# Some Properties of Bayesian Sensing Hidden Markov Models

George Saon[†] and Jen-Tzung Chien[‡]

[†]*IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598*
[‡]*National Cheng Kung University, Tainan, Taiwan 70101, ROC*

***Abstract*—In Bayesian sensing hidden Markov models (BSHMMs) the acoustic feature vectors are represented by a set of state-dependent basis vectors and by time-dependent sensing weights. The Bayesian formulation comes from assuming state-dependent zero mean Gaussian priors for the weights and from using marginal likelihood functions obtained by integrating out the weights. Here, we discuss two properties of BSHMMs. The first property is that the marginal likelihood is Gaussian with a factor analyzed covariance matrix with the basis providing a low-rank correction to the diagonal covariance of the reconstruction errors. The second property, termed automatic relevance determination, provides a method for discarding basis vectors that are not relevant for encoding feature vectors. This allows model complexity control where one can initially train a large model and then prune it to a smaller size by removing the basis vectors which correspond to the largest precision values of the sensing weights. The last property turned out to be useful in successfully deploying models trained on 1800 hours of data during the 2011 DARPA GALE Arabic broadcast news transcription evaluation.**

## I. Introduction

One of the goals of acoustic modeling for modern ASR systems is to achieve a rich representation with few trainable parameters. The underlying principle is that, among several representations of similar descriptive power, one should choose the one with the fewest free parameters requiring the least amount of training data to be robustly estimated. One avenue that is considered is to use a shared representation across phonetic states and state-specific "views" derived from the shared parameters. This is the case, for example, for tied Gaussian mixture models (GMMs), subspace precision and mean (SPAM) models [1] and, more recently, subspace GMMs [2]. In another approach, not necessarily orthogonal to the first, acoustic models are designed as an efficient approximation to more parameter-intensive, accurate representations such as full-covariance GMMs. Diagonal covariance GMMs, semi-tied covariance transforms [3] and factor-analyzed covariance GMMs [4] fall into this category.

Another option that is getting some traction in the literature is to rely on Bayesian learning techniques to address the overfitting issue when training on insufficient or noisy data. The uncertainties of HMM parameters, expressed through prior distributions, are incorporated in the model construction leading to a regularized model with superior recognition performance on mismatched test data [5]. The advantage of Bayesian learning in this context is that it can provide "error bars" or "distribution estimates" of the underlying parameters rather than providing "point estimates" which can be unreliable.

Bayesian sensing HMMs, introduced in [6], [7], combine both aforementioned ideas. On the one hand, as will be shown in this paper, they provide a parsimonious approximation to full-covariance Gaussian modeling through factor analyzed covariance matrices. The factors correspond to basis vectors that are used in a basis representation of speech features. The compressed representation comes from the fact that basis vectors which are not relevant for encoding feature vectors can be discarded through a procedure called automatic relevance determination (ARD) [8]. On the other hand, the model adopts a Bayesian treatment for the sensing weights which attempts to mitigate the risk of an overdetermined basis representation due to maximum likelihood (ML) point estimates of the weights. Instead, the model provides posterior distribution estimates of the sensing weights when conditioned on a feature vector and an HMM state. This Bayesian sensing scheme is important for compensating the model variations and for having a regularized basis representation.

The paper is organized as follows: in section II we review the model specification for Bayesian sensing HMMs, in section III we discuss two properties of these models, in section IV we provide some experimental results on a large scale LVCSR task, and in section V we summarize our findings.

## II. Bayesian Sensing Hidden Markov Models

In Bayesian sensing HMMs, acoustic feature vectors $\mathbf{x}_t \in \mathbb{R}^D$ are seen as samples generated from a state-dependent additive model

$$\mathbf{x}_t = \Phi_i \mathbf{w}_t + \boldsymbol{\epsilon}_t \tag{1}$$

where $\Phi_i = [\boldsymbol{\phi}_{i1}, \ldots, \boldsymbol{\phi}_{iN}]$, $\boldsymbol{\phi}_{ij} \in \mathbb{R}^D$, is the basis (or dictionary) for state $i$, $\mathbf{w}_t = [w_{t1}, \ldots, w_{tN}]^T$ is a time-dependent weight vector, and $\epsilon_t$ is a sample drawn from an independent Gaussian noise process with zero mean and a state-dependent precision matrix $R_i$, i.e. $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, R_i^{-1})$ or equivalently

$$p(\mathbf{x}_t|\mathbf{w}_t, \lambda_i) \propto |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \Phi_i\mathbf{w}_t)^T R_i (\mathbf{x}_t - \Phi_i\mathbf{w}_t)\right] \tag{2}$$

The sensing weight vector $\mathbf{w}_t$ is also assumed to be Gaussian distributed with zero mean and a state-dependent precision matrix $A_i$, that is

$$p(\mathbf{w}_t|\lambda_i) \propto |A_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{w}_t^T A_i \mathbf{w}_t\right] \tag{3}$$

where $\lambda_i = \{A_i, \Phi_i, R_i\}$ denotes the parameters for state $i$. Figure 1 displays the graphical model of BSHMMs for sequential frames $\{\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}\}$. Each frame $\mathbf{x}_t$ is generated by the BSHMM 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 $\{R_i\}$.

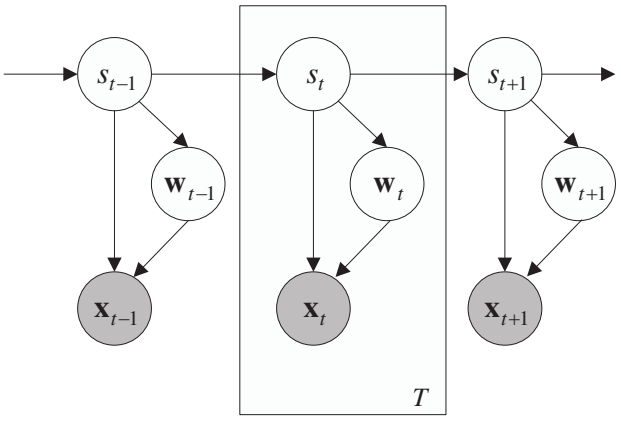


Fig. 1. Graphical model of BSHMMs. $s_{t-1}, s_t, s_{t+1}$, $\mathbf{w}_{t-1}, \mathbf{w}_t, \mathbf{w}_{t+1}$ and $\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}$ denote the HMM states, the sensing weights and the feature vectors at times $t-1, t, t+1$, respectively.

The key quantity for this model is the marginal likelihood or Bayesian predictive likelihood $p(\mathbf{x}_t|\lambda_i)$ which is obtained by marginalizing over the weight vector $\mathbf{w}_t$ and is proportional to [6]

$$\begin{aligned}
&\int_{\mathbb{R}^N} |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \Phi_i\mathbf{w}_t)^T R_i (\mathbf{x}_t - \Phi_i\mathbf{w}_t)\right] \\
&\cdot |A_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{w}_t^T A_i \mathbf{w}_t\right] d\mathbf{w}_t \\
&\propto |R_i|^{1/2}|A_i|^{1/2}|\Sigma_i|^{1/2} \exp\left[-\frac{1}{2}\mathbf{x}_t^T (R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i)\mathbf{x_t}\right] \\
&= |R_i|^{1/2}|A_i|^{1/2}|\Sigma_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t^T R_i \mathbf{x}_t - \mathbf{m}_{ti}^T \Sigma_i^{-1} \mathbf{m}_{ti})\right]
\end{aligned} \tag{4}$$

where $\Sigma_i \triangleq (\Phi_i^T R_i \Phi_i + A_i)^{-1}$, $\mathbf{m}_{ti} \triangleq \Sigma_i \Phi_i^T R_i \mathbf{x}_t$ are the *covariance matrix* and the *mean vector* of the posterior distribution $p(\mathbf{w}_t|\mathbf{x}_t, \lambda_i)$, respectively. The mean vector $\mathbf{m}_{ti}$ corresponds to the maximum *a posteriori* (MAP) estimate of the sensing weights at time $t$ for state $i$. In [6], we discuss the estimation of BSHMM parameters according to an ML type II criterion by maximizing the marginal likelihood of the training data, whereas in [7] we derive parameter updates under a maximum mutual information objective function.

## III. Properties of BSHMMs

In this section we discuss two properties of Bayesian sensing HMMs that were not addressed in [6], [7]. These properties shed some additional light on the functioning of the models and were found to be beneficial experimentally.

### A. Gaussians with factor analyzed covariances

The first property has to do with the fact that the marginal likelihood derived in (4) is a Gaussian function because the convolution of two Gaussian distributions is also a Gaussian distribution. This has two immediate consequences: 1) the matrix $R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i$ has to be symmetric positive definite (SPD) to be a valid precision matrix and 2) its determinant has to be equal to $|R_i||A_i||\Sigma_i|$. In order to gain more insight into the model, we provide alternative proofs for 1) and 2). To prove 1) we make use of the Woodbury matrix inversion lemma which states that

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{5}$$

where $A$, $U$, $C$ and $V$ denote matrices of compatible dimensions. Specifically, $A$ is $n \times n$, $U$ is $n \times k$, $V$ is $k \times n$ and $C$ is $k \times k$. Setting $A = R_i$, $U = R_i\Phi_i$, $V = \Phi_i^T R_i$ and $C = -\Sigma_i$, we get

$$\begin{aligned}
&S_i \triangleq (R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i)^{-1} \\
&= R_i^{-1} - R_i^{-1}R_i\Phi_i((-\Sigma_i)^{-1} + \Phi_i^T R_i R_i^{-1} R_i \Phi_i)^{-1}\Phi_i^T R_i R_i^{-1} \\
&= R_i^{-1} + \Phi_i A_i^{-1} \Phi_i^T
\end{aligned} \tag{6}$$

This result also follows from (1) due to the additivity of the covariance matrices of independent multivariate random variables, i.e.

$$S_i = \mathrm{Cov}(\mathbf{x}_t) = \mathrm{Cov}(\Phi_i\mathbf{w}_t) + \mathrm{Cov}(\boldsymbol{\epsilon}_t) = \Phi_i A_i^{-1}\Phi_i^T + R_i^{-1} \tag{7}$$

because $\boldsymbol{\epsilon}_t$ is independent of $\mathbf{w}_t$. The expression in (6) and (7) is a sum of two SPD matrices which makes the covariance matrix $S_i$ also SPD.

For $R_i$ diagonal, $S_i$ is a factor analyzed covariance where the factor loading matrix $\Lambda_i \triangleq \Phi_i A_i^{-1/2}$ can be seen as a rank-$N$ correction to $R_i^{-1}$ as shown in (7). $\Lambda_i$ is responsible for modeling the off-diagonal elements in $S_i$. The hope is that with relatively few factors, a good approximation of the covariance of the underlying Gaussian distribution can be obtained. Such a covariance structure has been proposed for general ASR in [4] and [9] and for noise-robust ASR in [10]. The difference here is that we have an additional term, $A_i$, which controls the importance of the columns of $\Phi_i$ as will be explained in the next section. Secondly, for (4) to be a Gaussian distribution, the following determinant equality has to hold

$$|R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i| = |R_i||A_i||\Sigma_i| \tag{8}$$

One way to prove this is by observing that $p(\mathbf{x}_t|\lambda_i)$ is a valid PDF therefore it has to integrate to 1. Another way of showing (8) involves the extended matrix of size $(D+N)\times(D+N)$

$$\begin{bmatrix} (R_i)_{D\times D} & (R_i\Phi_i)_{D\times N} \\ (\Phi_i^T R_i)_{N\times D} & \Sigma_i^{-1} = (\Phi_i^T R_i \Phi_i + A_i)_{N\times N} \end{bmatrix} \tag{9}$$

For the sake of clarity, we indicate the dimensions of the various submatrices. We consider the determinant identity for a partitioned matrix

$$\begin{vmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{vmatrix} = |B_{22}||B_{11} - B_{12}B_{22}^{-1}B_{21}| = |B_{11}||B_{22} - B_{21}B_{11}^{-1}B_{12}| \tag{10}$$

where, for example, the second equality can be demonstrated via the matrix identity

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} B_{11} & 0 \\ B_{21} & I \end{bmatrix} \begin{bmatrix} I & B_{11}^{-1}B_{12} \\ 0 & B_{22} - B_{21}B_{11}^{-1}B_{12} \end{bmatrix} \tag{11}$$

Applying (10) to the matrix in (9) yields

$$\begin{vmatrix} R_i & R_i\Phi_i \\ \Phi_i^T R_i & \Phi_i^T R_i \Phi_i + A_i \end{vmatrix} = \frac{|R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i|}{|\Sigma_i|} = |R_i||A_i| \tag{12}$$

which completes the proof of (8).

### B. Automatic relevance determination

In BSHMMs, the state-dependent priors of the time-dependent sensing weights are governed by a set of hyperparameters consisting of the precision matrices of the sensing weights. The precision matrices are iteratively reestimated from data using evidence maximization (ML type II estimation). More concretely, we remind the reader of the update formula for $A_i$ which was derived in [6]

$$A_i^{(k+1)} = \left[\Sigma_i + \frac{\sum_t \gamma_t(i)\mathbf{m}_{ti}\mathbf{m}_{ti}^T}{\sum_t \gamma_t(i)}\right]^{-1} \tag{13}$$

(13) was obtained by maximizing with respect to $A_i$ the auxiliary function

$$Q(\lambda|\lambda^{(k)}) = \sum_i \sum_t \gamma_t(i) \log p(\mathbf{x}_t|\lambda_i) \tag{14}$$

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 $X = \{\mathbf{x}_t\}$ and the current parameters $\lambda^{(k)}$. Recall that $\mathbf{m}_{ti}$ are the MAP estimates of the sensing weights at time $t$ (for state $i$) and $\Sigma_i$ is the covariance matrix of the posterior distribution $p(\mathbf{w}_t|\mathbf{x}_t, \lambda_i)$.

For the case of a diagonal matrix of sensing weights $A_i = \text{diag}(\alpha_{i1}, \ldots, \alpha_{iN})$, if many of the $\mathbf{m}_{ti}$'s are close to 0 for a particular dimension $j$, $\alpha_{ij}$ can become quite large. In the case of sufficiently large $\alpha_{ij}$, the corresponding weight parameter $w_{tj}$ is pegged at 0 by the dimension-specific prior $\mathcal{N}(0, \alpha_{ij}^{-1})$ implying an irrelevant basis $\boldsymbol{\phi}_{ij}$ for the Bayesian representation. This characteristic of the estimated $\alpha_{ij}$ is known as automatic relevance determination (ARD) [8]. The effect of a large $\alpha_{ij}$ on the factor analyzed covariance $S_i = (R_i - R_i\Phi_i\Sigma_i\Phi_i^T R_i)^{-1}$ can also be seen from (6) when rewritten as an expansion of rank-1 matrices

$$S_i = R_i^{-1} + \sum_{j=1}^{N} \frac{1}{\alpha_{ij}} \boldsymbol{\phi}_{ij}\boldsymbol{\phi}_{ij}^T \tag{15}$$

A large $\alpha_{ij}$ will effectively zero out the contribution of $\boldsymbol{\phi}_{ij}\boldsymbol{\phi}_{ij}^T$ in the expansion. Alternatively, one can use the trained $\alpha_{ij}$ for controlling model complexity. The technique consists in discarding the basis vectors $\boldsymbol{\phi}_{ij}$ for which the corresponding $\alpha_{ij}$ are greater than a predefined threshold and in retraining the compressed model. The compressed representation is more robust to mismatched or noisy conditions and provides better generalization on unseen data.

## IV. Experiments and Results

### A. Acoustic training and test data

The experiments were carried out on an Arabic broadcast news transcription task as part of the DARPA GALE program. The goal of the GALE program is to make foreign language (Arabic and Chinese) speech and text accessible to English-only speakers, particularly in military settings. A core component of GALE is automatic speech recognition, and research in this area spans multiple fields ranging from traditional speech recognition to improving the interfaces between speech recognition, machine translation and information extraction.

We used 1800 hours of manually transcribed Arabic broadcast news and broadcast conversations, coming from many different broadcasters and containing a mixture of Modern Standard Arabic (MSA) and dialectical Arabic. Results are presented on several testsets depending on the experiment and the availability of the data at the time the experiment was run: DEV'07 (2.5 hours), DEV'08 (3 hours), DEV'09 (3 hours), unsequestered portion of the phase 4 evaluation EVAL'09 (4.2 hours), unsequestered portion of the phase 5 evaluation EVAL'11 (3 hours).

### B. Front-end processing and acoustic modeling

The input speech is represented by PLP VTL-warped cepstra and a context window of 9 frames. The features are mean and variance normalized on a per speaker basis. An LDA transform is used to reduce the feature dimensionality to 40. The maximum-likelihood training of the acoustic model is

interleaved with estimation of a global semi-tied covariance transform [3].

Words in the recognition lexicon have a phonetic representation, and phones are modeled with 3-state left-to-right HMMs without state skipping. For the experiments described here, we used unvowelized (or graphemic) acoustic models where Arabic short vowels are not modeled explicitly. All acoustic models have pentaphone cross-word acoustic context and are speaker adaptively trained with feature-space MLLR (FMLLR) [3]. At test time, speaker adaptation is performed with VTLN, FMLLR and multiple regression tree-based MLLR transforms. The recognition vocabulary has 795K words and the decoding is done with 4-gram language models containing either 78M or 883M n-grams depending on the setup. The LMs were estimated with modified Kneser-Ney smoothing on a variety of corpora.

The acoustic models are discriminatively trained in both feature space [11] and model space using the boosted MMI criterion [12], [13]. For feature space discriminative training, all models use two-level transforms with offset features as described in [14]. Frame posteriors and offset features are provided by 4096 diagonal covariance Gaussians and form the input to the first-level transform. The temporal context spanned by the second-level transform is $\pm 8$ frames.

The baseline acoustic models have 5000 context-dependent HMM states and 800K 40-dimensional diagonal covariance Gaussians. The models have been sized for optimal performance after discriminative training during the GALE phase 4 and phase 5 evaluations [15], [16].

### C. Two extensions for acoustic modeling

Throughout the paper we assumed BSHMMs with a single Gaussian per state. In practice, similar to [6], [7], we consider a mixture model within each state where each component has its own basis and precision matrices of sensing weights and reconstruction errors. This leads to a new GMM in each state where each Gaussian has its own factor analyzed covariance matrix.

Second, we allow the Gaussians in the model to have non-zero means that are also trained. These means were introduced primarily in order to apply MLLR for speaker adaptation. The factor-analyzed covariances are assumed to be diagonal when computing the sufficient statistics for the MLLR transforms. In Table I, we compare BSHMMs with zero and non-zero mean Gaussians and conclude that the main benefit comes from being able to perform MLLR speaker adaptation on the non-zero means.

| Means | MLLR | DEV'07 | DEV'08 | DEV'09 |
|---|---|---|---|---|
| zero | no | 14.3% | 16.7% | 19.7% |
| non-zero | no | 14.2% | 16.4% | 19.6% |
| non-zero | yes | 13.6% | 16.0% | 18.9% |

TABLE I
COMPARISON OF WORD ERROR RATES FOR BSHMMS WITH ZERO AND NON-ZERO MEAN GAUSSIANS.

### D. BSHMM initialization and training

The first step in initializing the BSHMM parameters is to train a large acoustic model with 2.8M diagonal covariance Gaussians (and 5000 HMM states) using maximum likelihood. The means of the GMM for state $i$ are clustered using k-means. The means that are assigned to cluster center $(i, j)$ form the initial basis $\Phi_{ij}$ for mixture component $(i, j)$. The resulting number of mixture components for the BSHMM after the clustering step was 417K with, on average, 6.7 vectors per basis. The mixture component means $\boldsymbol{\mu}_{ij}$ are initialized to zero. The precision matrices $A_{ij}$ and $R_{ij}$ are assumed to be diagonal and are initialized to the identity matrix.

The ML type II training regime for the BSHMMs consists of 5 iterations with fixed HMM state alignments provided by the baseline HMMs followed by one Viterbi iteration where the data are re-aligned with the BSHMMs. In Table II, we compare the performance of the baseline 800K Gaussians models and the 2.8M Gaussians models used to seed the BSHMMs after ML training, and the BSHMMs after ML type II training.

| System | Nb. parameters | WER | | |
|---|---|---|---|---|
| | | DEV'07 | DEV'08 | DEV'09 |
| baseline 800K | 64.8M | 13.8% | 16.4% | 19.6% |
| baseline 2.8M | 226.8M | 14.1% | 16.2% | 19.3% |
| BSHMM 417K | 148.5M | 13.6% | 16.0% | 18.9% |

TABLE II
COMPARISON OF THE NUMBER OF FREE PARAMETERS AND WORD ERROR RATES FOR BASELINE ACOUSTIC MODELS AFTER ML TRAINING AND BSHMMS AFTER ML TYPE II TRAINING.

Additionally, we compare the number of free parameters for the three models in the second column of Table II. As can be seen, BSHMMs outperform both acoustic models even though it does not have the most parameters. We attribute this possibly to the more accurate modeling of the Gaussian covariance matrices and/or to the Bayesian parameter updates which provide an efficient form of smoothing. More experiments are needed to tease apart the relative contributions of the two components.

### E. Model compression

In this set of experiments, the acoustic models are built in the FMMI space, i.e. with discriminative feature-space transforms. As before, we first train a large 2.8M diagonal covariance Gaussian HMM with ML that is used to initialize the Bayesian sensing HMM parameters. Next, we perform ML type II estimation of the BSHMM parameters for 5 iterations with fixed state alignments followed by one Viterbi iteration. In Figure 2, we plot the histogram of the hyperparameters of the sensing weights (the $\alpha_{ij}$'s) after 6 training iterations.

We observe that all $\alpha_{ij}$'s have relatively small values because all of the basis vectors get used during training. This is due to the initialization of the basis vectors with Gaussian means which are estimated from the same training data. Model compression is performed by discarding 50% of the basis vectors $\boldsymbol{\phi}_{ij}$ corresponding to the largest $\alpha_{ij}$. This resulted in

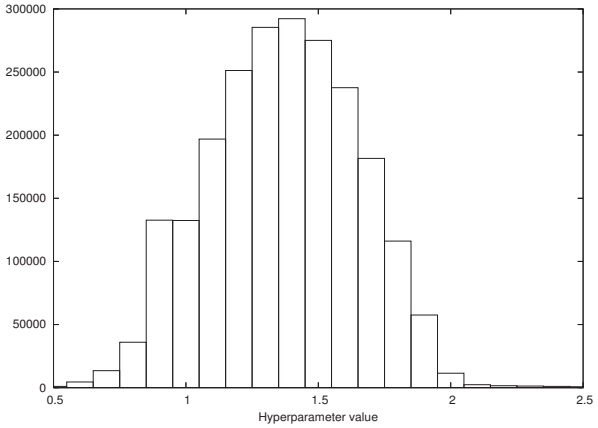


Fig. 2. Histogram of precision values $\alpha_{ij}$ of the sensing weights after 6 iterations of ML type II training.

a compressed model with approximately 91M free parameters (about 30% larger than the 800K baseline HMM). For both models (original and compressed), we perform four iterations of model-space discriminative training using the boosted MMI criterion as described in [7]. In Table III, we report the recognition results before and after model-space discriminative training.

| Model | Training | DEV'07 | DEV'08 | DEV'09 |
|---|---|---|---|---|
| original | ML type II | 12.0% | 13.9% | 17.4% |
| compressed | ML type II | 12.4% | 14.2% | 17.6% |
| original | boosted MMI | 10.7% | 11.9% | 15.0% |
| compressed | boosted MMI | 10.4% | 11.7% | 14.8% |

TABLE III
COMPARISON OF WORD ERROR RATES FOR ORIGINAL AND COMPRESSED BSHMMS BEFORE AND AFTER MODEL-SPACE DISCRIMINATIVE TRAINING WITH BOOSTED MMI. ALL MODELS HAVE DISCRIMINATIVE FEATURE-SPACE TRANSFORMS.

We conclude that the compressed models outperform the original ones after discriminative training even though they start from a higher word error rate after ML type II estimation. We have observed this phenomenon before in the context of standard HMMs with diagonal covariance Gaussians where optimal model size after ML training is not necessarily optimal after discriminative training. Unfortunately, discriminative training is significantly more expensive than ML estimation which makes it difficult to find the optimal model size.

### *F. GALE 2011 evaluation deployment*

In Table IV, we compare the word error rates of the baseline acoustic models and the compressed BSHMMs after discriminative training and cross-adaptation on the output of a system using vowelized acoustic models and subspace GMMs (SGMMs) [2]. Additionally, we report the performance of system combination using decision tree arrays as described in [17]. The DEV'09 word error rates in this table are much lower than the ones in Table III because of the large cross-adaptation gain. Also, the decodings here use the larger LM with 883M n-grams whereas in Tables I, II and III the smaller 78M n-grams LM was employed.

| System | DEV'09 | EVAL'09 | EVAL'11 |
|---|---|---|---|
| baseline 800K | 13.1% | 10.0% | 9.4% |
| BSHMM | 12.8% | 9.7% | 9.1% |
| system combination | 12.6% | 9.6% | 9.0% |

TABLE IV
COMPARISON OF WORD ERROR RATES FOR BASELINE ACOUSTIC MODELS AND BSHMMS AFTER DISCRIMINATIVE TRAINING AND CROSS-ADAPTATION ON THE OUTPUT OF A SYSTEM USING SUBSPACE GMMS.

The numbers reported in Table IV were obtained during the actual GALE 2011 evaluation run. The EVAL'11 testset corresponds to the previously unseen evaluation data for which there was no reference transcript available during the evaluation. The other testsets were part of the "shadow data" on which accuracy could be measured in order to optimize the overall system architecture. The system combination reported in the last line of Table IV formed one decoding branch out of three in the overall evaluation system diagram [16]. As can be seen, the BSHMMs achieved a 3% relative accuracy improvement over the baseline HMMs on the evaluation data (4% relative with system combination).

## V. CONCLUSION

In Bayesian sensing HMMs, the observations within each state are modeled by a mixture of Gaussians with factor analyzed covariance matrices. This allows a parsimonious representation of the acoustic features where an efficient approximation to the full covariance of the underlying Gaussian distributions can be achieved with relatively few factors or basis vectors. Furthermore, the complexity of the model in terms of the number of basis vectors can be controlled by discarding the least relevant factors corresponding to the largest sensing weight precision matrix values after training. This procedure leads to more compact models with superior performance after discriminative training. Such models have been deployed during the 2011 GALE Arabic broadcast news transcription evaluation and have shown gains on the evaluation data over state-of-the-art systems.

Future work will address a better initialization of the basis vectors via eigenanalysis of the covariances of a full covariance model. Additionally, we will look at various ways to improve discriminative training of the model parameters through better smoothing and by also updating the precision matrices of the sensing weights.

## ACKNOWLEDGMENT

This work was supported in part by DARPA under Grant HR0011-06-2-0001[1].

[1]Approved for Public Release, Distribution Unlimited. The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter 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.

## REFERENCES


[1] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. of ICSLP*, 2002, pp. 2177–2180.

[2] D. Povey *et al.*, "Subspace Gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010, pp. 4330–4333.

[3] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[4] R. Gopinath, B. Ramabhadran, and S. Dharanipragada, "Factor analysis invariant to linear transformations of data," in *Proc. ICSLP*, 1998, pp. 2223–2226.

[5] J.-T. Chien and S. Furui, "Predictive hidden Markov model selection for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 377–387, 2005.

[6] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models for speech recognition," in *Proc. of ICASSP*, 2011, pp. 5056–5059.

[7] ——, "Discriminative training for Bayesian sensing hidden Markov models," in *Proc. of ICASSP*, 2011, pp. 5316–5319.

[8] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[9] A.-V. Rosti and M. Gales, "Factor analysed hidden Markov models," in *Proc. ICASSP*, 2002, pp. 949–952.

[10] J.-T. Chien and C.-W. Ting, "Factor analyzed subspace modeling and selection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 239–248, 2008.

[11] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. of ICASSP*, 2005, pp. 961–964.

[12] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of ICASSP*, 2008, pp. 4057–4060.

[13] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Proc. of INTERSPEECH*, 2008, pp. 920–923.

[14] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proc. of Interspeech*, 2005, pp. 2977–2980.

[15] B. Kingsbury, H. Soltau, G. Saon, S. Chu, H.-K. Kuo, L. Mangu, S. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 GALE Arabic speech transcription system," in *Proc. of ICASSP*, 2011, pp. 4672–4675.

[16] L. Mangu, H.-K. Kuo, S. Chu, B. Kingsbury, G. Saon, H. Soltau, and F. Biadsy, "The IBM 2011 GALE Arabic speech transcription system," in *Proc. of IEEE ASRU*, 2011, submitted.

[17] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Proc. of IEEE Workshop on Spoken Language Technology (SLT)*, 2010, pp. 97–102.