AN EVIDENCE FRAMEWORK FOR BAYESIAN LEARNING OF CONTINUOUS-DENSITY HIDDEN MARKOV MODELS

Yu Zhang[1,2], *Peng Liu*[1], *Jen-Tzung Chien*[3] *and Frank Soong*[1]

[1]Microsoft Research Asia, Beijing, China
[2]Shanghai Jiao Tong University, Shanghai, China
[3]National Cheng Kung University, Tainan, Taiwan
zhangyu@apex.sjtu.edu.cn,{pengliu,frankkps}@microsoft.com, jtchien@mail.ncku.edu.tw

ABSTRACT

We present an evidence Bayesian framework, which can learn both the prior distributions and posterior distributions from data, for continuous-density hidden Markov models (CDHMM). The goal of this study is to build the *regularized* CDHMMs to improve model generalization, and achieve desirable recognition performance for unknown test speech. Under this framework, we develop an EM iterative procedure to estimate the marginal distribution or the evidence function for *exponential family distributions*. By adopting the variational Bayesian inference, we derive an empirical Bayesian solution to CDHMM parameters and their hyperparameters. Such a regularized CDHMM compensates the model uncertainty and the ill-posed conditions. Compared with maximum likelihood (ML) or other Bayesian approaches with heuristic hyperparameters, the proposed approach can utilize available data more effectively. The experiments on noisy speech recognition using Aurora2 show that the proposed Bayesian approach performs better than the baseline ML CDHMMs especially with mismatched test data or limited training data.

Index Terms— hidden Markov model, evidence framework, variational Bayesian

1. INTRODUCTION

Robust acoustic modeling plays an important role for speech recognition when the collected training data are sparse and noisy. The ill-posed conditions severely hampers in the trained hidden Markov models (HMMs) to recognize test data robustly and model uncertainty deteriorates the recognition performance. Accordingly, we are motivated to present an evidence framework of continuous-density HMMs (CDHMMs). This framework assures the model generalization by fulfilling Bayesian regularization theory. Under this framework, the marginalization of likelihood function over the uncertainty of HMM parameters is calculated, and acts as the objective function to be optimized to build the regularized CDHMMs. Compared with the point estimate of CDHMMs in maximum likelihood (ML) training, the regularized CDHMMs are known as the distribution estimate, which is inherently robust to the variations of model distributions. This idea fulfills Mackay's evidence framework [1, 2]. Therefore, the regularized CDHMMs can achieve better classification performance by using insufficient or noisy training data.

In implementing model regularization, the selection of suitable prior distribution or its hyperparameters is critical. In general, there are two approaches, subjective Bayesian and objective Bayesian, which are useful to select priors. In former approach, the priors are built based on some background knowledge while in the latter approach, also referred as the empirical Bayes, the priors are automatically learned from training data. In speech recognition systems using Bayesian learning, it is popular to estimate hyperparameters based on some intuitive data statistics and optimization metrics [3, 4]. The collection of validation data is usually required. However, under the evidence framework, the hyperparameters are selected from training data, and the resulting evidence is maximized to assure model generalization.

In previous studies, the evidence framework [1, 2] has been applied to linear regression model, support vector regression model [5], and neural networks. This study applies the evidence frameworks to exponential family distributions and CDHMMs, and shows their effectiveness in characterizing the model uncertainty from data. Different from [1, 5, 2], a marginal likelihood using CDHMMs is calculated without a Laplace approximation. Owing to the missing labels of state and mixture component, we present a variational expectation-maximization (EM) algorithm [6, 7] to estimate the hyperparameters of Gaussian mean vector, covariance matrix, and mixture weights. These hyperparameters are iteratively updated by EM procedure according to the variational inference with decomposition of CDHMMs and missing labels. We also illustrate this evidence framework by using graphical models [8] of the regularized CDHMMs and their variational models. In the experiments of noisy speech recognition, the proposed method outperforms baseline ML method, and the improvement is significant in presence of insufficient training data.

2. EVIDENCE FRAMEWORK FOR EXPONENTIAL FAMILY DISTRIBUTIONS

We begin by discussing the evidence framework for the basic component distributions used in CDHMMs. Most of them, such as the Gaussian distribution, multinomial distribution for mixture weights and transition probabilities, can be grouped into the exponential family. Hence, we study the generic solution for the exponential family. Supposing that K distributions, which take the same form but respectively governed by parameters $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_K$, share an identical prior distribution governed by the hyperparameter $\boldsymbol{\eta}$. (Obviously, setting individual priors for them is a special case.) Based upon to the evidence framework, we can obtain the best $\hat{\boldsymbol{\eta}}$ in the sense of

The work was done during the first author's internship in Microsoft Research Asia.

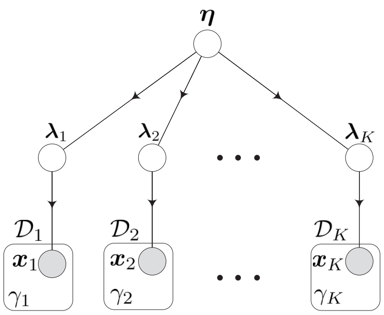

Fig. 1. A graphical model of the evidence framework

maximum type II likelihood:

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \prod_{i=1}^{K} \int p(\mathcal{D}_i|\boldsymbol{\lambda}_i)p(\boldsymbol{\lambda}_i|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\lambda}_i \tag{1}$$

where $\mathcal{D}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,\gamma_i}\}$ represents the observed data set of the i^{th} distribution.

A graphical representation of such a problem is shown in Fig. 1. We can observe that Eq.(1) can be regarded as a maximization of the data likelihood with respect to $\boldsymbol{\eta}$, by marginalizing out the model parameters $\boldsymbol{\lambda}_i$. Hence, we solve it with the EM algorithm by treating $\boldsymbol{\lambda}_i$ as latent variables. In the E-step, we evaluate the following auxiliary function:

$$\mathcal{Q}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\text{old}}) = \sum_{i=1}^{K} \int p(\boldsymbol{\lambda}_i|\mathcal{D}_i, \boldsymbol{\eta}^{\text{old}}) \ln p(\mathcal{D}_i, \boldsymbol{\lambda}_i|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\lambda}_i \tag{2}$$

As shown in the graphic model, $\mathcal{D}_i$ and $\boldsymbol{\eta}$ are independent given $\boldsymbol{\lambda}_i$, i.e., $\mathcal{D}_i \perp \boldsymbol{\eta}|\boldsymbol{\lambda}_i$. Based on this property, we can simplify the logarithm term in the integrand:

$$\ln p(\mathcal{D}_i, \boldsymbol{\lambda}_i|\boldsymbol{\eta}) = \ln p(\mathcal{D}_i|\boldsymbol{\lambda}_i) + \ln p(\boldsymbol{\lambda}_i|\boldsymbol{\eta}) \tag{3}$$

With an adopted conjugate prior, the posterior $p(\mathcal{D}_i|\boldsymbol{\lambda}_i, \boldsymbol{\eta})$ takes the same form as its prior. Hence, we can represent it by $p(\boldsymbol{\lambda}_i|\tilde{\boldsymbol{\eta}}_i^{\text{old}})$, where $\tilde{\boldsymbol{\eta}}_i^{\text{old}}$ is the posterior parameter of $\boldsymbol{\lambda}_i$ after observing the data set $\mathcal{D}_i$. In this context, by substituting Eq.(3) into Eq.(2), we have:

$$\mathcal{Q}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\text{old}}) = \sum_{i=1}^{K} \int p(\boldsymbol{\lambda}_i|\tilde{\boldsymbol{\eta}}_i^{\text{old}}) \ln p(\boldsymbol{\lambda}_i|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\lambda}_i + C \tag{4}$$

where C is a constant independent of $\boldsymbol{\eta}$.

In the M-step, we maximize $\mathcal{Q}$ to find $\boldsymbol{\eta}^{\text{new}}$ based upon the concrete form of $p(\boldsymbol{x}_i|\boldsymbol{\lambda}_i)$. In this study, aiming at a more general solution, we focus on distributions in the exponentially family, which can be represented in a general form[1]:

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}_i) = h(\boldsymbol{x}_i)g(\boldsymbol{\lambda}_i)\exp[\boldsymbol{\lambda}_i^{\top} u(\boldsymbol{x}_i)] \tag{5}$$

where $h(\boldsymbol{x})$ is some function of $\boldsymbol{x}$, $g(\boldsymbol{\lambda})$ is a normalization term and $u(\boldsymbol{x})$ is sufficient statistics. To facilitate the mathematical derivation, we choose the conjugate prior in Bayesian learning:

$$p(\boldsymbol{\lambda}_i|\boldsymbol{\chi}_i, \nu_i) = f(\boldsymbol{\chi}_i, \nu_i)g(\boldsymbol{\lambda}_i)^{\nu_i}\exp(\nu_i\boldsymbol{\lambda}_i^{\top}\boldsymbol{\chi}_i) \tag{6}$$

For convenience, here we decompose the hyperparameter $\boldsymbol{\eta}$ into $(\boldsymbol{\chi}, \nu)$ which are hyperparameters of exponential distribution family and $f(\boldsymbol{\chi}, \nu)$ is a normalization term to ensure a valid pdf.

In the E-step, we can calculate the posterior distribution $\boldsymbol{\lambda}_i$ with sufficient statistics and the hyperparameter:

$$\tilde{\nu}_i \;=\; \nu + \gamma_i, \quad \tilde{\boldsymbol{\chi}}_i = \frac{\sum_{n=1}^{\gamma_i} \boldsymbol{u}(\boldsymbol{x}_{i,n}) + \nu\boldsymbol{\chi}_i}{\tilde{\nu}_i} \tag{7}$$

By substituting Eqs.(5), (6) and (7) into Eq.(4) and maximizing it, we obtain $\boldsymbol{\eta}^{\text{new}}$ in the M-step:

$$\langle \boldsymbol{\lambda}, \ln[g(\boldsymbol{\lambda})]\rangle_{\boldsymbol{\eta}^{\text{new}}} = \frac{1}{K}\sum_{i=1}^{K} \langle \boldsymbol{\lambda}, \ln[g(\boldsymbol{\lambda})]\rangle_{\tilde{\boldsymbol{\eta}}_i^{\text{old}}} \tag{8}$$

In general, this implicit equation can be solved by the Newton method. As shown below, for most of the parameters used in CDHMMs, we have closed-form solutions.

3. EVIDENCE FRAMEWORK FOR CDHMMS

Now we study the evidence framework for CDHMMs. Because the most popular output distributions used in CDHMMs are Gaussian mixture models (GMMs), we consider this specific case in this paper. However, with the general solution proposed in the above section, we can easily extend the results to other kinds of output distributions.

In the training phase, when applying the evidence framework to CDHMMs, we cannot derive a concise EM algorithm to jointly deal with the latent variables of the model parameters as well as the hidden Gaussian component sequence. As we know, in Bayesian training, various approximated approaches such as variational Bayes [9] and quasi-Bayes [10] has been studied to approximate the joint posterior. Here we follow the variational Bayesian approach.

3.1. Variational Bayesian Training for CDHMMs

In CDHMM with GMM output distributions, given the sequential observation $\boldsymbol{x}_1^T$, we calculate $p(\boldsymbol{\lambda}, \boldsymbol{q}_1^T|\boldsymbol{x}_1^T)$ in the E-step. Here $\boldsymbol{\lambda}$ denotes the CDHMM parameters set, and $\boldsymbol{q}_1^T$ denotes the underlying Gaussian component sequence. Because exact evaluation of the posterior is intractible, in variational Bayesian, we assume the posterior can be decomposed into:

$$p(\boldsymbol{\lambda}, \boldsymbol{q}_1^T|\boldsymbol{x}_1^T, \boldsymbol{\eta}^{\text{old}}) \approx p(\boldsymbol{\lambda}|\boldsymbol{x}_1^T, \boldsymbol{\eta}^{\text{old}})p(\boldsymbol{q}_1^T|\boldsymbol{x}_1^T, \boldsymbol{\eta}^{\text{old}}) \tag{9}$$

It leads to a minor revision of the conventional Baum-Welch algorithm for estimating CDHMM, and the difference is to use the following quantity instead of the corresponding component distribution probability:

$$p'(\boldsymbol{x}_t|\boldsymbol{q}_t = i) = \exp\left\{\int \ln p(\boldsymbol{x}_t|\boldsymbol{\lambda}_i)p(\boldsymbol{\lambda}_i|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\lambda}_i\right\} \tag{10}$$

We shall give the concrete form of it when discussing the CDHMM parameters in the following section. Based upon it, occupancies of all the Gaussian components can be obtained in the Baum-Welch procedure to collect statistics. Given the statistics, it is straightforward to apply the proposed evidence based Bayesian training for the assumed exponential family.

Accordingly, the resultant occupancy $\gamma_{it} = p(\boldsymbol{q}_t = i|\boldsymbol{x}_1^T, \boldsymbol{\lambda}, \boldsymbol{\eta}^{old})$ for each Gaussian components i at time t, derived by the Baum-Welch algorithm, can be used to collect the following statistics:

$$\begin{aligned} \gamma_i \;&=\; \sum_{t=1}^{T}\gamma_{it}, \quad \gamma_i(\boldsymbol{x}) = \sum_{t=1}^{T}\gamma_{it}\boldsymbol{x}_t \\ \gamma_i(\boldsymbol{x}\boldsymbol{x}^{\top}) \;&=\; \sum_{t=1}^{T}\gamma_{it}\boldsymbol{x}_t\boldsymbol{x}_t^{\top} \end{aligned} \tag{11}$$

3.2. CDHMM Parameter Update

With the statistics collected in the E-step of variational Bayesian procedure, we can apply the EM based maximum evidence algorithm proposed in section 2 to CDHMMs parameters, and the concrete update algorithm is shown below.

To give a clear view of the algorithm, we first give a conceptual pseudo code of the algorithm. The whole training procedure is shown in Table 1. Without setting any knowledge-based prior, the process can automatically train Bayesian models on a given data set by iteratively updating priors and corresponding posteriors. The update formulas for concrete CDHMM parameters is provided in the following sections.

Table 1. The pseudo code of evidence framework based Bayesian training for CDHMMs

```
iteration loop:
    variational E-step:
        conduct Baum-Welch on the training set, by using
        Eq.(10) instead of Gaussian probabilities, and
        collect statistics $\gamma_i, \gamma_i(\boldsymbol{x}), \gamma_i(\boldsymbol{x}\boldsymbol{x}^\top)$
    variational M-step:
        maximum evidence E-step:
            calculate $\tilde{\boldsymbol{\eta}}_i^{\text{old}}$ for all the CDHMM parameters
        maximum evidence M-step:
            solve $\boldsymbol{\eta}^{\text{new}}$ with Eq.(15) and
while the evidence gap is larger than a threshold
```

3.2.1. *Gaussian parameters*

For Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{R}_i^{-1})$, we have $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{R}_i^{-1}\}$, and the corresponding conjugate prior takes a Gaussian-Wishart form as:

$$p(\boldsymbol{\mu}_i, \boldsymbol{R}_i|\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\mu}_i; \boldsymbol{\mu}_0, \beta_0^{-1}\boldsymbol{R}_i^{-1})\mathcal{W}(\boldsymbol{R}_i; \boldsymbol{R}_0, \nu_0) \qquad (12)$$

where the hyperparameter $\boldsymbol{\eta}$ is collectively defined by $\{\boldsymbol{\mu}_0, \boldsymbol{R}_0, \beta_0, \nu_0\}$. Accordingly, in VB training, the revised probability of Eq.(10) can be calculated as:

$$\begin{aligned} \ln p'(\boldsymbol{x}_t|\boldsymbol{q}_t = i) \quad &= \quad -\frac{1}{2}\{D(\ln\pi + \frac{1}{\tilde{\beta}_i} - \Psi(\frac{\tilde{\nu}_i}{2}) + \ln\tilde{\nu}_i) \\ - \ln|\tilde{\boldsymbol{R}}_i| \quad &+ \quad (\boldsymbol{x}_t - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{R}}_i(\boldsymbol{x}_t - \tilde{\boldsymbol{\mu}}_i)\} \qquad (13) \end{aligned}$$

where $\Psi(\nu) \equiv \frac{\partial}{\partial\nu}\ln\Gamma(\nu)$ is a digamma function, and D is the dimension of $\boldsymbol{x}$.

By aligning Gaussian distribution with the general exponential form of Eq. (5), and substituting the concrete form into Eqs. (7) and (8), we obtain the maximum evidence EM formulas for Gaussian parameters:

1. Maximum evidence E-step:

$$\begin{aligned} \tilde{\beta}_i^{\text{old}} &= \beta_0^{\text{old}} + \gamma_i, \quad \tilde{\nu}_i^{\text{old}} = \nu_0^{\text{old}} + \gamma_i \\ \tilde{\boldsymbol{\mu}}_i^{\text{old}} &= \frac{\beta_0^{\text{old}}\boldsymbol{\mu}_0^{\text{old}} + \gamma_i(\boldsymbol{x})}{\tilde{\beta}_i^{\text{old}}} \\ \tilde{\boldsymbol{R}}_i^{\text{old}} &= \tilde{\nu}_i\Big\{\nu_0^{\text{old}}(\boldsymbol{R}_0^{\text{old}})^{-1} + \gamma_i(\boldsymbol{x}\boldsymbol{x}^\top) - \frac{\gamma_i(\boldsymbol{x})\gamma_i^\top(\boldsymbol{x})}{\gamma_i} + \\ &\quad \frac{\beta_0^{\text{old}}}{\gamma_i\tilde{\beta}_i^{\text{old}}}\left[\gamma_i(\boldsymbol{x}) - \gamma_i\boldsymbol{\mu}_0\right]\left[\gamma_i(\boldsymbol{x}) - \gamma_i\boldsymbol{\mu}_0\right]^\top\Big\}^{-1} \qquad (14) \end{aligned}$$

2. Maximum evidence M-step:

$$\begin{aligned} \boldsymbol{R}_0^{\text{new}} &= \frac{1}{K}\sum_{i=1}^{K}\tilde{\boldsymbol{R}}_i^{\text{old}}, \quad \boldsymbol{\mu}_0^{\text{new}} = \frac{(\boldsymbol{R}_0^{\text{new}})^{-1}}{K}\sum_{i=1}^{K}\tilde{\boldsymbol{R}}_i^{\text{old}}\tilde{\boldsymbol{\mu}}_i^{\text{old}} \\ \frac{1}{\beta_0^{\text{new}}} &= \frac{1}{K}\left[\sum_{i=1}^{K}(\frac{1}{\tilde{\beta}_i^{\text{old}}} + \frac{(\boldsymbol{\mu}_0^{\text{new}} - \tilde{\boldsymbol{\mu}}_i^{\text{old}})^\top\tilde{\boldsymbol{R}}_i^{\text{new}}(\boldsymbol{\mu}_0^{\text{new}} - \tilde{\boldsymbol{\mu}}_i^{\text{old}})}{D})\right] \\ \nu_0^{\text{new}} &= \Phi^{-1}\left[\frac{1}{K}\sum_{i=1}^{K}(\Phi(\tilde{\nu}_i^{\text{old}}) + \frac{1}{D}\ln\frac{|\tilde{\boldsymbol{R}}_i^{\text{old}}|}{|\boldsymbol{R}_0^{\text{new}}|})\right] \qquad (15) \end{aligned}$$

where $\Phi(\nu) \equiv \Psi(\nu/2) - \ln(\nu/2)$.

3.2.2. *Mixture weights*

The mixture weights used in GMMs follow a multinomial distribution, which is also a member of exponential family. By adopting the corresponding conjugate prior, i.e., Dirichlet distribution, we can also use the general solution of Eqs. (7) and (8) to solve it. Because of space limitation, the detailed solution is omitted here.

3.3. Bayesian predictive classification

In testing, given a Bayesian version of CDHMMs, we should make use of the posterior distribution of model parameters instead of their point estimates. The method is usually referred to as Bayesian predictive classification (BPC) [4]. Strictly apply BPC in decoding is cumbersome, and in this study we follow the approximation used in [7], which marginalizes the model parameter on each individual frame and calculates the probability of the resultant Student-t distribution instead of the original Gaussian distribution [3, 7].

4. EXPERIMENTS

The evidence framework of CDHMMs was tested on Aurora2, a connected digit recognition task [11]. Meanwhile, whole-word HMMs were built for each of the eleven digits ranging from 'zero' to 'nine', and 'oh', and 3-component GMMs were adopted as the output distributions for all the states, with all the covariance matrices set to be diagonal. In Bayesian training, all the Gaussian components belonging to the same GMM share an identical prior distribution. Because we mainly focus on the Gaussian components in this study, we didn't apply evidence framework to mixture weights and transition probabilities and only set fixed prior for them, following [7].

In Table 2, we compare the word recognition accuracies of ML trained models and evidence trained models. It can be observed when the mismatch between training and testing set is small, i.e., at a low signal-to-noise (SNR) ratio, the ML training achieves slightly better performance. But as the mismatch becomes large, the maximum evidence Bayesian approach yields better results.

Because with the full training set, data is sufficient for the relatively small number of whole word models, the gap between ML and Bayesian training is not distinctively different. Hence, we also studied the difference between ML and proposed training in case of insufficient training data. First, we compared the average word accuracy on the testing set of three systems, in both clean training and multi training cases, and plotted the results in Fig. 2 and Fig. 3, respectively. The three systems are: 1. Evidence framework based training Beyesian training; 2. Conventional Bayesian training with manually set prior using the proposed method in [7]. In this method, $\boldsymbol{\mu}_0, \boldsymbol{R}_0$ are derived with data statistics [7], and β_0, ν_0 are experimentally determined. We tried

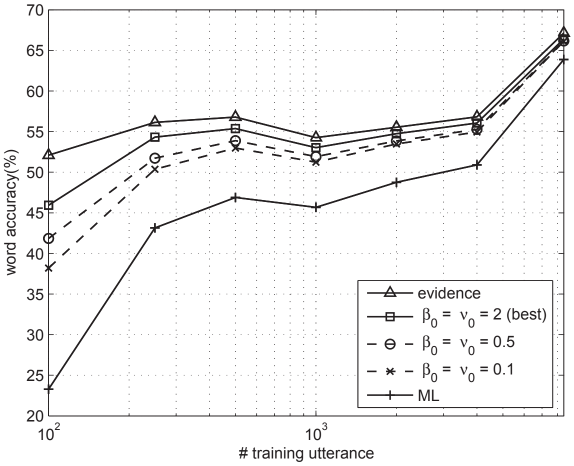

Fig. 2. Performance comparison with a variable size of clean training data

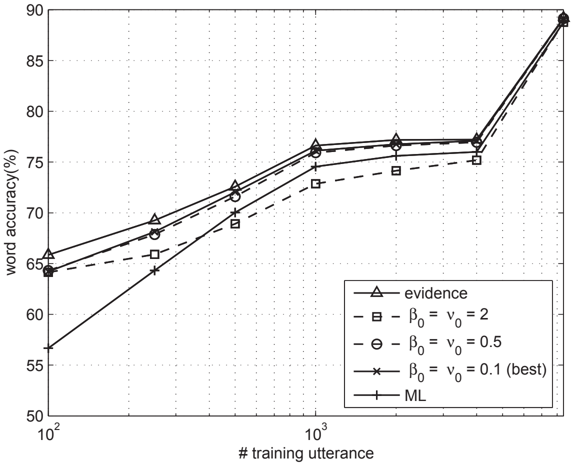

Fig. 3. Performance comparison with a variable size of multi-conditional training data

$\beta_0 = \nu_0 = 0, 0.001, 0.05, 0.1, 0.5, 1, 2, 10$ but only plotted the best result in solid line, as well as other two representative results in dashed line. We can observe that the evidence framework outperforms not only the ML training, but also the state-of-the-art Bayesian approach with manually set priors. Note that in clean training and multi training, the best β_0, ν_0 differs significantly, and inappropriate setting of them can sometimes lead to even worse performance than the ML system. Obviously, it is hard to make a good suggestion on how to manually set the hyperparameters. However, the evidence framework is always shown the best performance without any heuristic setting of hyperparameter.

Table 2. Word accuracy (%) comparison on Aurora2

	clean train		multi train	
SNR	ML	evidence	ML	evidence
clean	**99.15**	98.98	**98.46**	98.42
20db	**97.23**	97.16	97.66	**97.79**
15db	92.31	**92.70**	97.05	**97.24**
10db	75.05	**77.15**	95.31	**95.64**
5db	42.21	**44.73**	89.14	**89.68**
0db	22.49	**22.59**	64.75	**65.62**
average	65.86	**66.87**	90.86	**91.20**

5. CONCLUSIONS AND FUTURE WORK

Based upon the evidence framework, we propose a training algorithm for CDHMMs, which automatically learns the priors as well as their posteriors from data. We first derive an EM solution for the exponential family distributions, and extend the algorithm to deal with CDHMMs by using an variational Bayesian procedure. Experimental results show that in comparison with ML training, the evidence framework leads to better regularization of the models, hence better robustness in case of mismatched or limited training data.

Note that the evidence framework is promising for insufficient training data. In our future research, we shall investigate the proposed algorithm in more complex tri-phone HMMs to find a better trade-off between number of model parameters and reliable estimates.

6. REFERENCES

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

[2] D. J. C. MacKay, “Bayesian interpolation”, *Neural Computation*, 4(3), pp. 415-447, 1992.

[3] J.-T. Chien and S. Furui, “Predictive hidden Markov model selection for speech recognition”, *IEEE Trans. SAP*, 13(3), pp. 377-387, 2005.

[4] Q. Huo, H. Jiang and C.-H. Lee, “A Bayesian predictive classification approach to robust speech recognition”, *Proc. of ICASSP*, pp. 1547-1550, 1997.

[5] J. T.-Y. Kwok, “The evidence framework applied to support vector machines”, *IEEE Trans. Neural Networks*, 11(5), pp.1162-1173, 2000.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statist. Society (B)*, 39, pp. 1-38, 1977.

[7] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda, “Variational Bayesian estimation and clustering for speech recognition”, *IEEE Trans. SAP*, 12(4), pp. 365-381, 2004.

[8] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L Saul, “An introduction to variational methods for graphical models”, *Machine Learning*, 37, pp. 183-233, 1999.

[9] H. Attias, “A Variational Bayesian Framework for Graphical Models”, *NIPS12*, MIT Press, 2000.

[10] Q. Huo and C.-H. Lee, “A study of on-line quasi-Bayes adaptation for CDHMM-based speech recognition”, *Proc. of ICASSP*, pp. 705-708, 1996.

[11] H. G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition under noisy conditions”, *Proc. of ISCAITRW ASR2000*, pp. 181-188, 2000.