ABSTRACT
We introduce Bayesian sensing hidden Markov models (BS-HMMs) to represent speech data based on a set of state-dependent basis vectors. By incorporating the prior density of sensing weights, the relevance of a feature vector to different bases is determined by the corresponding precision parameters. The BS-HMM parameters, consisting of the basis vectors, the precision matrices of sensing weights and the precision matrices of reconstruction errors, are jointly estimated by maximizing the likelihood function, which is marginalized over the weight priors. We derive recursive solutions for the three parameters, which are expressed via maximum a posteriori estimates of the sensing weights. Experimental results on an LVCSR task show consistent gains over conventional HMMs with Gaussian mixture models for both ML and discriminative training scenarios.

Index Terms— Speech recognition, Bayesian learning, basis representation, acoustic model

1. INTRODUCTION
There have been many technologies successfully developed to represent or encode a general signal based on a set of over-determined basis vectors. It is beneficial to use a relatively small set of relevant basis vectors to represent the underlying signal. The overfitting problem in such a transform coding is alleviated. Model regularization can be thus achieved in pattern recognition systems, e.g. face recognition and speech recognition, where the overtraining issue always exists in presence of heterogeneous environments using sparse data, noisy data, unlabeled data, etc. The regularized model is crucial to achieve good recognition performance on unknown test data. Recently, sparse coding techniques have been developed for feature extraction of speech signals based on an over-determined dictionary using the exemplar-based method [1],[2]. This paper proposes a new Bayesian representation of speech data and introduces Bayesian sensing HMMs (BS-HMMs) for large vocabulary continuous speech recognition (LVCSR).

In standard HMMs, the sequential data in a Markov state are assumed to be conditionally independent and are represented by state-dependent Gaussian mixture models (GMMs). The corresponding state sequence is determined according to the accumulated likelihood function and the state transition probability. However, the maximum likelihood (ML) estimates of HMM parameters are prone to be overtrained. We question whether GMMs are suitable for the representation of training data as well as for the generalization to unknown test data. Predictive HMMs [3] were accordingly proposed to compensate the overfitting problem by incorporating the uncertainties of HMM parameters, expressed by prior distributions, in a Bayesian model comparison [4]. Another approach was given by buried Markov models [5] which relaxed the conditional independence assumption for the representation of speech features. A set of state-dependent basis vectors was trained to express the conditionally dependent feature vectors. In yet another approach, subspace Gaussian mixture models [6] were constructed to represent speech features by using the state-dependent weights and a common large-scale GMM structure. The feature representation was seen as sensing based on different subspaces of a global GMM.

In this study, we address the basis representation of speech features for hidden Markov modeling and present the Bayesian sensing framework to ensure model regularization for speech recognition. The resulting BS-HMMs are constructed by a set of basis vectors, the precision matrix of sensing weights, and the precision matrix of reconstruction errors. The precision matrix of weights naturally reflects how relevant the input feature is encoded by the basis vectors similar to the perspective of relevance vector machine (RVM) [7]. Importantly, we maximize the marginal likelihood of the training data over random weights and jointly estimate the three sets of parameters. Multivariate solutions are derived by maximum likelihood (ML) type II estimation and expressed through recursive formulas. These formulas are interpreted in terms of the mean vector and covariance matrix of the a posteriori distribution of the sensing weights. The maximum a posteriori (MAP) estimate of the sensing weights plays a central role in BS-HMMs. Experimental results on an LVCSR task show consistent improvements over standard HMMs with Gaussian mixture models.

2. BAYESIAN SENSING HIDDEN MARKOV MODELS
2.1. Model Construction
The concept of Bayesian sensing is incorporated within the HMM framework by viewing the D-dimensional feature vector $x_t \in \mathbb{R}^D$ as a measurement based on a set of state-dependent basis vectors $\Phi_i = [\phi_{i1}, \ldots, \phi_{iN}]$ for state $i$, which may suffer from an overfitting problem. We assume that the reconstruction error between measurement $x_t$ and its representation $\Phi_i w_t$, where $w_t = [w_{t1}, \ldots, w_{tN}]^T$, is Gaussian distributed with zero mean and a state-dependent precision matrix $R_i$. Furthermore, the sensing weight vector $w_t$ is assumed to be Gaussian distributed with zero mean and a state-dependent precision matrix $A_i$. Accordingly, the state likelihood function is expressed as

$$p(x_t|w_t, \lambda) \propto |R_i|^{1/2} \exp \left[ -\frac{1}{2} (x_t - \Phi_i w_t)^T R_i (x_t - \Phi_i w_t) \right]$$

(1)

Each frame $x_t$ is generated by the BS-HMM parameters $\lambda = \{\pi_i, a_{ik}, A_i, \Phi_i, R_i\}$ consisting of initial state probabilities $\{\pi_i\}$,
state transition probabilities \( \{a_{ik}\} \), precision matrices of sensing weights \( \{A_i\} \), basis vectors \( \{\Phi_i\} \) and precision matrices of residuals \( \{R_i\} \). In this study, the BS-HMM parameters are estimated by performing ML type II estimation, namely maximizing the marginal likelihood calculated by marginalizing the uncertainty of the weight vectors \( w_i \). Without loss of generality, the marginalization over the other parameters is not considered.

The sensing weight vector \( w_i \) is prone to be sharply distributed. This is because a multivariate \( t \) distribution with zero mean is naturally formed by considering a Wishart distribution as the prior of the precision matrix of \( w_i \). Such a hierarchical prior leads to a sparse representation of the speech feature \( x_t \) by using a small amount of relevant basis vectors of \( \Phi_i \). Bayesian dictionary learning for HMMs is achieved in a similar way to Bayesian compressive sensing [8]. The spirit of RVM [7] is embedded in BS-HMMs where each speech feature vector is represented by relevant basis vectors. Irrelevant basis vectors are de-emphasized.

### 2.2. Auxiliary Function

To find the ML type II estimates of BS-HMM parameters, we apply the EM algorithm and maximize the auxiliary function

\[
Q(\lambda|\lambda^{(k)}) = \sum_{i} \sum_{t} \gamma_t(i) \log \int_{R^N} p(x_t|w_t, \lambda_t) p(w_t|A_i) dw_t,
\]

where \( \gamma_t(i) = p(s_t = i|X, \lambda^{(k)}) \) is the posterior probability of being in state \( i \) at time \( t \) given the observation sequence \( X = \{x_t\} \) and the current parameters \( \lambda^{(k)} \). The key issue in this E-step is to calculate the marginal likelihood or Bayesian predictive likelihood \( p(x_t|\lambda_t) \) which is marginalized over the weight vector \( w_t \) and is proportional to

\[
\int_{R^N} |R_t|^{1/2} \exp \left[ -\frac{1}{2}(x_t - \Phi_t w_t)^T R_t(x_t - \Phi_t w_t) \right] \cdot |A_t|^{1/2} \exp \left[ -\frac{1}{2}w_t^T A_t w_t \right] dw_t
\]

\[
= |R_t|^{1/2} |A_t|^{1/2} \exp \left[ -\frac{1}{2}x_t^T R_t x_t \right] \int_{R^N} \exp \left[ -\frac{1}{2}w_t^T \Sigma^{-1}_t w_t - (x_t^T R_t x_t) \Sigma^{-1}_t w_t \right. \]

\[
+ (x_t^T R_t x_t) \Sigma^{-1}_t (x_t^T R_t x_t) \]

\[
\left. - (x_t^T R_t x_t) \Sigma^{-1}_t (x_t^T R_t x_t) \right] dw_t \]

\[
\propto |R_t|^{1/2} |A_t|^{1/2} |\Sigma_t|^{1/2} \exp \left[ -\frac{1}{2}x_t^T (R_t - R_t \Phi_t \Sigma_t \Phi_t^T R_t) x_t \right] \]

\[
= |R_t|^{1/2} |A_t|^{1/2} |\Sigma_t|^{1/2} \exp \left[ -\frac{1}{2}x_t^T (R_t - \Phi_t \Sigma_t \Phi_t^T R_t) x_t \right] \]

\[
\frac{1}{2} \sum_{t} \gamma_t(i) \log |R_t| + |A_t| + |\Sigma_t| - \frac{1}{2}x_t^T R_t x_t + m_t \Sigma_t^{-1} m_t
\]

(5)

The new notations \( m_t \) and \( \Sigma_t \) are the mean vector and the covariance matrix of the posterior distribution \( p(w_t|x_t, \lambda_t) \), respectively. The precision matrix for \( w_t \) is changed from \( A_t \) of the a priori density \( p(w_t) \) to \( \Phi_t^T R_t \Phi_t + A_t \) of the a posteriori distribution \( p(w_t|x_t) \). The difference \( \Phi_t^T R_t \Phi_t \) comes from the likelihood function \( p(x_t|w_t) \) and is caused by the CS measurement \( \Phi_t w_t \). This is reasonable since Bayesian learning performs subjective inference which increases the model precision.

### 2.3. Solutions to BS-HMM Parameters

The ML type II BS-HMM parameters are estimated in the M step. By plugging (3) into the auxiliary function in (2), new BS-HMM parameters \( \lambda = \{\pi_t, a_{ik}, A_t, \Phi_t, R_t\} \) are estimated by maximizing the auxiliary function for state \( i \)

\[
\sum_{t} \gamma_t(i) \log |R_t| + |A_t| + |\Sigma_t| - \frac{1}{2}x_t^T R_t x_t + m_t \Sigma_t^{-1} m_t
\]

(6)

The value of \( \lambda_t \) that maximizes the auxiliary function satisfies

\[
\hat{A}_t = \Sigma_t + \frac{1}{\sum_t \gamma_t(i)} \frac{\partial}{\partial \hat{A}_t} \log |R_t| \Phi_t^T R_t + A_t = 2R_t \Phi_t^T R_t \Sigma_t \Phi_t^T R_t x_t
\]

(7)

Note that (7) is an implicit or recursive solution to \( A_t \) since the RHS is a function of \( A_t \).

To find the new estimate of the basis vectors \( \Phi_t \), we maximize (5) by taking the gradient of the terms related to \( \Phi_t \) and setting it to zero which leads us to

\[
\frac{\partial}{\partial \Phi_t} \log |R_t| \Phi_t^T R_t + A_t = 2R_t \Phi_t^T R_t \Sigma_t \Phi_t^T R_t x_t
\]

(8)

where the gradients of the two terms are derived as the \( D \times N \) matrices given by

\[
\frac{\partial}{\partial \Phi_t} \log |R_t| \Phi_t^T R_t + A_t = 2R_t \Phi_t^T R_t \Sigma_t \Phi_t^T R_t x_t
\]

(9)

The new estimate \( \Phi_t \), which is a \( D \times N \) matrix, satisfies

\[
\sum_{t} \gamma_t(i) R_t |\Phi_t \Sigma_t - \Phi_t m_t \Sigma_t^{-1} m_t + x_t m_t^T| = 0
\]

(10)
Similar to the solution for $A_i$, the new estimate of $\Phi_i$ can be written in an implicit form, i.e.

$$\hat{\Phi}_i = \left[ \sum_i \gamma(i) x_i m_{i1}^T \right] \left[ \sum_i \gamma(i) (\Sigma_i + m_{i1} m_{i1}^T) \right]^{-1}$$

$$= \frac{\sum_i \gamma(i) x_i m_{i1}^T}{\sum_i \gamma(i)} \cdot A_i \doteq \hat{\Sigma}(\Phi_i)$$

(11)

Again, we reuse the notations $m_{i1}$ and $A_i$ to express the solution to $\Phi_i$.

To find the new estimate of the precision matrix of reconstruction errors $R_i$, we similarly maximize the auxiliary function (5) with respect to $R_i$ and obtain

$$\sum_i \gamma(i) \left[ R_i^{-1} - \Phi_i \Sigma_i \Phi_i^T - x_i x_i^T + \frac{\partial}{\partial R_i} (x_i^T R_i \Phi_i \Sigma_i \Phi_i^T R_i x_i) \right] = 0$$

(12)

where the gradient in the last term of the LHS is derived by

$$\frac{\partial}{\partial R_i} (x_i^T R_i \Phi_i \Sigma_i \Phi_i^T R_i x_i) = \frac{\partial}{\partial R_i} \text{Tr}\{\Sigma_i \Phi_i^T R_i x_i x_i^T R_i \Phi_i\}$$

$$= -\Phi_i \Sigma_i \Phi_i^T R_i x_i x_i^T R_i \Phi_i + \Phi_i \Sigma_i \Phi_i^T R_i x_i x_i^T R_i \Phi_i + \Phi_i \Sigma_i \Phi_i^T R_i x_i x_i^T R_i + x_i x_i^T$$

(13)

Equivalently, (12),(13) are arranged as

$$\hat{R}_i^{-1} = \Phi_i \Sigma_i \Phi_i^T + \frac{\sum_i \gamma(i) (x_i - \Phi_i m_{i1})(x_i - \Phi_i m_{i1})^T}{\sum_i \gamma(i)}$$

(14)

which is also a recursive solution to $R_i$ since the RHS of (14) depends on $R_i$. Importantly, we express the new estimate $\hat{R}_i$ by using the notations $m_{i1}$ and $\Sigma_i$ which have been defined and calculated for the solutions to $A_i$ and $\Phi_i$. Note that the RHS of (14) is symmetric positive definite which is required for the estimation of $R_i$.

Such consistent solutions to $A_i$, $\Phi_i$, and $R_i$, expressed as implicit equations, do not require a gradient descent implementation. The convergence of parameter estimation is empirically observed. Moreover, the computational cost of estimating the three parameters is significantly reduced due to the shared calculation of $m_{i1}$ and $\Sigma_i$.

To our knowledge, the multivariate solutions for BS-HMMs are novel and differ significantly from those of RVM [7]. Additionally, they are specifically designed for HMM-based sequential pattern recognition.

2.4. Extensions for Acoustic Modeling

The following two extensions were found beneficial for improving the performance of BS-HMMs for acoustic modeling. First, we consider a mixture model of basis vectors within each state. Feature vectors are assumed to be generated from several state-dependent subspaces constructed by different sets of basis vectors. Mixture component $j$ for state $i$ has precision matrix of sensing weights $A_{ij}$, basis $\Phi_{ij}$ and precision matrix of reconstruction errors $R_{ij}$.

Second, we allow the reconstruction error for mixture component $(i, j)$ to have a non-zero mean $\mu_{ij}$ that can be estimated from data. Accordingly, (7),(11),(14) are modified by replacing $x_i$ with $x_i - \mu_{ij}$. The ML estimate for $\mu_{ij}$ is the same as for GMM-based HMMs. The primary reason for introducing mixture component means is to be able to use MLLR for speaker adaptation.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We experimented with the mixture model of BS-HMMs on an English broadcast news transcription task. The training data consists of 50 hours of transcribed audio and we report results on the DEV04f test set which has 2 hours of speech and 22.6K words.

The input speech is encoded by VTLN-warped LP cepstral frames which are mean and variance normalized on a per speaker basis. Every 9 consecutive frames are spliced together and projected to 40 dimensions by an LDA transform followed by a decorrelating semi-tied covariance (STC) transform [9]. Words in the recognition lexicon are represented as sequences of phones, and phones are modeled with 3-state left-to-right HMMs without state skipping. All acoustic models have pentaphone cross-word acoustic context and are speaker Adaptively trained with feature-space MLLR (FMLLR) [9]. At test time, speaker adaptation is performed with VTLN, FMLLR and multiple regression tree-based MLLR transforms. The recognition vocabulary has 90K words and the decoding is done with a 4-gram language model containing 4M ngrams which was trained with modified Kneser-Ney smoothing.

The baseline GMM-HMMs and the proposed BS-HMMs have comparable context decision trees containing 2200 leaves. We carefully optimized the baseline models by varying the number of Gaussians per state and we report results for two configurations: 64 Gaussians per state and 128 Gaussians per state. For the BS-HMMs, the number of mixture components per state is empirically set to the number of Gaussians per state divided by 4 resulting in two configurations: 16 and 32 components per state, respectively.

We found that the performance of the BS-HMMs is quite sensitive to parameter initialization. The best results were obtained by partitioning the Gaussian means of the baseline GMM for a given state using k-means into clusters. The clusters become the bases $\{\Phi_{ij}\}$ for the BS-HMM mixture components. Note that the BS-HMMs have 25% fewer parameters than the corresponding GMM-HMMs because of fewer mixture component variances. The means $\mu_{ij}$ are initialized to zero. The precisions $A_{ij}$ are initialized to the identity matrix and are assumed to be diagonal. The training regime for the BS-HMMs consists of 5 iterations with fixed HMM state alignments followed by two Viterbi iterations where the data are re-aligned with the BS-HMMs.

3.2. Experimental Results

In Figure 1, we plot the average and standard deviation of the estimated precisions of the sensing weights as a function of the training iteration for the 16 components per state model. We find that the precisions tend to increase for the first few iterations until they stabilize. This implies that more basis vectors become less relevant to the observation data as the training proceeds.

In Table 1, we compare the performance of the baseline GMM-HMMs and the proposed BS-HMMs for maximum likelihood and
Table 1. Comparison of word error rates for ML and discriminatively trained (DT) GMM-HMMs and BS-HMMs with 64/16 and 128/32 mixture components per state, respectively.

<table>
<thead>
<tr>
<th>System</th>
<th>Training</th>
<th>64 mix/state</th>
<th>128 mix/state</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-HMM</td>
<td>ML</td>
<td>23.6%</td>
<td>23.6%</td>
</tr>
<tr>
<td>BS-HMM</td>
<td>ML</td>
<td>22.6%</td>
<td>22.2%</td>
</tr>
<tr>
<td>GMM-HMM</td>
<td>DT</td>
<td>18.7%</td>
<td>19.0%</td>
</tr>
<tr>
<td>BS-HMM</td>
<td>DT</td>
<td>18.2%</td>
<td>18.4%</td>
</tr>
<tr>
<td>Cross-adaptation</td>
<td>DT</td>
<td>17.6%</td>
<td>17.8%</td>
</tr>
</tbody>
</table>

Fig. 1. Evolution of the mean and standard deviation of the precisions of the sensing weights as a function of the training iteration.

Table 1. Evolution of the mean and standard deviation of the precisions of the sensing weights as a function of the training iteration.

4. CONCLUSION

We described a new modeling paradigm called Bayesian sensing hidden Markov model for speech recognition. Compared to standard HMMs with GMM state emissions, we address the problem of parameter overfitting and model regularization in an HMM framework by drawing ideas from relevance vector machines such as the automatic determination of relevant basis vectors. Three sets of state-dependent parameters are jointly estimated: basis vectors, precision matrices for the sensing weights and for the reconstruction errors. Multivariate solutions are obtained in a consistent manner and are expressed through recursive formulas. These formulas are interpreted in terms of mean vectors and covariance matrices of posterior distributions of the sensing weights. We have tested the proposed models on a large vocabulary continuous speech recognition task and shown gains over both ML and discriminatively trained state-of-the-art GMM-HMMs. Additional improvements were obtained by cross-adapting the baseline GMM-HMMs on the BS-HMM output. Future work will address the application of these models to larger datasets and other sequential pattern recognition problems.

5. ACKNOWLEDGMENT

The authors wish to thank Vaibhava Goel, John Hershey, David Nahamoo, Michael Picheny and Hagen Soltau from the IBM Speech Recognition group for useful discussions.

This work was partially supported by DARPA contract No. HR0011-06-2-0001. The views, opinions, and findings contained in this article are those of the authors, and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

6. REFERENCES


