investigate the effect of the control parameter \( \rho \) for the prior distribution related to ISMAPLR, SMAPLR, and the proposed method.

<table>
<thead>
<tr>
<th>Table 2: Experimental setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
</tr>
<tr>
<td>Feature vector</td>
</tr>
<tr>
<td>Frame length</td>
</tr>
<tr>
<td>Frame shift</td>
</tr>
<tr>
<td>Window type</td>
</tr>
<tr>
<td>CMN</td>
</tr>
<tr>
<td>No. of categories</td>
</tr>
<tr>
<td>HMM topology</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Training method</td>
</tr>
<tr>
<td>Language model</td>
</tr>
<tr>
<td>Vocabulary size</td>
</tr>
<tr>
<td>Perplexity</td>
</tr>
<tr>
<td>OOV rate</td>
</tr>
</tbody>
</table>

We performed SMAPLR adaptation using various values of control parameter \( \rho \) for scaling matrix \( V \). The control parameter \( \rho \) plays an important role for estimating the transformation matrix of the current node in the SMAPLR approach. If the control parameter \( \rho \) is large\(^1\), the transformation matrix of the parent node has a great influence on the transformation matrix estimation of the current node. Otherwise, its impact is small. The value of the control parameter can directly affect the recognition performance especially when there are few adaptation data. Therefore, we investigated the recognition performance with various values of control parameter \( \rho \). The results for unsupervised adaptation are shown in Table 3.

When large amounts of adaptation data were used for the experiment, the results did not differ greatly. However, for a small amount of adaptation data, the results were very different and depended on the \( \rho \) value. The obtained results were similar to those obtained for the original SMAPLR [3]. We obtained the best results for a \( \rho \) value of 100. When \( \rho \) was larger than 100, the recognition performance was completely degraded for all numbers of adaptation utterances. Using the value of 100, ISMAPLR was performed first. In the next section, we describe our experimental results.

#### 4.3. Unsupervised adaptation experiment

To confirm the effect of the VBLR adaptation in the normalized feature space, we performed experiments using ISMAPLR, SMAPLR, VBLR, ISMAPLR+SMAPLR and the proposed ISMAPLR+VBLR methods. The experimental results are shown in Table 4. The speaker independent (initial) model has a word error rate (WER) of 22.4% for the test data.

First, we compare the results of SMAPLR and VBLR. We found that the VBLR approach produced similar results to the SMAPLR, which determines the control parameter \( \rho \) and the occupancy threshold experimentally. We also found that ISMAPLR+VBLR provided similar results to ISMAPLR+SMAPLR. We note that the occupancy threshold \( \tau \) and control parameter \( \rho \) for SMAPLR were fine-tuned from the preliminary experiments. We obtained the best performance when we used only 1 utterance for the adaptation. This result also shows that the variational lower bound can achieve a better approximation of the marginalized log likelihood especially for a small amount of data [12, 13]. Through these experiments, we found the effectiveness of VBLR adaptation in a normalized feature space without tuning the hyper-parameters.

\(^1\)The control parameter \( \rho \) defined in this paper is the reciprocal of the control parameter of SMAPLR [3].
Table 3: Word error rate comparison with various values of control parameter $\rho$ (%).

<table>
<thead>
<tr>
<th>Adaptation Method</th>
<th>Number of adaptation utterances (average length in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>SMAPLR</td>
<td>27.7</td>
</tr>
<tr>
<td>10</td>
<td>23.7</td>
</tr>
<tr>
<td>100</td>
<td>21.8</td>
</tr>
<tr>
<td>1000</td>
<td>21.9</td>
</tr>
<tr>
<td>ISMAPLR</td>
<td>29.9</td>
</tr>
<tr>
<td>10</td>
<td>25.6</td>
</tr>
<tr>
<td>100</td>
<td>21.6</td>
</tr>
<tr>
<td>1000</td>
<td>21.8</td>
</tr>
</tbody>
</table>

Table 4: Word error rate comparison of unsupervised adaptation (%).

<table>
<thead>
<tr>
<th>Adaptation Method</th>
<th>Number of adaptation utterances (average length in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>fSMAPLR</td>
<td>21.6</td>
</tr>
<tr>
<td>SMAPLR</td>
<td>21.8</td>
</tr>
<tr>
<td>VBLR</td>
<td>21.7</td>
</tr>
<tr>
<td>fSMAPLR+SMAPLR</td>
<td>21.8</td>
</tr>
<tr>
<td>fSMAPLR+VBLR</td>
<td>21.5</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, we proposed the VBLR approach in a normalized feature space. We first performed ISMAPLR for feature space normalization. We then employed the VBLR in the normalized feature space. The experimental results revealed the effectiveness of the proposed method. VBLR provided similar results to those obtained using SMAPLR for which fine-tuned hyperparameters were used. Future work will include feature space VBLR (fVBLR) and fVBLR followed by VBLR methods in feature and model space.

6. Acknowledgements

The authors would like to thank Shinnji Watanabe at Mitsubishi Electric Research Laboratories (MERL), Massachusetts, USA, for his valuable comments on this work.

7. References