BAYESIAN LARGE MARGIN HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

Jung-Chun Chen and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{jcchen,chien}@chien.csie.ncku.edu.tw

ABSTRACT

This paper presents a Bayesian learning approach to large margin classifier for hidden Markov model (HMM) based speech recognition. We build the Bayesian large margin HMMs (BLM-HMMs) and improve the *model generalization* for handling unknown test environments. Using BLM-HMMs, the variational Bayesian HMM parameters are estimated by maximizing lower bound of a marginal likelihood over the uncertainties of HMM parameters. The Bayesian large margin estimation is performed with frame selection mechanism, and is illustrated to meet the objective of support vector machines, i.e. maximal class margin and minimal training errors. The new objective function is not only interpreted as a discriminative criterion, but also feasible to deal with model selection and adaptive training. Experiments on phone recognition show that BLM-HMMs perform better than other generative and discriminative models.

Index Terms— Bayesian learning, model generalization, large margin classifier, hidden Markov models

1. INTRODUCTION

Support vector machines (SVMs) [11] are known as a powerful mechanism for general pattern recognition, and have been successfully applied for speech recognition [6]. SVMs perform the large/maximum margin classification using the kernel features and support tokens, and are also called the *sparse kernel machines* [2]. The principle of SVMs is to maximize the minimum misclassification margin. The correct classification can be made when a new sample falls near margin region. The margin is surrounded by support vectors. In [6][10], the large margin HMM (LM-HMM) parameters was estimated for large margin classification of speech signal. Using LM-HMMs, the support tokens, which are correctly classified, are selected to adjust the decision boundaries, and so those correctly recognized utterances leave away boundaries as far as possible. LM-HMMs were presented to improve the generalization of discriminative training using the minimum classification error (MCE) method [7]. In LM-HMMs, some empirical parameters in sparse learning procedure should be tuned. The learning method using the generalized probabilistic descent algorithm is prone to converge at local optimum.

To activate the capabilities of *model regularization* and *adaptive learning* [14], we are motivated to introduce the Bayesian theory into the large margin classification [8], and present the Bayesian large margin (BLM) classifier for HMM-based speech recognition. In this BLM-HMM framework, the margin is seen as a logarithm of ratio of posterior distributions of target model to competing model. A variational Bayesian inference [1][2][12] is developed to establish the large margin classifier by decomposing the variational models due to different HMM parameters and latent variables. The graphical models of BLM-HMM and its variational models are demonstrated. A new objective function is constructed for model training from the sequence data and is illustrated to meet the properties of SVMs. In the experiments, we conduct the evaluation of speech recognition by using TIMIT database and obtain the improvement of proposed BLM-HMMs compared to the maximum likelihood HMMs (ML-HMMs), the MCE-HMMs and the LM-HMMs.

2. LARGE MARGIN HMMS

In a standard speech recognizer, we choose the most likely word sequence corresponding to an input utterance $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ by the maximum *a posteriori* decision

$$\hat{W} = \arg\max_{W} p(W \mid X) = \arg\max_{W} p(X \mid W, \lambda) p(W) \qquad (1)$$

where λ is the acoustic model and $p(W)$ is the language model that we don't consider in this work. The discriminant function is specified by $\log p(X \mid W, \lambda)$. In a LM classifier, the separation margin of an utterance X_i with true word identity W_i is defined as a log likelihood ratio of discriminant functions of true model to the most competing model

$$d_{\mathrm{LM}}(X_i) = \log p(X_i \mid \lambda_{W_i}) - \max_{W_j \in \Omega_W, j \neq i} \log p(X_i \mid \lambda_{W_j}) \qquad (2)$$

where Ω_W denotes the space of word sequence. If $d(X_i) > 0$, the classification is correct; otherwise a wrong decision is made. LM-HMM parameters $\lambda_{\mathrm{LM}} = \{\lambda_i\}$ [6][10] are trained by selecting the support tokens from training utterances D with a preset positive number ε

$$\Psi_{\mathrm{LM}} = \{X_i \mid X_i \in D \text{ and } 0 \le d_{\mathrm{LM}}(X_i) \le \varepsilon\} \qquad (3)$$

and maximizing the minimum margin due to support tokens by

$$\lambda_{\mathrm{LM}} = \arg\max_{\lambda} \min_{X_i \in \Psi_{\mathrm{LM}}} d_{\mathrm{LM}}(X_i) =$$

$$= \arg\min_{\lambda} \max_{X_i \in \Psi_{\mathrm{LM}}, W_j \in \Omega_W, j \neq i} [\log p(X_i \mid \lambda_{W_j}) - \log p(X_i \mid \lambda_{W_i})] \qquad (4)$$

$$\approx \arg\min_{\lambda} \log \left[\sum_{X_i \in \Psi_{\mathrm{LM}}, W_j \in \Omega_W, j \neq i} \exp \begin{bmatrix} \eta \log p(X_i \mid \lambda_{W_j}) \\ -\eta \log p(X_i \mid \lambda_{W_i}) \end{bmatrix} \right]^{1/\eta}.$$

The selected tokens are correctly classified, and relatively close to the separation boundary. This minimax optimization performs the so-called *sparse learning* since only support

tokens are put into model training. Accordingly, LM-HMMs fulfill the spirit of SVMs [11], which are popular in many pattern recognition applications. In (4), the optimization is approximated by considering the margin with smoothing parameter η significantly larger than 1.

Considering the continuous-density HMMs with state observation probability $p(\mathbf{x}\,|\,\lambda_i)$, $1 \le i \le I$, composed of a mixture of K Gaussian densities

$$\sum_{k=1}^{K} \omega_{ik} \cdot (2\pi)^{-d/2} \,|\, r_{ik} \,|^{1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_{ik})^T r_{ik} (\mathbf{x} - \mu_{ik}) \right] \quad (5)$$

the discriminant function of an input utterance $X = \{\mathbf{x}_t\}$ is produced by

$$\log p(X \,|\, \lambda_W) = \log \sum_S \sum_L p(X, S, L \,|\, \lambda_W) \approx \log \pi_{\hat{s}_1} + \sum_t \left[\log a_{\hat{s}_{t-1}\hat{s}_t} + \log \omega_{\hat{s}_t \hat{l}_t} - \frac{1}{2} (\mathbf{x}_t - \mu_{\hat{s}_t \hat{l}_t})^T r_{\hat{s}_t \hat{l}_t} (\mathbf{x}_t - \mu_{\hat{s}_t \hat{l}_t}) \right] \quad (6)$$

where $\{\pi_i, a_{im}, \omega_{ik}, \mu_{ik}, r_{ik}\}$ are the initial state probabilities, transition probabilities, mixture weights, mean vectors and precision matrices, respectively. Here, the Viterbi approximation is applied by adopting the decoded state sequence $\hat{S} = \{\hat{s}_t\}$ and mixture component sequence $\hat{L} = \{\hat{l}_t\}$. The objective function in (4) can be expanded for optimization. To avoid the unlimitedly increasing margin, Jiang et al. [6] imposed the limitations of linear and quadratic terms in objective function and derived the LM-HMM parameters based on the penalized gradient descent algorithm. The imposed constraints acted as the penalization terms in the extended objective function. Starting from the seed model trained by MCE algorithm [7], an iterative optimization procedure is performed to find LM-HMM parameters [6].

3. BAYESIAN LARGE MARGIN HMMS

In general, LM-HMMs are trained as the *point estimates*, where the model uncertainties are not considered. However, in real-world applications, we suffer from the problems of *ill-posed modeling* and *environmental mismatch* between training and test data [3]. The static LM-HMM parameters are not fitted to the unknown variations in test environments. The model space and the model complexity are not well controlled. To deal with the issue of model adaptation and selection [4][5][12], we are motivated to develop a *Bayesian framework* for LM-HMMs.

3.1. Bayesian large margin estimation

From Bayesian viewpoint, the model uncertainty is considered in expressing the separation margin. The model parameters of Bayesian LM-HMMs (denoted by BLM-HMMs) λ_{BLM} are estimated by a minimax procedure according to a Bayesian criterion based on the posterior distributions

$$\sum_{X_i \in \Psi_{\text{BLM}}, W_j \in \Omega_W, j \neq i} \exp[\log p(\lambda_{W_j} \,|\, X_i) - \log p(\lambda_{W_i} \,|\, X_i)] \quad (7)$$

where the set of support tokens Ψ_{BLM} is collected similar to (3) except that the posterior distribution is adopted in calculation of separation margin $d_{\text{BLM}}(X_i)$. The smoothing parameter η in LM estimation (4) is neglected in BLM estimation (7). The model uncertainty, which is characterized by a prior density, is merged in the posterior distribution. However, the true posterior distribution with latent variables is unknown in speech recognition. We can apply the variational Bayesian (VB) method [1][2][12][13] and use a variational distribution $q(\lambda_W \,|\, X)$ to approximate the true distribution $p(\lambda_W \,|\, X)$. The variational distribution is estimated by an approximate inference method through the maximization of a lower bound of logarithm of marginal likelihood. The lower bound is obtained via the Jensen's inequality as given by [2][12]

$$\begin{aligned} \log p(X) &= \log \int \sum_{S,L} p(X, S, L \,|\, \lambda_W) p(\lambda_W) d\lambda_W \\ &\ge \int \sum_{S,L} q(S, L, \lambda_W \,|\, X) \log \frac{p(X, S, L \,|\, \lambda_W) p(\lambda_W)}{q(S, L, \lambda_W \,|\, X)} d\lambda_W \\ &= \int q(\lambda_W \,|\, X) \left[\sum_{S,L} q(S, L \,|\, X) \log \frac{p(X, S, L \,|\, \lambda_W) p(\lambda_W)}{q(\lambda_W \,|\, X)} \right] d\lambda_W \\ &\quad - \sum_{S,L} q(S, L, X) \log q(S, L \,|\, X). \end{aligned} \quad (8)$$

In (8), the factorization $q(S, L, \lambda_W \,|\, X) = q(S, L \,|\, X) q(\lambda_W \,|\, X)$ is applied with (S, L) being latent variables. By taking differential of right hand side of (8) with respect to $q(\lambda_W \,|\, X)$ and setting it to zero, we obtain the optimal variational model $\tilde{q}(\lambda_W \,|\, X)$ which approximates the true model $p(\lambda_W \,|\, X)$ with the smallest Kullback-Leibler divergence [2]. As a result, BLM-HMMs are implemented by inferring a new discriminant function $\tilde{q}(\lambda_W \,|\, X)$ attaining a lower bound and maximizing the function to fulfill the minimax estimation in (7). At the same time, the variational posterior distribution $\tilde{q}(S, L \,|\, X)$ can be calculated. Due to the incomplete data problem in HMMs, the expectation-maximization (EM) steps should be performed to update the current estimate λ to new estimate λ' in an VB-EM procedure [1][12][14].

3.2. Implementation in BLM-HMMs

In the variational inference, we specify the prior densities of parameters π_i, a_{im} and ω_{ik} to be Dirichlet densities with hyperparameters ϖ_i, ϕ_{im} and φ_{ik}, respectively, and the prior density of Gaussian mean μ_{ik} and precision r_{ik} to be a normal-Wishart density

$$\begin{aligned} p(\mu_{ik}, r_{ik} \,|\, m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}) &= |r_{ik}|^{(\alpha_{ik} - d)/2} \\ &\times \exp\left[-\frac{\tau_{ik}}{2} (\mu_{ik} - m_{ik})^T r_{ik} (\mu_{ik} - m_{ik}) \right] \exp\left[-\frac{1}{2} \text{tr}(u_{ik} r_{ik}) \right] \end{aligned} \quad (9)$$

with hyperparameters $\alpha_{ik} > d - 1$, $\tau_{ik} > 0$, μ_{ik} being $d \times 1$, and u_{ik} being $d \times d$ and positive definite [4]. By combining the likelihood function in (6) and the conjugate prior in (9), the posterior distribution $p(\lambda \,|\, X)$ can be written as a product of posterior distributions for individual HMM parameters. Here, we infer the variational posterior distribution $q(\lambda \,|\, X)$, or equivalently infer the variational distributions for four sets of HMM parameters $\{\pi_i, a_{im}, \omega_{ik}, (\mu_{ik}, r_{ik})\}$. The variational inference proceeds by maximizing the lower bound in (8) with respect to $q(\lambda \,|\, X)$. The optimal VB distribution is inferred by [12]

$$\tilde{q}(\lambda \,|\, X) \propto p(\lambda \,|\, \{\varpi_i, \phi_{im}, \varphi_{ik}, m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}\})$$

$$\times \exp\left[\sum_{S,L} \widetilde{q}(S,L \mid X) \log p(X,S,L \mid \lambda)\right]$$
$$= \prod_{i,m,k} \widetilde{q}(\{\pi_i\} \mid X)\, \widetilde{q}(\{a_{im}\} \mid X)\, \widetilde{q}(\{\omega_{ik}\} \mid X)\, \widetilde{q}(\{\mu_{ik}, r_{ik}\} \mid X) \quad (10)$$
$$= \prod_{i,m,k} p(\{\pi_i\} \mid \{\widetilde{\varpi}_i\}) p(\{a_{im}\} \mid \{\widetilde{\phi}_{im}\}) p(\{\omega_{ik}\} \mid \{\widetilde{\varphi}_{ik}\})$$
$$\times p(\{\mu_{ik}, r_{ik}\} \mid \{\widetilde{m}_{ik}, \widetilde{\tau}_{ik}, \widetilde{\alpha}_{ik}, \widetilde{u}_{ik}\})$$

which combines the prior density and the likelihood measure calculated by the variational occupation probability model $\widetilde{q}(S,L \mid X)$. The variational distributions $\widetilde{q}(\{\pi_i\} \mid X)$, $\widetilde{q}(\{a_{im}\} \mid X)$ and $\widetilde{q}(\{\omega_{ik}\} \mid X)$ with new hyperparameters $\{\widetilde{\varpi}_i, \widetilde{\phi}_{im}, \widetilde{\varphi}_{ik}\}$ can be found in [12]. We focus on the variational inference for Gaussian means and precisions $\{\mu_{ik}, r_{ik}\}$, and optimize the minimax criterion in (7) which can be expressed as

$$\widetilde{q}(\{\mu_{ik}, r_{ik}\} \mid X) \propto p(\{\mu_{ik}, r_{ik}\} \mid \{m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}\}) \times \exp\left[\sum_{i,k,t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik} \log p(\mathbf{x}_t \mid \mu_{ik}, r_{ik})\right] \quad (11)$$

where Ψ_{BLM} is the set of support tokens in BLM-HMM training and $\widetilde{\varsigma}_{tik} = \widetilde{q}(s_t = i, l_t = k \mid X, \lambda)$ denotes the variational occupation probability of $\mathbf{x}_t$ staying at state s_t and mixture component l_t by using current estimate λ . This variational distribution is expressed as a product of normal-Wishart distributions with the updated hyperparameters

$$\widetilde{\tau}_{ik} = \tau_{ik} + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik} \quad (12)$$

$$\widetilde{m}_{ik} = \frac{\tau_{ik} m_{ik} + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik} \mathbf{x}_t}{\tau_{ik} + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik}} \quad (13)$$

$$\widetilde{u}_{ik} = \left[u_{ik} + \tau_{ik} m_{ik} m_{ik}^T + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik} \mathbf{x}_t \mathbf{x}_t^T - \left(\tau_{ik} + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{tik}\right) \widetilde{m}_{ik} \widetilde{m}_{ik}^T\right] \quad (14)$$

$$\widetilde{\alpha}_{ik} = \alpha_{ik} + \sum_{t \in \Psi_{\mathrm{BLM}}} \widetilde{\varsigma}_{ikt} \ . \quad (15)$$

In (12)-(15), the support vector tokens are considered in updating procedure. The BLM-HMM parameters are obtained by maximizing the variational distribution $\widetilde{q}(\mu_{ik}, r_{ik} \mid X)$ and finding $\hat{\mu}_{ik} = \widetilde{m}_{ik}$ and $\hat{r}_{ik}^{-1} = (\widetilde{\alpha}_{ik} - d)^{-1} \widetilde{u}_{ik}$ [4]. The graphical representations of BLM-HMM and its variational model are illustrated in Figures 1 and 2, respectively. In addition, the variational occupation probability $\widetilde{q}(S,L \mid X) = \{\widetilde{q}(s_t = i, l_t = k \mid X)\}$ is calculated in a VB-EM procedure [1][12][14]. In [12], the forward-backward algorithm was developed to calculate the variational occupation probability in HMMs. This study applies the Viterbi algorithm to decode the best state sequence $\hat{S} = \{\hat{s}_t\}$ and mixture component sequence $\hat{L} = \{\hat{l}_t\}$ by using variational posterior distributions $\widetilde{q}(\lambda \mid X)$. The Viterbi approximation is realized by $\widetilde{\varsigma}_{tik} = \widetilde{q}(s_t = i, l_t = k \mid X) = \delta(\hat{s}_t - i) \cdot \delta(\hat{l}_t - k)$.

3.3. Relations to SVM objective function

Importantly, we focus on developing BLM classifier for HMM-based speech recognition. In [8], the evidence framework of SVMs was addressed with a Bayesian interpretation of large margin classification. Here, we select support tokens by investigating BLM distance of $X_i = \{\mathbf{x}_{it}\}$ calculated by variational posterior distributions from two word models W_i and W_j . The distance at each frame $\mathbf{x}_{it}$ is yielded by

$$d_{\mathrm{BLM}}^{ij}(\mathbf{x}_{it}) = \log \widetilde{q}(\lambda_{W_i} \mid \mathbf{x}_{it}) - \log \widetilde{q}(\lambda_{W_j} \mid \mathbf{x}_{it}) = \log \widetilde{q}(\mu_{s_{it} l_{it}}, r_{s_{it} l_{it}} \mid \mathbf{x}_{it}) - \log \widetilde{q}(\mu_{s_{jt} l_{jt}}, r_{s_{jt} l_{jt}} \mid \mathbf{x}_{it}). \quad (16)$$

If $d_{\mathrm{BLM}}^{ij}(\mathbf{x}_{it}) > 0$, $\mathbf{x}_{it}$ is correctly recognized. If $d_{\mathrm{BLM}}^{ij}(\mathbf{x}_{it}) < 0$, $\mathbf{x}_{it}$ is misclassified. The support tokens for BLM classifier are selected from training data as in (3). Such a sparse learning process is comparable of implementing the variational occupation probability by a soft approximation [8]

$$\widetilde{q}(s_t = i, l_t = k \mid \mathbf{x}_{it}) \cong \exp(-[-d_{\mathrm{BLM}}^{ij}(\mathbf{x}_{it})]_+) = \exp(-\xi_t) \quad (17)$$

where $[b]_+ = b$ if $b > 0$ and $[b]_+ = 0$ if $b < 0$. This is different from the hard approximation by Viterbi algorithm. In (17), ξ_t is defined as a misclassification measure or the training error due to a frame $\mathbf{x}_{it}$. With this equation, the objective function for training Gaussian parameters (μ_{ik}, r_{ik}) of BLM-HMMs in (7) is rewritten as a variational distribution and can be related to SVM objective function by

$$-\log \widetilde{q}(S, L, \mu_{ik}, r_{ik} \mid X_i) = -\log \widetilde{q}(\mu_{ik}, r_{ik} \mid X_i) - \log \widetilde{q}(S, L \mid X_i) = \frac{\widetilde{\tau}_{ik}}{2} (\mu_{ik} - \widetilde{m}_{ik})^T r_{ik} (\mu_{ik} - \widetilde{m}_{ik}) + \sum_t \xi_t + \text{constant} . \quad (18)$$

Maximizing the variational posterior distribution with Gaussian mean and precision is equivalent to jointly minimizing the Mahalanobis distances using the updated hyperparameters and the sum of training errors [5][8]. The negative distance is known as a class margin. The discrimination information from incorrectly recognized samples is applied to assure good performance of using BLM-HMMs in speech recognition.

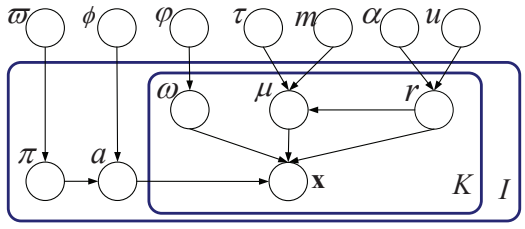

Figure 1 Graphical model of BLM-HMM

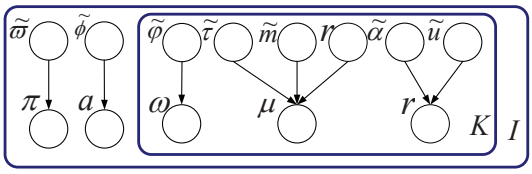

Figure 2 Graphical model of variational BLM-HMM

4. EXPERIMENTS

4.1. Experimental setup and implementation

In the experiments, the proposed method was evaluated by phone recognition using TIMIT database. There were 39 phone models trained by the maximum-likelihood (ML) method by using HTK tools. The 39-dimensional feature vector was extracted for each frame and was composed of 12 MFCCs and one log energy, and their first and second derivatives. 4614 utterances were selected as training data, and 1680 utterances were used as test data. Each phone was represented by a three-state HMM. The number of mixture components was changed to be 4, 8 and 16 for evaluation. The covariance matrix was assumed to be diagonal. For comparative study, MCE discriminative training [7] was implemented with a learning rate selected in a range between 0.002 and 0.004. The estimated MCE-HMMs were used as the seed models for training of LM-HMMs and BLM-HMMs. The initial hyperparameters in BLM-HMMs were selected from the optimal HMM estimates using $\tau_{ik} = n_{ik}$, $\alpha_{ik} = d + n_{ik}$, $m_{ik} = \hat{\mu}_{ik}$ and $u_{ik} = n_{ik}\hat{r}_{ik}^{-1}$, where n_{ik} is the number of frames staying at component k and state i in last iteration of Viterbi decoding. The variational inference was performed for Gaussian parameters while the remaining HMM parameters was unchanged. The parameter ε in selection of support tokens in LM-HMMs and BLM-HMMs was specified by $\varepsilon = \log(0.5n_i)$ where n_i denotes the number of frames in X_i . Using BLM-HMMs, the VB-EM procedure [1][14] was implemented by updating the variational occupation probability $\widetilde{\varsigma}_{tik}$ and the hyperparameters of variational distributions $\widetilde{\tau}_{ik}, \widetilde{m}_{ik}, \widetilde{u}_{ik}, \widetilde{\alpha}_{ik}$ in EM iterations. The convergence was met when the improvement rate of variational distributions $\widetilde{q}(\{\mu_{ik}, r_{ik}\} \mid X)$ was less than 1%.

Table 1 Comparison of phone error rates of HMMs trained by ML, MCE, LM and BLM methods

	ML	MCE	LM	BLM
$K = 4$	40.51%	39.97%	39.34%	38.36%
$K = 8$	39.17%	38.49%	38.27%	37.29%
$K = 16$	38.27%	37.58%	37.45%	36.43%

4.2. Experimental results

Table 1 compares the phone error rates by using different training criteria; ML, MCE, LM and BLM. The number of mixture components K in each HMM state is changed in the evaluation. We find that the discriminative training methods by MCE and LM consistently outperform baseline ML method. LM obtained slight improvement compared to MCE. Among these four training algorithms, the lowest phone error rate 36.43% is obtained by BLM-HMMs with 16 mixture components. The error reduction of BLM-HMMs compared to LM-HMMs is 2.72%, which is not significant. One reason is that the initial hyperparameters and the BLM-HMM parameters were calculated from the same training data. No validation data was used. Also, the initial hyperparameters were empirically determined. Such hyperparameters did not provide too much prior information for characterizing the variations in test environments. The improvement should be significant if the hyperparameters are estimated according to MacKay's evidence framework [8][9][14] and learned from the adaptation data.

5. CONCLUSIONS

This paper presented a Bayesian learning method for large margin HMM based speech recognition. The variational Bayesian approach was applied to build the empirical posterior distribution from training data. Importantly, the mechanism of SVM was embedded for sparse learning of HMM parameters. The preliminary experiments on TIMIT phone recognition showed the improvement by using BLM training compared to ML, MCE and LM training. The performance was considerably affected by the prior parameters. In the future, we are extensively evaluating the contributions of the margin and the priors and examining the performance in noisy speech recognition. We will develop the BLM-HMM adaptive training where the prior information is estimated from new training data, and also the BLM-HMM model adaptation where the model parameters are updated by using adaptation data. Due to the benefits of Bayesian learning, we will present the model selection solution to control the HMM structure and the goodness of support tokens in BLM-HMMs. The kernel method applied to BLM-HMMs shall be also investigated.

6. REFERENCES

[1] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD thesis, University College London, UK, 2003.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science, 2006.

[3] J.-T. Chien and G.-H. Liao, "Transformation-based Bayesian predictive classification using online prior evolution", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, pp. 399-410, 2001.

[4] J.-T. Chien and S. Furui, "Predictive Hidden Markov Model Selection for Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 377-387, 2005.

[5] J.-T. Chien and J.-C. Chen, "Recursive Bayesian linear regression for adaptive classification", *IEEE Trans. Signal Processing*, 2009.

[6] H. Jiang, X. Li and C. Liu, "Large margin hidden Markov models for speech recognition'', *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp.1584-1595, 2006.

[7] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 2, pp. 257–265, 1997.

[8] J. T.-Y. Kwok, "The evidence framework applied to support vector machines", *IEEE Trans. Neural Networks*, vol. 11, no. 5, pp.1162-1173, 2000.

[9] D. J. C. MacKay, "Bayesian interpolation", *Neural Computation*, vol. 4, no. 3, pp. 415-447, 1992.

[10] F. Sha and L. K. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition", in *Proc. of ICASSP*, vol. 1, pp. 265-268, 2006.

[11] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[12] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 4, pp. 365-381, 2004.

[13] K. Yu and M. J. F. Gales, "Bayesian Adaptive Inference and Adaptive Training", *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1932-1943, 2007.

[14] Y. Zhang, P. Liu, J.-T. Chien and F. Soong, "An evidence framework for Bayesian learning of continuous-density hidden Markov models", in *Proc. of ICASSP*, 2009.