# BAYESIAN ADAPTATION IN HMM TRAINING AND DECODING USING A MIXTURE OF FEATURE TRANSFORMS

*Stavros Tsakalidis, Spyros Matsoukas*

BBN Technologies
10 Moulton St., Cambridge, MA 02138
{stavros,smatsouk}@bbn.com

**ABSTRACT**

Adaptive training under a Bayesian framework addresses some limitations of the standard Maximum Likelihood approaches. Also, the adaptively trained system can be directly used in unsupervised inference. The Bayesian framework uses a distribution of the transform rather than a point estimate. A continuous transform distribution makes the integral associated with the Bayesian framework intractable and therefore various approximations have been proposed. In this paper we model the transform distribution via a mixture of transforms. Under this model, the likelihood of an utterance is computed as a weighted sum of the likelihoods obtained by transforming its features based on each of the transforms in the mixture, with weights set to the transform priors. Experimental results on Arabic broadcast news exhibit increased likelihood on acoustic training data and improved speech recognition performance on unseen test data, compared to speaker independent and standard adaptive models.

***Index Terms***— adaptive training, Bayesian inference

## 1. INTRODUCTION

The task of acoustic modeling is to provide a stochastic model that captures the phonetically relevant variation of the speech signal. One of the prominent problems in modeling the process of speech communication is that of *unwanted* variability, due to the wide range of speakers and acoustic conditions in the speech signal. A well-established technique aiming at reducing the unwanted variability within the training corpus is adaptive training.

Adaptive training tackles variability by hypothesizing two models: a model of phonetically relevant variation and a model of speaker or acoustic variations, where each speaker or acoustic variation is modeled by a separate transform. Regardless of the precise formulation of the transform, these techniques, until recently, have been based on maximum likelihood (ML) estimation, such as Speaker Adaptive Training (SAT) [1], or discriminative criteria, such as Maximum Mutual Information SAT (DSAT) [2] and Minimum Phone Error SAT (MPE-SAT) [3]. Although these adaptive training techniques alleviate the problem of unwanted variability, their application in speech recognition systems exhibits some limitations.

One issue is that during decoding, the transforms used in these adaptive training models are being discarded and a new set of transforms is being estimated with respect to the speakers in the test set. Consequently, the transforms are not fully integrated in the training and decoding procedure. Moreover, it has been shown that discriminative versions of supervised adaptive training and supervised adaptation perform better than those found by ML training [2, 3, 4]. However, discriminative training for unsupervised adaptation is not as effective as ML training due to the errors in the training hypothesis [4]. Therefore, since the training-set transforms are not integrated into the decoding procedure, discriminatively-trained transforms cannot benefit the decoding process.

Recently, adaptive training procedures within a Bayesian framework have been proposed [5] which provide an integrated training and decoding procedure. Under this framework, the parameters of the transform are assumed to be random variables and therefore are described by their probability density function. The marginal likelihood of each hypothesis is calculated by integrating out over the transform distribution. Thus, the Bayesian approach provides a framework that allows the transform model, estimated in training, to be used directly in decoding. The integration of the transform model in decoding enables the use of discriminatively-trained transforms into the decoding process. Furthermore, since there is no need to perform adaptation to the test set, issues with limited adaptation data and over-tuning to the hypothesis can be avoided. However, in the general case, the integral associated with the transform distribution is intractable and therefore approximations are required [5].

In this work, we propose a special case of the Bayesian-based adaptive procedure [5] that does not involve the use of approximations in the calculation of the integral. The intractable integral is avoided by modeling the transform distribution by a finite mixture of point transform estimates. The transforms employed in this model are feature-based, that is, they act on the observation features. Under this model, the likelihood of an utterance is computed as a weighted sum of the likelihoods obtained by transforming its features based on each of the transforms in the mixture, with weights set to the transform priors. The proposed adaptive training procedure is termed Bayesian Speaker Adaptive Training (BSAT). BSAT applies an incremental training of the transforms in a way similar to the Gaussian mixture training. Contrary to SAT, BSAT does not rely on fixed speaker-clusters but rather allows the transforms to act on the training and test data dynamically. We will describe in subsequent sections the BSAT framework and derive the estimation procedure under the ML criterion. Also, we will provide alternative transform splitting techniques that gradually increase the number of transforms in the mixture. Finally, we will present experimental results in Arabic broadcast news that assess the effectiveness of BSAT.

## 2. BSAT FRAMEWORK

The goal of speech recognition is to find the word sequence $w^*$ that has the highest posterior probability, given the sequence of observa-

This material is based upon work supported by the GALE program of the Defense Advanced Research Projects Agency (IPTO) under Contract No. HR0011-06-C-0022.

tions $X = \{x_1, \ldots, x_N\}$

$$w^* = \underset{w}{\operatorname{argmax}}\, P(w \mid X) \tag{1}$$

The posterior probability $P(w \mid X)$ depends on the choice of parameterization for the acoustic and language model. The language model parameters are the n-gram probabilities. The acoustic model parameters consist of the hidden Markov model (HMM) state transition probabilities, state distributions (mixture weights, Gaussian means and variances), and speaker transforms. Let $\mathcal{L}$,$\mathcal{M}$, and $\mathcal{T}$ denote the sets of language model, HMM, and speaker transform parameters, respectively. Then, Equation 1 can be written as follows

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmax}} \int_{\mathcal{T},\mathcal{M},\mathcal{L}} \frac{P(w, X, \mathcal{T}, \mathcal{M}, \mathcal{L})}{P(X)}\, d\mathcal{T}\, d\mathcal{M}\, d\mathcal{L} \qquad (2) \\ &= \underset{w}{\operatorname{argmax}} \int_{\mathcal{T},\mathcal{M},\mathcal{L}} P(w, X \mid \mathcal{T}, \mathcal{M}, \mathcal{L}) P(\mathcal{T}, \mathcal{M}, \mathcal{L})\, d\mathcal{T}\, d\mathcal{M}\, d\mathcal{L} \end{aligned}$$

For a known sequence of observations, the marginal distribution $P(X)$ is constant. Therefore it was ignored since it does not affect the criterion of Equation 2. Assuming that the model parameters are independent and that there is only a single HMM and single LM, i.e., $P(\mathcal{M}) = \delta(\mathcal{M} - \lambda_H)$ and $P(\mathcal{L}) = \delta(\mathcal{L} - \lambda_L)$, where $\delta(\cdot)$ denotes the Dirac delta function, we get

$$w^* = \underset{w}{\operatorname{argmax}} \int_{\mathcal{T}} P(w, X \mid \mathcal{T}, \lambda_H, \lambda_L) P(\mathcal{T})\, d\mathcal{T} \tag{3}$$

In the general case, the integral over the transform distribution in Equation 3 is intractable and therefore approximations are required [5]. Here we assume that $P(\mathcal{T})$ has the discrete form

$$P(\mathcal{T}) = \sum_{k=1}^{M} v_k \delta(\mathcal{T} - \tau_k) \tag{4}$$

where $v_k$ are the mixture coefficients that satisfy $\sum_{k=1}^{M} v_k = 1$. Hence, the transform distribution in BSAT is a mixture of Dirac delta functions with $\tau_k$ as the mode. The mixture coefficients $v_k$ can be thought of as prior probabilities of the transform component $\tau_k$. The integral in Equation 3 becomes tractable and reduces to

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmax}} \sum_{k=1}^{M} v_k P(w, X \mid \tau_k, \lambda_H, \lambda_L) \\ &= \underset{w}{\operatorname{argmax}}\, P(w \mid \lambda_L) \sum_{k=1}^{M} v_k P(X \mid w, \tau_k, \lambda_H) \qquad (5) \end{aligned}$$

Note that the transform component remains constant over the whole observation sequence. In other words, the transform does not vary from one time instance to another. The values $P(w \mid \lambda_L)$ and $\sum_{k=1}^{M} v_k P(X \mid w, \tau_k, \lambda_H)$ are provided from the language model and acoustic model respectively. In the following section we will show how the ML criterion can be used to estimate the parameters of the acoustic model.

## 3. MAXIMUM LIKELIHOOD ESTIMATION OF BSAT PARAMETERS

Maximum likelihood estimation of the acoustic model parameters is performed using the expectation-maximization (EM) [6] algorithm where the function to maximize, based on Equation 5, is

$$L(X, w) = \sum_{k=1}^{M} v_k P(X \mid w, \tau_k, \lambda_H) \tag{6}$$

The entire parameter set for the BSAT model is defined as $\theta = \{\lambda_H, v_1, \ldots, v_M, \tau_1, \ldots, \tau_M\}$. Let $S = \{s_1, \ldots, s_N\}$ denote the sequence of unobserved HMM state sequences and the random scalar variable $k$ denote the unobserved component of the transform mixture. The auxiliary function of interest is

$$\begin{aligned} Q(\hat{\theta}, \theta) &= E\left\{ \log P(X, S, k \mid \hat{\theta}) \mid X, \theta \right\} \\ &= \sum_{S,k} P(S, k \mid X, \theta) \log P(X, S, k \mid \hat{\theta}) \qquad (7) \end{aligned}$$

where $\theta$ are the current parameter estimates that we use to evaluate the expectation and $\hat{\theta}$ are the new parameters that we optimize to increase $Q(\hat{\theta}, \theta)$. Note that without loss of generality, we represent the entire acoustic observation training data by $X$ even when it consists of independent utterances. Given that

$$P(X, S, k \mid \hat{\theta}) = P(X \mid S, k, \hat{\theta}) P(k \mid S, \hat{\theta}) P(S \mid \hat{\theta}) \tag{8}$$

and since the transform component $k$ remains constant over the utterance, that is $P(k \mid S, \hat{\theta}) = P(k \mid \hat{\theta})$, we may write the auxiliary function as (ignoring all terms not involving $\hat{\theta}$)

$$Q(\hat{\theta}, \theta) = \sum_{S,k} P(S, k \mid X, \theta) \left( \log \hat{v}_k + \log P(X \mid S, k, \hat{\theta}) \right) \tag{9}$$

Here we used the shorthand $\hat{v}_k = P(k \mid \hat{\theta})$ since the mixture coefficients $\hat{v}_k$ can be thought of as prior probabilities of each transform component.

Our goal is to estimate the transform priors, transforms and Gaussian parameters under the ML criterion. This estimation is performed as a three-stage iterative procedure. We first maximize the ML criterion with respect to the transform priors while keeping the transforms and Gaussian parameters fixed. Subsequently, we estimated the transforms using the updated values of the transform priors. Finally, we compute the Gaussian parameters using the updated values of the transform and transform priors.

### 3.1. Transform prior estimation

In the first part of the three stage estimation procedure we fix the Gaussian parameters and transforms and maximize the auxiliary function $Q(\hat{\theta}, \theta)$ with respect to the transform priors. To find the update formula for the transform prior $v_k$ we use the first term in Equation 9 (denoted by $Q_1(\hat{\theta}, \theta)$)

$$\begin{aligned} Q_1(\hat{\theta}, \theta) &= \sum_{S,k} P(S, k \mid X, \theta) \log \hat{v}_k \\ &= \sum_{k} P(k \mid X, \theta) \log \hat{v}_k \qquad (10) \end{aligned}$$

Adding the Lagrangian multiplier and setting the derivative of $Q_1(\hat{\theta}, \theta)$ with respect to $\hat{v}_k$ equal to zero, we get

$$\begin{aligned}\hat{v}_k &= P(k \mid X, \theta) \\ &= \frac{v_k P(X \mid \tau_k, \lambda_H)}{\sum_j v_j P(X \mid \tau_j, \lambda_H)}\end{aligned} \tag{11}$$

where $P(X \mid \tau_k, \lambda_H)$ is the likelihood of the sequence of observations $X$ under transform $\tau_k$ and can be computed via a forward pass.

### 3.2. Transform estimation

In the second part of the estimation procedure we use the updated values of the transform priors and maximize the auxiliary function $Q(\hat{\theta}, \theta)$ with respect to the transforms. To find the update formula for the transforms $\tau_k$ we use the second term in Equation 9 (denoted by $Q_2(\hat{\theta}, \theta)$)

$$\begin{aligned}Q_2(\hat{\theta}, \theta) &= \sum_{S,k} P(S, k \mid X, \theta) \log P(X \mid S, k, \hat{\theta}) \\ &= \sum_{k,s} \sum_{t=1}^{N} P(s_t = s, k \mid X, \theta) \log q(x_t \mid s, \hat{\tau}_k) \\ &= \sum_{k,s} \sum_{t=1}^{N} \gamma_{s,k}(t; \theta) \log q(x_t \mid s, \hat{\tau}_k)\end{aligned} \tag{12}$$

Here, $\gamma_{s,k}(t; \theta)$ is the posterior probability of being at state $s$ in frame $t$, given transform $k$, the observations and transcript. This posterior can be obtained using the forward-backward algorithm using the new transform prior estimates and the old transform and Gaussian model parameters. Also, $q(x_t \mid s, \hat{\tau}_k)$ is the emission density of state $s$ under transform $\hat{\tau}_k$ . Although this Bayesian modeling framework is quite general and can be extended to a variety of normalization techniques, in this paper we study only feature-based acoustic normalization in HMMs, which is commonly termed as Constrained Maximum Likelihood Linear Regression (CMLLR) [7].

The CMLLR technique applies affine transforms to the $m$ dimensional observation vector $x$ so that a normalized feature vector is found as $Ax + b$, where $A$ is a nonsingular $m \times m$ matrix and $b$ is an $m$-dimensional vector. The emission density of state $s$ is assumed to be Gaussian and is therefore reparameterized as

$$q(\zeta \mid s, \tau_k) = \frac{|A_k|}{\sqrt{(2\pi)^m |\Sigma_s|}} \exp\{-\frac{1}{2}(\tau_k \zeta - \mu_s)^T \Sigma_s^{-1} (\tau_k \zeta - \mu_s)\}$$

Here, $\tau_k$ denotes the extended transformation matrix $[b_k \;\; A_k]$; $\zeta$ is the extended observation vector $[1 \;\; x^T]^T$; and $\mu_s$ and $\Sigma_s$ are the mean and variance for the observation distribution of state $s$. The $\Sigma_s$ are constrained to be diagonal covariance matrices. The reestimation formula for the transform $\tau_k$ is found by differentiating $Q_2$ with respect to $[\hat{\tau}_k]$ and solving for its zeros, where $[\tau_k]_i$ denotes the $i$th row of $\tau_k$. A detailed derivation of the transformation parameters is contained in the work of Gales [7] where each row of $\tau_k$ is optimized given the current value of all the other rows.

### 3.3. Gaussian parameter estimation

This section describes the estimation scheme for the Gaussian means and variances under the ML criterion. With the transforms and transform priors estimated as described in sections 3.1 and 3.2 the Gaussian parameters can be derived in a similar fashion. That is, by taking the gradient of $Q_2(\hat{\theta}, \theta)$, given by Equation 12, with respect to the Gaussian mean and variance and solving for its zeros gives

$$\hat{\mu}_s = \frac{\sum_{k=1}^{M} \sum_{t=1}^{N} \gamma_{s,k}(t; \theta) \tau_k \zeta_t}{\sum_{k=1}^{M} \sum_{t=1}^{N} \gamma_{s,k}(t; \theta)} \tag{13}$$

and

$$\hat{\Sigma}_s = \frac{\sum_{k=1}^{M} \sum_{t=1}^{N} \gamma_{s,k}(t; \theta) \tau_k \zeta_t \zeta_t^T \tau_k^T}{\sum_{k=1}^{M} \sum_{t=1}^{N} \sum_k \gamma_{s,k}(t; \theta)} - \hat{\mu}_s \hat{\mu}_s^T \tag{14}$$

Here, the posterior $\gamma_{s,k}(t; \theta)$ is estimated using the new transform and transform prior estimates and the old Gaussian model parameters.

### 3.4. Transform splitting

BSAT treats the mixture of transforms similar to Gaussian mixture training. The process that increases the number of Gaussian components in a mixture is called Gaussian mixture splitting [8]. In this process, the Gaussian with the maximum variance is split in two by a random perturbation of the mean vector. The splitting process is repeated until the required number of components is obtained. The Gaussian with the largest variance is the one for which the training data likelihood is minimum. Therefore, splitting this Gaussian is intuitively consistent with the ML criterion.

While Gaussian splitting is a straightforward procedure, transform splitting is an open question. A transform is a full matrix and the computation of its variance is impractical. The challenge is to find meaningful perturbations of a transform to obtain the initial estimates. A random perturbation of the transform parameters may not be the most judicious choice. As an alternative to random perturbation, we considered to cluster the utterances and estimate initial transforms for each cluster. We investigated two possible utterance clustering schemes: bias clustering and feature clustering.

The idea behind bias clustering is to group the utterances based on the similarity of the transform that is most suitable for each utterance under the ML criterion. In this way, the transforms estimated over each cluster of utterances can capture distinct acoustic phenomena and increase the discriminatory capability of the features across clusters. However, since the estimation of a full matrix for each utterance is impractical, we estimate only a bias term for each utterance. The bias term is estimated on top of the transform chosen to be splited. Then we cluster the utterances based on the similarity of the biases and finally we estimate an initial transform for each cluster of utterances.

Preliminary experimental results showed that the resulting biases have almost zero variance. This behavior is attributed mainly to the application, in the front-end, of mean normalization of the features. As we will describe in Section 5.2, we perform cepstral mean normalization over each speaker turn. We observed that the speaker turns, in the acoustic training corpus used in this work (see Section 5.1), are relatively short and effectively equivalent to the average length of the utterances. Hence, given that we already have normalized the utterances, the bias term of the feature-normalizing transform becomes nearly zero. This greatly diminishes the discriminatory capability of the bias term.

In an effort to improve upon bias clustering we also employed the commonly used k-means clustering procedure to cluster the utterances. In our approach, termed feature clustering, each sample in the k-means algorithm corresponds to an utterance. The centroids

are defined as Gaussian distributions estimated from the sequence of feature vectors. The distance measure between a sample (utterance) and a centroid (Gaussian) is defined as the likelihood of the utterance evaluated by the Gaussian associated with the centroid.

Although in BSAT transform splitting can be applied concurrently with Gaussian splitting, in this paper we are incrementing only the transforms according to the following iterative scheme:

1. Initialize the HMM parameters of BSAT from a speaker independent (SI) model that has already a mixture of Gaussian components as emission densities. In the first iteration, BSAT uses a single transform component, initialized by the identity matrix.
2. At the first 5 iterations split every transform in two; in subsequent iterations split only the top 75% transforms, according to the amount of training data associated with each transform. This approach effectively creates a set of transforms that has a fairly uniform distribution of associated training data.
3. Smooth the newly created transforms with their corresponding parent transform.
4. Use the resulting transforms as initial estimates.

## 4. EFFICIENT BSAT TRAINING AND DECODING

According to Equation 5 the likelihood of an utterance is computed as a weighted sum of the likelihoods obtained by transforming its features based on each of the transforms in the mixture, with weights set to the transform priors. However, this procedure increases the computational load and time efficiency by a factor that equals the number of transforms in the mixture. To alleviate this problem, it is possible to extend the standard decoding procedure by adding a new dimension in the search space that corresponds to the transforms. That is, we can run a *synchronous* decoding procedure by considering all transforms in the mixture in parallel and use a beam threshold to prune transforms dynamically. Given that, preliminary experimental results on the evaluation set (defined in Section 5) showed that the dynamic range of the likelihoods obtained under each transform in the mixture is very wide, the search space after the first few steps can be dramatically reduced. Similarly, we can apply a synchronous training procedure during the forward-backward algorithm by considering all transforms in the mixture in parallel.

Although, the aforementioned synchronous decoding procedure is time efficient and practical, for faster turnaround in this paper we adopted a different decoding procedure that only required to rescore lattices of word hypotheses. The lattices were created by running unadapted decoding using the baseline SI model. As we mentioned above the dynamic range of the likelihoods obtained under each transform is very wide. Thus, the weighted sum of likelihoods, defined in Equation 5, can be approximated by the highest likelihood in the mixture with almost no loss

$$\sum_{k=1}^{M} v_k P(X \mid w, \tau_k, \lambda_H) \approx \operatorname*{argmax}_k v_k P(X \mid w, \tau_k, \lambda_H) \quad (15)$$

Therefore, we can rescore the SI lattice only once using the transform in the mixture that yields the highest likelihood. Then, the key issue is to predict for each utterance which transform yields the highest likelihood. This was achieved by using the BSAT model to compute the likelihood of the utterance under each transform via a forward pass given the 1-best SI hypothesis. The transform with the highest likelihood on the 1-best SI hypothesis was used to rescore the SI lattice. The decoding procedure used for BSAT in this paper is summarized in the following steps:

1. Run unadapted decoding using the SI baseline model and create the word lattice. Also, find the 1-best SI hypothesis.
2. Using the BSAT model compute the likelihood of the utterance under each transform via a forward pass given the 1-best SI hypothesis. Select the transform in the mixture that yields the highest likelihood.
3. Rescore the SI lattice using the BSAT model and the transform selected in step 2.

## 5. EXPERIMENTAL SETUP

### 5.1. Training and test data

The acoustic training corpus used in this work consists of 150 hours of Arabic broadcast news speech data. These include 28 hours of data from the FBIS corpus, 67 hours of data selected from the TDT4 Arabic corpus available from LDC and the remaining 55 hours of speech data selected from an in-house broadcast news database that contains data from various sources. The language model training corpus is a pool of around 400 million-word text. It includes the data from the Gigaword Arabic corpus, TDT4 Arabic corpus, and a few other sources. We also downloaded some data from the website of Aljazeera. All the training data cover various time periods from 1994 to October 2003. To evaluate the recognition performance, we used the BBN 2005 Arabic development set (bnat05) as the test set. It consists of 3.8 hours of data from 9 episodes broadcast by *Aljazeera*, *Dubai Television* and *Lebanese Broadcasting Corporation* in November 2003. The decoding lexicon was created using morphological decomposition [9] and consisted of 64K words selected based on the occurring frequency in the 400M-word language training corpus.

### 5.2. System architecture

The baseline system uses a PLP front-end, computing 14 cepstral coefficients and normalized energy per frame of speech. Cepstral mean normalization was performed over each speaker turn. The actual 60-dimensional features used in acoustic model training are produced by applying LDA+MLLT on sets of 9 contiguous cepstral frames (135 dimensions). The baseline decoding experiments were carried out in a multi-pass search strategy. The forward pass uses a simple acoustic model, a State Tied Mixture (STM) model with 190 state clusters, and a bigram language model, and outputs the most likely wordends at each frame together with their scores. The backward pass then uses the output of the forward pass to guide a Viterbi beam search with a state clustered within-word quinphone acoustic model (1548 state clusters) and a trigram language model. A lattice is also generated. Finally, we do lattice rescoring using a state clustered cross-word quinphone model with 1762 state clusters. The top scoring hypothesis represents the system's recognition output.

## 6. EXPERIMENTAL RESULTS

### 6.1. Speaker independent decoding

Initial baseline experiments were performed to measure the performance of SI models by varying the number of Gaussian mixtures. The rationale behind these experiments was twofold. First, recall that in BSAT we incrementally build the mixture of transforms. Therefore, for fair comparison a SI system should have relatively the same

| Gaussians / State | 12 | 21 | 30 | 64 | 128 |
|---|---|---|---|---|---|
| Unadapted WER | 23.6 | 22.5 | 21.8 | 20.9 | 20.5 |

**Table 1**. Word Error Rate (%) of baseline SI systems trained by varying the number of Gaussian components per state as evaluated on the bnat05 test set. Results are reported without unsupervised speaker adaptation.

number of total parameters as the BSAT system. Second, we wanted to assess the effectiveness of BSAT as we vary the complexity, and therefore the power, of the seed SI models. That is, we explored whether a SI system with a relative large number of Gaussian mixtures is capable to model both speech and non-speech variabilities. Table 1 summarizes the performance of SI systems as we increase the number of Gaussian mixtures. Not surprisingly, the performance of the SI models increases as the number of Gaussian components per state cluster increases.

### 6.2. BSAT unadapted decoding

We then conducted a series of experiments to assess the effectiveness of BSAT as proposed in Section 2. Throughout the BSAT experiments we used a single-class CMLLR transform. We first investigated the BSAT training behavior under the three different transform splitting procedures: random perturbation, bias clustering and feature clustering. Following the transform splitting procedures, described in Section 3.4, all BSAT training experiments were seeded by the 12 Gaussian mixture SI model. Figure 1 shows the log-likelihood per frame across training iterations of each BSAT model. It is apparent that all three splitting procedures yield comparable likelihood. These results show that bias clustering and feature clustering do not improve upon the naive random perturbation procedure.

We then run BSAT decoding using the three BSAT systems described above. Note that BSAT decoding, described in section 4, is an integrated process, where the training-set transforms are being directly used in decoding. To distinguish from the standard adaptation in decoding, where a set of transforms is being estimated with respect to the speakers in the test set, we term this procedure *unadapted* BSAT decoding. The top part of Table 2 compares the performance of the three BSAT systems as we increment the transform mixture. It is apparent that all three transform splitting procedures yield comparable WER, which is consistent with their behavior during training. BSAT yields a 4% relative gain over the SI baseline by using only 16 transforms, and an overall 7% by using 98 transforms. We also observe that a further increase in the number of transforms (i.e. more than 98) does not give additional gains.

Finally, we trained a BSAT model seeded by the 64 Gaussian mixture SI model in order to explore whether the gains from BSAT hold in the presence of a large number of Gaussian components. The bottom part of Table 2 shows the WER on the bnat05 set as we increment the transform mixture. Since all three transform splitting procedures yielded the same performance when BSAT was seeded by the 12 Gaussian mixture SI model, for this experiment we have only used feature clustering. Two observations can be made from the results: First, the BSAT gains (7% relative overall) remain, even when BSAT is seeded from the much larger SI model. Second, unlike the BSAT model with 12 Gaussians mixtures, the BSAT model with 64 Gaussians mixtures consistently increases its performance on the test set as the size of the transform mixture increases. Finally, the BSAT system seeded by the 64 Gaussian mixture SI model outperforms the 128 Gaussian SI model by 5% relative, even though

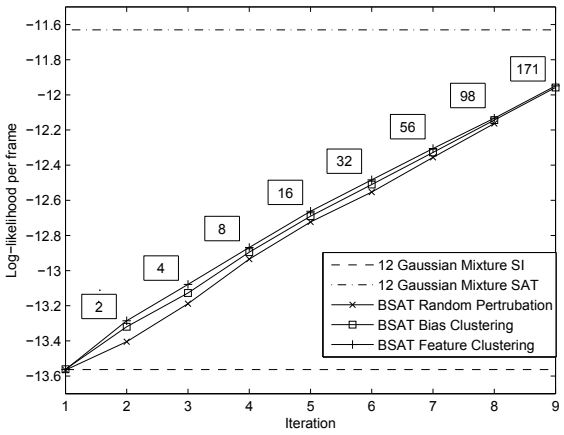


**Fig. 1**. Log-likelihood per frame against iteration number of BSAT models for the the three different transform splitting procedures: random perturbation, bias clustering and feature clustering. The numbers in boxes indicate the number of transforms. The likelihood of the seed 12 Gaussian mixture SI model and the SAT model is shown for reference. The SAT model uses 3394 speaker clusters.

the BSAT system has 30% fewer total parameters (Gaussians and transforms) relative to the 128 Gaussian SI model.

### 6.3. BSAT adapted decoding

In the previous section we explored the efficacy of BSAT in unadapted decoding, where the training-set transforms were integrated into the decoding procedure. However, it is well known that significant gains in performance can be obtained if we reduce the mismatch between the training and test speech data during recognition. This is the goal of the standard transformed-based acoustic normalization and adaptation techniques in decoding, where the transforms are estimated from the speech to be recognized. The amount of the test data used for the estimation of the transformation parameters is usually much lower than the one used for training the initial models. This sometimes leads to poorly trained transforms and over-fitting problems to the initial hypothesis.

To avoid such problems in adapted decoding, we can further utilize the BSAT transforms. For example, the BSAT transforms can be used as initial estimates in CMLLR or as prior information in Maximum a Posteriori Linear Regression (MAPLR) [10] adaptation algorithm. Furthermore, we can introduce lower order adaptation parameters that can act on the BSAT transforms with the goal to adapt them on a test speaker. Although all these issues are of great interest they are not addressed in this paper.

All adapted decoding experiments described in this section estimate first a single-class CMLLR transform and then a 2-class MLLR [11] transform for each speaker in the test set. Table 3 compares the recognition performance of SAT and BSAT models by incorporating unsupervised speaker adaptation on the bnat05 test set. The two baseline SAT models used 3394 speaker clusters and were seeded by the 12 and 64 Gaussian mixture SI models of section 6.1, respectively. The corresponding BSAT models, selected for decoding, were the ones that gave the best unadapted performance according to the results of Table 2. That is, we selected the 12 Gaussian mixture BSAT model with 171 transform components and the 64 Gaussian

| Gaussians / State | Unadapted SI WER | Transform Splitting | Unadapted BSAT WER | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | #Transforms | | | | | |
| | | | 16 | 56 | 98 | 171 | 299 | 523 |
| 12 | 23.6 | Random Perturbation | 22.5 | 22.1 | 22.0 | 22.0 | - | - |
| | | Bias Clustering | 22.6 | 22.1 | 22.0 | 22.0 | - | - |
| | | Feature Clustering | 22.3 | 22.1 | 21.9 | 21.9 | - | - |
| 64 | 20.9 | Feature Clustering | 20.1 | 19.9 | 19.8 | 19.7 | 19.6 | 19.5 |

**Table 2**. Unadapted Word Error Rate (%) results of BSAT systems trained by varying the number of Gaussian and transform mixtures as evaluated on the bnat05 test set. The Word Error Rate (%) results of the seed SI models are shown as the baseline.

mixture BSAT model with 523 transform components.

The BSAT adapted decoding experiments of this section used a modified version of the BSAT unadapted decoding described in section 4 and is summarized in the following steps:

1. Using the BSAT model compute the likelihood of the utterance under each transform via a forward pass given the 1-best BSAT unadapted hypothesis. Select the transform in the mixture that yields the highest likelihood.
2. Estimate a single-class CMLLR transform on top of the BSAT transform found in step 1.
3. Apply the cascaded BSAT and CMLLR transforms.
4. Estimate a 2-class MLLR transform for the BSAT model given the transformed features from step 3.
5. Adapt the Gaussian parameters of the BSAT model using the MLLR transform found in step 4 and transform the features of the utterance by the cascaded BSAT and CMLLR transform.
6. Rescore the lattice created via adapted decoding of the SAT model.

| Gaussians / State | Adapted WER | |
|---|---|---|
| | SAT | BSAT |
| 12 | 19.4 | 19.1 |
| 64 | 17.8 | 17.3 |

**Table 3**. Adapted Word Error Rate (%) results of SAT and BSAT systems trained under various number of Gaussian components as evaluated on the bnat05 test set. Results are reported with unsupervised speaker adaptation.

The results of Table 3 show that adapted decoding with the BSAT models outperforms by 1.5% and 3% relative the SAT models, in systems with 12 and 64 Gaussian components per state cluster, respectively. Moreover, the results indicate that BSAT is more effective when we use more Gaussian components per state cluster. Finally, note that the 12 Gaussian BSAT model uses a mixture of 171 transforms whereas the 12 Gaussian SAT model uses 3394 transforms. However, the training likelihood of the BSAT system is close to the training likelihood of the SAT model, as illustrated in the Figure 1. The same observation is true for the BSAT and SAT systems with 64 Gaussian components per state cluster.

## 7. DISCUSSION

In this paper we proposed a Bayesian adaptive training and decoding technique that uses a mixture of feature-based transforms. We presented ML reestimation formulae for the parameters of the BSAT model and developed various transform splitting procedures. Also, we discussed training and decoding approximations needed for their effective application. BSAT experimental results on Arabic broadcast news exhibited increased likelihood on acoustic training data and improved speech recognition performance on unseen test data, compared to speaker independent and standard adaptive models.

The Bayesian framework adopted in this work allowed the full integration of the transforms into the training and decoding procedure. Therefore, discriminatively-trained transforms which directly aim to minimize recognition performance criteria can be used in decoding. Hence, in the future we are planning to develop BSAT under discriminative criteria. Moreover, we are planning to address several issues in BSAT training and decoding, such as, improved transform splitting techniques, the concurrent splitting of transforms and Gaussians, and more elaborate adaptive decoding procedures.

## 8. REFERENCES


[1] T. Anastasakos et al., "A compact model for speaker-adaptive training," in *ICSLP*, 1996, pp. 1137–1140.

[2] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation," *IEEE Trans. Spch. & Aud. Proc.*, vol. 13, no. 3, pp. 367–376, May 2005.

[3] L. Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," in *ASRU*, 2003.

[4] L. Wang and P. C. Woodland, "MPE-based discriminative linear transform for speaker adaptation," in *ICASSP*, 2004.

[5] K. Yu and M. J. F. Gales, "Bayesian adaptation and adaptively trained systems," in *ASRU*, 2005.

[6] A. P. Dempster, A. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comp. Spch. & Lang.*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[8] A. Sankar, "Experiments with a Gaussian merging-splitting algorithm for HMM training for speech recognition," in *Proc. of he DARPA Wkshp.*, 1998, pp. 99–104.

[9] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological decomposition for arabic broadcast news transcription," in *ICASSP*, 2006, pp. 1089–1092.

[10] C. Chesta, O. Siohan, and C. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *Eurospeech*, 1999, pp. 211–214.

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Spch. & Lang.*, vol. 9, pp. 171–185, Apr. 1995.