BAYESIAN ADAPTATION AND ADAPTIVELY TRAINED SYSTEMS

K. Yu and M.J.F. Gales

Engineering Department, Cambridge University
Trumpington St. Cambridge, CB2 1PZ, U.K.

ABSTRACT

As the use of found data increases, more systems are being built using adaptive training. Here transforms are used to represent unwanted acoustic variability, e.g. speaker and acoustic environment changes, allowing a canonical model that models only the "pure" variability of speech to be trained. Adaptive training may be described within a Bayesian framework. By using complexity control approaches to ensure robust parameter estimates, the standard point estimate adaptive training can be justified within this Bayesian framework. However during recognition there is usually no control over the amount of data available. It is therefore preferable to be able to use a full Bayesian approach to applying transforms during recognition rather than the standard point estimates. This paper discusses various approximations to Bayesian approaches including a new variational Bayes approximation. The application of these approaches to state-of-the-art adaptively trained systems using both CAT and MLLR transforms is then described and evaluated on a large vocabulary speech recognition task.

1. INTRODUCTION

Adaptive training [1, 2] has become popular as the use of *found data*, such as Broadcast News, has increased. In these approaches two sets of model parameters are extracted from the training data. The first set is the *canonical* model parameters, $\mathcal{M}$, which represent the underlying acoustic data variability. The second set, the *transform* model parameters, $\mathcal{T}$, represent any unwanted variability, such as speaker and acoustic condition changes. A separate transform is used to represent each homogeneous block of data, e.g. from a particular speaker/environment combination. Adaptive training is usually derived from a maximum likelihood perspective. However it is closely linked with Bayesian approaches where the model and transform parameters are treated as random variables and marginalised out. In common with many Bayesian schemes this marginalisation is intractable with HMM-based speech recognition systems, though for the model parameters a variational approximation has been examined [3]. However by appropriately controlling the complexity of the system during training, for example using a minimum occupancy threshold when constructing the decision tree and limiting the number of components and transforms, the standard point estimates used in adaptive training can be justified. However during recognition, it is not possible to control the amount of available data, if any, to estimate the test adaptation transform so full Bayesian approaches may be required.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no offi cial endorsement should be inferred.

There have been a number of approaches investigated to allow robust estimates of transform parameters with limited training data. One standard approach is to use Maximum a Posteriori (MAP) estimation [4]. Though these approaches yield more robust estimates than Maximum Likelihood (ML) training, they still only yield a point estimate of the transform. An alternative approach, Bayesian predictive adaptation, employs a real distribution rather than a point estimate[5, 6, 7]. However though a distribution is estimated, to allow the integration to be computed it is usual to assume that the transform can change from frame to frame, the frame independent assumption discussed in [8]. Another approach to make the integration tractable is to use a system to obtain the Viterbi state/component sequence [9]. Though this allows the integral to be computed it may be sensitive to the precise state/component alignments used. Recently, Variational Bayes (VB) approaches [10] have become popular, for example they have been applied to training standard HMM model parameters [3] and scaled mean bias in adaptation [11]. This approach does not introduce crude assumption and has a solid mathematical basis in the sense that it yields a strict lower bound on the likelihoods. This paper examines the application of a VB approach to Bayesian adaptation, along with other approximate schemes. Bayesian adaptation and adaptive training are described in a unified framework, motivating the use of adaptively trained systems to find appropriate priors. The forms of approximation are then applied to the specific tasks of estimating transforms within a Cluster Adaptive Training (CAT) [12] framework and Maximum Likelihood Linear Regression (MLLR) [13] transforms.

2. BAYESIAN SCHEMES AND ADAPTIVE TRAINING

This section reviews the basic theory behind adaptive training within a Bayesian framework, similar to that in [8]. Given a set of training data $\mathcal{O} = \{\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(S)}\}$ where $\mathbf{O}^{(s)}$ represents a particular homogeneous block, the aim of adaptive training is to estimate the transform and model set parameter distributions that maximise

$$p(\mathcal{O}|\mathcal{H}) = \int_{\mathcal{M}} \prod_{s=1}^{S} \left(p(\mathbf{O}^{(s)}|\mathcal{H}, \mathcal{M}) \right) p(\mathcal{M})\, d\mathcal{M} \tag{1}$$

where

$$p(\mathbf{O}^{(s)}|\mathcal{H}, \mathcal{M}) = \int_{\mathcal{T}} p(\mathbf{O}^{(s)}|\mathcal{H}, \mathcal{M}, \mathcal{T}) p(\mathcal{T})\, d\mathcal{T} \tag{2}$$

$p(\mathcal{M})$ and $p(\mathcal{T})$ are the distributions for the canonical model and transform parameters[1] respectively and $\mathcal{H}$ is the transcription for

[1] The distribution of the transform parameters is dependent on the model set, for clarity of notation this dependence has been dropped.

the training data. Normally HMMs, with Gaussian mixture model (GMM) as the state output distributions, are used as the underlying acoustic model. Letting $\Lambda = \{\mathcal{T}, \mathcal{M}\}$,

$$p(\mathbf{O}^{(s)}|\mathcal{H}, \Lambda) = \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P(\boldsymbol{\theta}|\mathcal{M}) \prod_t b(\mathbf{o}_t|\Lambda, \theta_t) \tag{3}$$

where $\boldsymbol{\Theta}$ is the set of all possible component sequences[2] for $\mathcal{H}$, $P(\boldsymbol{\theta}|\mathcal{M})$ is the distribution of a particular component sequence $\boldsymbol{\theta}$, $b(\mathbf{o}_t|\Lambda, \theta_t)$ is the Gaussian distribution at component θ_t.

The direct optimisation of this expression is highly complex. To overcome this problem, a joint distribution, $q(\boldsymbol{\theta}, \Lambda)$, over the component sequence, $\boldsymbol{\theta}$, and parameters, Λ, is introduced. Applying Jensen's inequality yields

$$\log p(\mathcal{O}|\mathcal{H}) \geq \left\langle \log \frac{p(\mathcal{O}, \boldsymbol{\theta}|\Lambda, \mathcal{H}) p(\Lambda)}{q(\boldsymbol{\theta}, \Lambda)} \right\rangle_{q(\boldsymbol{\theta}, \Lambda)} \tag{4}$$

where $< f(x) >_{q(x)}$ denotes the expectation of function $f(x)$ with respect to the distribution of $q(x)$. The above becomes equality when

$$q(\boldsymbol{\theta}, \Lambda) = P(\boldsymbol{\theta}|\mathcal{O}, \mathcal{H}, \Lambda) p(\Lambda|\mathcal{O}, \mathcal{H}) \tag{5}$$

Using equation 5 is impractical, so approximations are used, this is the class of approaches known as *variational Bayes* [10, 3]. These approximations mean that a lower bound rather than the actual likelihood is optimised. The tightness of the bound is dependent on the form used.

When building speech recognition systems it is possible to control the complexity of the system being trained. For example minimum occupancies may be applied during the construction of the decision tree, and homogeneous blocks clustered together, to ensure robust estimates of the parameters. Using these standard approaches it is usual to approximate the distributions over the model parameters using point estimates, the distribution is a delta function. Effectively the variances of the parameter estimates is assumed to be sufficiently small that this approximation is reasonable. With this approach it is possible to use the Expectation Maximisation (EM) algorithm. The joint distribution at iteration $k+1$ is obtained using the previous iteration's estimate, $\hat{\Lambda}_k$. The lower bound in equation 4 is usually written as an *auxiliary function*, $\mathcal{Q}(\Lambda_{k+1}, \hat{\Lambda}_k)$. Thus

$$q(\boldsymbol{\theta}, \Lambda) = P(\boldsymbol{\theta}|\mathcal{O}, \mathcal{H}, \hat{\Lambda}_k) \tag{6}$$

This allows the standard iterative adaptive training formulation to be used [1], which interleaves estimating the model parameters and the transform parameters. Each iteration is guaranteed to increase the right-hand side of equation 4, hence decreasing the difference between the approximation and the likelihood using the "true" point estimate. Eventually when the value of current estimate of the parameters is the "true", either ML or MAP, estimate, the point form of the equality in equation 5 is obtained.

It is interesting to compare this lower bound approximation for recognition to standard iterative approaches such as iterative MLLR. In iterative MLLR, a transform is estimated using the 1-best hypothesis. This transform is then used to re-recognise the data and the process repeated if necessary. When used in recognition the lower bound in equation 4 requires that for *each hypothesis* a transform is estimated and used to obtain an estimate of the log-likelihood for that hypothesis. This should result in a tighter lower bound for the hypothesis, other than the 1-best, than iterative MLLR. This should reduce the inherent biases to the adaptation transcription seen in iterative MLLR.

One effect of ensuring that there is sufficient data to obtain robust estimates is that the distribution over the model parameters, $p(\Lambda)$, is not normally required. The only aspect of the adaptively trained system required for recognition is the estimated model parameters $\hat{\mathcal{M}}$. However for many situations there may be limited, or even no, test adaptation data available. For these situations it is useful to also extract the distribution of the transform parameters. From the training data, a point estimate of the transform for each of the homogeneous blocks, $\{\hat{\mathcal{T}}^{(1)}, \ldots, \hat{\mathcal{T}}^{(S)}\}$, is found. This data may be directly used to estimate the prior. The precise form of this prior is important. It is preferable to use a conjugate prior distribution as it commonly leads to tractable mathematical forms [10]. For the form of bound in equation 4, this is often a single Gaussian. Unfortunately a single component is not always powerful, for example in the case of CAT [2], the distribution of the interpolation weights may be highly bimodal. For these instances it makes sense to use a GMM as the prior distribution for the transform parameters. Using GMMs further complicates the training. Consider the N-component prior distribution of the form

$$p(\mathcal{T}) = \sum_{n=1}^{N} c_n \mathcal{N}(\mathcal{T}; \boldsymbol{\mu}_{\mathcal{T}}^{(n)}, \boldsymbol{\Sigma}_{\mathcal{T}}^{(n)}) \tag{7}$$

where c_n is the component prior. For MAP estimation it is no longer possible to directly apply the prior, which is possible in the case of a single component [14]. Instead Jensen's inequality must again be used. At iteration $k+1$

$$\log p(\mathcal{T}) \geq \left\langle \frac{\log c_n p(\mathcal{T}_{k+1}|n)}{q_k(n)} \right\rangle_{q_k(n)} \tag{8}$$

where $q_k(n) = P(n|\hat{\mathcal{T}}_k)$. Though this will decrease the tightness of the bound, hence requiring additional iterations, it allows the MAP estimates of the transform parameters to be simply obtained. Substituting this into the transform estimation auxiliary function yields (ignoring constants)

$$\begin{aligned} \mathcal{Q}(\mathcal{T}_{k+1}, \hat{\mathcal{T}}_k) &= \langle \log p(\mathbf{O}, \boldsymbol{\theta}|\mathcal{T}_{k+1}, \mathcal{H}) \rangle_{P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \hat{\mathcal{T}}_k)} \\ &\quad + \sum_n P(n|\hat{\mathcal{T}}_k) \log p(\mathcal{T}_{k+1}|n) \end{aligned} \tag{9}$$

The precise effect of this on the MAP estimation is shown in detail in section 4. It is worth noting that equation 9 is also the lower bound used for inference or recognition.

3. BAYESIAN ADAPTATION APPROXIMATIONS

The previous section discussed adaptive training and estimation of the prior transform distribution. This section examines how this transform distribution may be used in situations where there is no adaptation data available to estimate the transform. The aim is to select the hypothesis that maximises

$$p(\mathbf{O}|\mathcal{H}) = \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H}, \mathcal{T}) p(\mathcal{T}) \, d\mathcal{T} \tag{10}$$

where $\mathbf{O}$ is assumed to belong to a single homogeneous block. In the same fashion as adaptive training, this integral is intractable unless for example the point estimates from the previous section are

[2] Using the component sequence as the hidden variable sequence is for deriving formulae for updating component parameters. This is a natural extension to using the state sequence.

used. Approximations to this integral are discussed in this section. Though no test adaptation data is used, equation 10 results in a very different recognition process to standard speaker-independent (SI) recognition. For a homogeneous block, the transform, representation of a speaker/acoustic condition, is constrained to remain constant. In contrast, for SI recognition it can usually change from frame to frame (the standard HMM assumption).

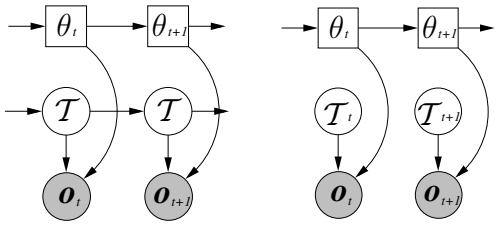
3.1. Sampling approximation

Sampling approaches are a standard method for approximating intractable integrals. The basic idea is to draw samples from the distribution and use the average integral function value to approximate the real probabilistic expectation. Thus

$$p(\mathbf{O}|\mathcal{H}) \approx \frac{1}{N}\sum_{n=1}^{N} p(\mathbf{O}|\mathcal{H},\hat{\mathcal{T}}_n) \quad (11)$$

where N is the total number of samples, $\hat{\mathcal{T}}_n$, independently drawn from $p(\mathcal{T})$. In the limit as $N \rightarrow \infty$ this will tend to the true integral. There is a fundamental issue associated with this form of approximation. As the number of transform parameters increases the number of samples required to obtain good estimates dramatically increases. As a separate decode is required for each sample to find the final best hypothesis, this approach is only applicable to systems with small numbers of adaptation parameters such as CAT. It is worth noting that in contrast to using the ML or MAP point estimates there is no iterative process required, the likelihoods are simply estimated from the samples drawn from $p(\mathcal{T})$.

3.2. Frame-independent (FI) assumption



1(a) Strict Adaptation **1(b)** FI Assumption

Fig. 1. Dynamic Bayesian networks

Rather than approximating the integral, an alternative approach is used to alter the dynamic Bayesian network (DBN) associated with the recognition process. Figure 1(a) shows the DBN for decoding with adaptively trained systems. Here there is a first-order Markov state/component process and the constraint that the transform is the same for all frames of the same homogeneous block. Mathematically this yields the integral in equation 10. If the constraint that the transform is the same for all observations is relaxed, then the DBN in figure 1(b) will be obtained. This allows the transform to vary from time instance to time instance and hence will be referred to as the *frame-independent* assumption[3], see for example [8, 6]. Using this approximation yields

$$p(\mathbf{O}|\mathcal{H}) \approx \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathcal{M}) \prod_t \int_{\mathcal{T}} b(\mathbf{o}_t|\mathcal{T},\theta_t) p(\mathcal{T})\, d\mathcal{T} \quad (12)$$

[3]This resultant distribution is referred to as a *predictive distribution* [6].

With the appropriate form of $p(\mathcal{T})$, this frame-level integral is tractable [8, 6]. In common with the sampling scheme, no iterative estimation scheme is required and standard decoding may be used. However for the no adaptation data case, this FI assumption is very close to the standard SI system. Unless a multiple-component prior is the used, the results with FI approximation will be similar to the SI performance.

3.3. Variational Bayes (VB) approximation

For the FI approximation it is not possible to state how close the approximation is to the actual likelihood. In contrast, if the variational Bayes approximation in equation 4 is used, it is guaranteed to yield a lower bound on the actual likelihood. One simple variational approximation is to use a point estimate, as in equation 6. However it would be preferable to use a distribution. In order to make the calculation tractable, the distributions of the component posterior and the transform posterior are assumed to be conditionally independent. Thus

$$q(\boldsymbol{\theta},\mathcal{T}) = q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H}) q(\mathcal{T}|\mathbf{O},\mathcal{H}) \quad (13)$$

For simplicity of notation these distributions will be denoted as $q(\boldsymbol{\theta})$ and $q(\mathcal{T})$. The aim is to now obtain forms of $q(\boldsymbol{\theta})$ and $q(\mathcal{T})$ that maximise the RHS of equation 4, hence making the lower bound as tight as possible. This is the *Variational Bayesian EM* (VBEM) algorithm [10]. In the same fashion as EM, VBEM is an iterative process. At each iteration the bound is guaranteed to become tighter. The process is:

1) Initialise: $q_0(\mathcal{T}) = p(\mathcal{T})$, $k = 1$.

2) VB expectation (VBE): the optimal variational posterior component distribution can be shown to be

$$q_k(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H})} \exp\left(\langle \log p(\mathbf{O},\boldsymbol{\theta}|\mathcal{T},\mathcal{H}) \rangle_{q_{k-1}(\mathcal{T})} \right) \quad (14)$$

where $\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H})$ is the normalisation term to make $q_k(\boldsymbol{\theta})$ a valid distribution, $q_k(\mathcal{T})$ is the variational transform distribution of the k^{th} iteration. As $\log p(\mathbf{O},\boldsymbol{\theta}|\mathcal{T},\mathcal{H})$ can be factorised, the expectation with respect to $q_k(\mathcal{T})$ can be done at frame-level in the logarithm domain. This allows $q_k(\boldsymbol{\theta})$ to be viewed as a posterior component sequence distribution of a model set with a modified Gaussian component

$$\tilde{b}(\mathbf{o}_t|\theta_t) = \exp\left(< \log b(\mathbf{o}_t|\mathcal{T},\theta_t) >_{q_{k-1}(\mathcal{T})} \right) \quad (15)$$

$\tilde{b}(\mathbf{o}|\theta)$ is sometimes referred to as a pseudo-distribution because it is not necessarily correctly normalised. $\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H})$ can be simply calculated using the standard forward algorithm with $\tilde{b}(\mathbf{o}|\theta)$,

$$\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H}) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} P(\boldsymbol{\theta}|\mathcal{M}) \prod_t \tilde{b}(\mathbf{o}_t|\theta_t) \quad (16)$$

3) VB maximisation (VBM): find $q_k(\mathcal{T})$ by maximising

$$\mathcal{Q}\left(q_k(\boldsymbol{\theta}), q_k(\mathcal{T})\right) = \langle \log p(\mathbf{O},\boldsymbol{\theta}|\mathcal{T},\mathcal{H}) \rangle_{q_k(\boldsymbol{\theta}) q_k(\mathcal{T})} + \int_{\mathcal{T}} q_k(\mathcal{T}) \log \frac{p(\mathcal{T})}{q_k(\mathcal{T})}\, d\mathcal{T} \quad (17)$$

unless converged **goto (2)**, $k = k + 1$.

In contrast to the point estimates where after convergence of the EM algorithm the equality constraint for the point estimate version of equation 4 applies, using the approximation of equation 13 means that the equality will not be achieved. Furthermore it is still not possible to marginalise over the parameters with this variational approximation. Hence for decoding this VB lower bound, $\mathcal{Q}\left(q_k(\mathcal{T})\right)$, must be computed. This can be achieved using the final estimate of the transform distribution (iteration k), $q_k(\mathcal{T})$,

$$\mathcal{Q}\left(q_k(\mathcal{T})\right) = \log \mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H}) + \int_{\mathcal{T}} q_k(\mathcal{T}) \log \frac{p(\mathcal{T})}{q_k(\mathcal{T})}\, d\mathcal{T} \quad (18)$$

This lower bound is used for *inference*, under the assumption that the ordering of $\log p(\mathbf{O}|\mathcal{H})$ is similar to that of the VB lower bound. This assumption is expected to be better when the VB lower bound is close enough to the marginal likelihood. Hence the importance of making the lower bound as tight as possible.

Similar to the MAP approach, if $p(\mathcal{T})$ is a mixture model, the VB lower bound with $p(\mathcal{T})$ can not be directly optimised. A posterior component weight has to be introduced. The logarithm of the marginal likelihood is then approximated by

$$\begin{aligned}\log p(\mathbf{O}|\mathcal{H}) \;&\geq\; \left\langle \log \frac{c_n}{q_k(n)} \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H},\mathcal{T}) p(\mathcal{T}|n)\, d\mathcal{T} \right\rangle_{q_k(n)} \\ &\geq \left\langle \mathcal{Q}\big(q_{k-1}(\mathcal{T}|n)\big) \right\rangle_{q_k(n)} + \left\langle \log \frac{c_n}{q_k(n)} \right\rangle_{q_k(n)} \end{aligned} \quad (19)$$

To simplify the calculation of $q(n)$, the components of the prior are assumed to be independent of each other. Hence the component sequence $\boldsymbol{\theta}$ in $\mathcal{Q}(q(\mathcal{T}|n))$ may alter from an prior component to another. Thus $q_k(n)$ is calculated in the VBE step by

$$q_k(n) = \frac{c_n \exp\left(\mathcal{Q}\left(q_{k-1}(\mathcal{T}|n)\right)\right)}{\sum_n c_n \exp\left(\mathcal{Q}\left(q_{k-1}(\mathcal{T}|n)\right)\right)} \quad (20)$$

In the VBM step, the auxiliary function for $q(\mathcal{T}|n)$ is similar to equation 17 except for using the n^{th} component of $p(\mathcal{T})$ and $q(\mathcal{T})$. Note that $q(\boldsymbol{\theta})$ here is calculated based on the complete $q(\mathcal{T})$ rather than a particular component.

4. APPLICATION TO MLLR AND CAT

Section 3 introduced the general form of various approximation schemes. In this section, these schemes are applied to two specific types of transforms. Cluster Adaptive Training (CAT) [2] and Maximum Likelihood Linear Regression (MLLR) [13].

4.1. Cluster Adaptive Training

In CAT the transform used to represent the unwanted variability is a set of interpolation weights that operate on the means of a set of clusters, or *eigenvoices*. The estimate of the adapted mean for component m, $\hat{\boldsymbol{\mu}}^{(m)}$, is given by

$$\hat{\boldsymbol{\mu}}^{(m)} = \sum_{p=1}^{P} \lambda_p \boldsymbol{\mu}_p^{(m)} = \mathbf{M}^{(m)} \boldsymbol{\lambda} \quad (21)$$

where $\hat{\boldsymbol{\mu}}^{(m)}$ is the adapted mean vector of Gaussian component m, $\mathbf{M}^{(m)} = [\boldsymbol{\mu}_1^{(m)}, \cdots, \boldsymbol{\mu}_P^{(m)}]$ is the cluster mean vector for component m, P is the number of clusters, and $\boldsymbol{\lambda}$ is a $P \times 1$ interpolation weight vector. As the number of interpolation weights used is small, typically only 2 or 3, it is possible to use the sampling approaches discussed in section 11.

Though the MAP estimate for the single component prior case has already been derived for CAT [12], the multiple component case has not been considered. Using the variational approximation in equation 9, the MAP estimate of $\boldsymbol{\lambda}$ at iteration k is

$$\hat{\boldsymbol{\lambda}}_k = \left(\sum_{n=1}^{N} q(n) \boldsymbol{\Sigma}_{\mathcal{T}}^{(n)-1} + \mathbf{G} \right)^{-1} \left(\sum_{n=1}^{N} q(n) \boldsymbol{\Sigma}_{\mathcal{T}}^{(n)-1} \boldsymbol{\mu}_{\mathcal{T}}^{(n)} + \mathbf{k} \right) \quad (22)$$

where the standard CAT sufficient statistics are used

$$\mathbf{G} = \sum_m \sum_t \gamma_m(t) \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \mathbf{M}^{(m)} \quad (23)$$

$$\mathbf{k} = \sum_m \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \left(\sum_t \gamma_m(t) \mathbf{o}_t \right) \quad (24)$$

$\gamma_m(t)$ is the component posterior derived from $P(\boldsymbol{\theta}|\mathbf{O},\mathcal{H},\hat{\boldsymbol{\lambda}}_{k-1})$.

For the frame independent assumption it is necessary to obtain the predictive distribution. For both the CAT case here and the MLLR transform in the next section, if the original distribution and prior are both GMMs, then the resultant predictive distribution will also be a GMM. Thus for a particular Gaussian component m

$$\int_{\mathcal{T}} b(\mathbf{o}|\mathcal{T},m) p(\mathcal{T})\, d\mathcal{T} = \sum_{n=1}^{N} c_n \mathcal{N}(\mathbf{o}; \tilde{\boldsymbol{\mu}}^{(mn)}, \tilde{\boldsymbol{\Sigma}}^{(mn)}) \quad (25)$$

where the prior distribution is given in equation 7. For CAT the parameters of this distribution can be shown to be

$$\begin{aligned} \tilde{\boldsymbol{\mu}}^{(mn)} &= \mathbf{M}^{(m)} \boldsymbol{\mu}_{\mathcal{T}}^{(n)} \\ \tilde{\boldsymbol{\Sigma}}^{(mn)} &= \mathbf{M}^{(m)} \boldsymbol{\Sigma}_{\mathcal{T}}^{(n)} \mathbf{M}^{(m)T} + \boldsymbol{\Sigma}^{(m)} \end{aligned} \quad (26)$$

It is interesting to note that even if the prior and original Gaussian distribution both have diagonal covariance matrices, the resultant predictive distribution has a full covariance matrix.

Using VB as the approximation it is necessary to find the pseudo-distribution, $\tilde{b}(\mathbf{o}|\theta)$ given in equation 15. For CAT the pseudo distribution for component m is [4]

$$\begin{aligned}\log \tilde{b}(\mathbf{o}|m) = \sum_{n=1}^{N} q(n) \Big(& \log \mathcal{N}(\mathbf{o}; \mathbf{M}^{(m)} \tilde{\boldsymbol{\mu}}_{\mathcal{T}}^{(n)}, \boldsymbol{\Sigma}^{(m)}) \\ & - \frac{1}{2} \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{T}}^{(n)} \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \mathbf{M}^{(m)}) \Big) \end{aligned} \quad (27)$$

The final variational distribution required is $q(\mathcal{T})$. For both CAT and MLLR this distribution can also be shown to be a GMM

$$q(\mathcal{T}) = \sum_{n=1}^{N} q(n) \mathcal{N}(\mathcal{T}; \tilde{\boldsymbol{\mu}}_{\mathcal{T}}^{(n)}, \tilde{\boldsymbol{\Sigma}}_{\mathcal{T}}^{(n)}) \quad (28)$$

where the component parameters are given by

$$\tilde{\boldsymbol{\Sigma}}_{\mathcal{T}}^{(n)} = \left(\boldsymbol{\Sigma}_{\mathcal{T}}^{(n)-1} + \mathbf{G} \right)^{-1} \quad (29)$$

$$\tilde{\boldsymbol{\mu}}_{\mathcal{T}}^{(n)} = \tilde{\boldsymbol{\Sigma}}_{\mathcal{T}}^{(n)} \left(\boldsymbol{\Sigma}_{\mathcal{T}}^{(n)-1} \boldsymbol{\mu}_{\mathcal{T}}^{(n)} + \mathbf{k} \right) \quad (30)$$

where $\mathbf{G}$ and $\mathbf{k}$ are the standard sufficient statistics given in equations 23 and 24, except that $\gamma_m(t)$ is calculated based on the pseudo-distribution given in equation 27. The component weight of the GMM, $q(n)$, is updated using equation 20.

[4] Iteration index k is dropped for clarity of notation.

4.2. Maximum Likelihood Linear Regression

In MLLR a linear transformation of the mean parameters of the model set is used to represent the unwanted variability. Thus

$$\hat{\boldsymbol{\mu}}^{(m)} = \mathbf{A}\boldsymbol{\mu}^{(m)} + \mathbf{b} = \mathbf{W}\boldsymbol{\xi}^{(m)} \tag{31}$$

where $\boldsymbol{\xi}^{(m)} = [\boldsymbol{\mu}^{(m)T} \quad 1]^T$ is the extended mean vector, $\mathbf{W} = [\mathbf{A} \quad \mathbf{b}]$ is the extended linear transform. A GMM may be used as the prior distribution, but now each row of the transform is assumed to be independent given the prior component. Thus

$$p(\mathcal{T}) = \sum_{n=1}^{N} c_n \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d; \boldsymbol{\mu}_{\mathbf{w}_d}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{w}_d}^{(n)}) \tag{32}$$

where D is the size of the original mean vector, $\mathbf{w}_d^T$ is the d^{th} row of $\mathbf{W}$. This row-independent assumption is consistent with the diagonal covariance matrix used for HMM systems [8]. For MLLR transforms, the parameters of the Gaussian component of the predictive distribution have already been derived in [6, 8] and are not reproduced here. The single component prior form of MAP was presented in [14]. The multiple component prior MAP estimate is a straightforward extension, yielding forms similar to that for CAT given in equation 22.

For the VB approximation, the pseudo distribution is first required. Again this can be shown to an unnormalised distribution, where component m has the form

$$\log \tilde{b}(\mathbf{o}|m) = \sum_{n=1}^{N} q(n)\Big(\log \mathcal{N}(\mathbf{o}; \tilde{\mathbf{W}}_{\boldsymbol{\mu}}^{(n)} \boldsymbol{\xi}^{(m)}, \boldsymbol{\Sigma}^{(m)}) - \frac{1}{2} \sum_{d=1}^{D} \frac{\boldsymbol{\xi}^{(m)T} \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d}^{(n)} \boldsymbol{\xi}^{(m)}}{\boldsymbol{\sigma}_{dd}^{(m)}} \Big) \tag{33}$$

where $\tilde{\mathbf{W}}_{\boldsymbol{\mu}}^{(n)} = [\tilde{\boldsymbol{\mu}}_{\mathbf{w}_1}^{(n)T}, \cdots, \tilde{\boldsymbol{\mu}}_{\mathbf{w}_D}^{(n)T}]^T$ is the mean of the n^{th} transform prior component. Given the pseudo-distribution, $q(\mathcal{T})$ can be updated. This is similar to equation 28, but modified to reflect the independence assumption between rows of the transform shown in equation 32. The n^{th} component's mean and variance for row d can be shown to be

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d}^{(n)} &= \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{(n)-1} + \mathbf{G}^{(d)} \right)^{-1} \\ \tilde{\boldsymbol{\mu}}_{\mathbf{w}_d}^{(n)} &= \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d}^{(n)} \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{(n)-1} \boldsymbol{\mu}_{\mathbf{w}_d}^{(n)} + \mathbf{k}^{(d)} \right) \end{aligned} \tag{34}$$

where $\boldsymbol{\mu}_{\mathbf{w}_d}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathbf{w}_d}^{(n)}$ are the parameters of the n^{th} prior component, $\mathbf{G}^{(d)}$ and $\mathbf{k}^{(d)}$ have the same form as the standard statistics used in MLLR estimation

$$\mathbf{G}^{(d)} = \sum_t \sum_m \frac{\gamma_m(t)}{\boldsymbol{\sigma}_{dd}^{(m)}} \boldsymbol{\xi}^{(m)} \boldsymbol{\xi}^{(m)T} \tag{35}$$

$$\mathbf{k}^{(d)} = \sum_t \sum_m \frac{\gamma_m(t) o_{t,d}}{\boldsymbol{\sigma}_{dd}^{(m)}} \boldsymbol{\xi}^{(m)} \tag{36}$$

but the component posterior, $\gamma_m(t)$, is based on the pseudo-distribution given in equation 33.

5. EXPERIMENTAL RESULTS

The performance of the various Bayesian adaptation approximations were evaluated on a large vocabulary speech recognition system, conversational telephone speech task. The training data set consists of 5446 speakers (2747 female, 2699 male), about 295 hours of data. The performance was evaluated on the 2003 evaluation test dataset, `eval03`, consisting of 144 speakers (77 female, 67 male), about 6 hours of data. All systems used a 12-dimensional PLP front-end with log energy and its first, second and third derivatives with Cepstral mean and variance normalisation and VTLN. An HLDA transform was then applied to reduce the feature dimension to 39. Though the use of normalisation techniques may reduce the possible gain from adaptation, it gave a more realistic baseline. A standard decision-tree state-clustered triphones with an average of 16 Gaussian components per state was constructed as the starting point for the adaptive training. This is the baseline speaker-independent (SI) model. Two adaptively trained systems were then built. The first was a 2-cluster CAT system, initialised using gender information [12]. For this CAT system a 2-component prior was estimated from the training transforms. The second was a SAT system constructed using MLLR, where a single component prior was estimated from the training data. For the CAT system a global transform was used and for the SAT system separate speech and silences transforms were used, the priors for which were independently estimated.

As Viterbi decoding is not possible for the variational approximation in equation 4, N-best rescoring is employed for recognition. A 150-best list was generated using the SI system. All results shown are based on this N-best list. Though the use of N-best lists can limit performance differences, using spot-checks on for example the frame independent configuration this was not found to be a major problem. To illustrate the effects of these Bayesian approaches to adaptation, the homogeneous blocks considered here were based on a single utterance, not as in the standard case on a side basis. For the `eval03` test set the average utterance length was 3.13 seconds, compared to the average side length of 153.75 seconds. This dramatically limits the available data for estimating transforms.

Approx.	$q(\boldsymbol{\theta}, \mathcal{T})$ basis	`eval03` CAT	`eval03` SAT
Speaker Independent		32.83	
ML	$\hat{\mathcal{T}}_0^{\text{ML}}$	32.83	33.44
	$\hat{\mathcal{T}}_1^{\text{ML}}$	32.19	35.16
MAP	$\hat{\mathcal{T}}_0^{\text{MAP}}$	32.85	33.47
	$\hat{\mathcal{T}}_1^{\text{MAP}}$	32.17	31.76
Sampling	—	32.16	—
FI	—	32.48	32.90
VB	$q_0(\mathcal{T})$	32.99	34.12
	$q_1(\mathcal{T})$	32.14	31.50

Table 1. % WER 295hr CAT and SAT systems with 2-mixture and 1-mixture Gaussian prior transform distribution respectively

Table 1 shows the performance of both point estimate and Bayesian approximation techniques. The baseline performance for the SI model was 32.83% on this task. The first set of experiments used the ML and MAP point estimates discussed in section 2. If no prior information is used then the ML estimate of the CAT [2] and MLLR [13] are obtained. However in contrast to the standard approaches a transform is computed for *every* one of the N-best hypothesis, as this corresponds to the point estimate of the variational-approximation in equation 4. As the estimation of the transform parameters relies on EM, the approximation $q(\boldsymbol{\theta}, \mathcal{T})$

must be considered. Two sets of results were obtained. The zeroth iteration used the SI model component posteriors for CAT and the posteriors using an identity transformation for SAT. The first iteration then used the transforms estimated from the zeroth stage, $\hat{\mathcal{T}}_1$, to obtain the posteriors. For CAT, where very little data is required to estimate the transforms, the performance on the first iteration was better than the SI model, in this case by about 0.7% absolute. On the zeroth iteration the performance was about the same as the SI system. This illustrates the sort of degradation that can result when the bound is too loose. For SAT, where an MLLR transform must be estimated, the performance at the first iteration was about 2.3% absolute worse than that of the SI system. this was expected as the transform parameters were estimated using an average of only 300 frames. This problem is partially solved using MAP estimation. For SAT this gave about 1.1% absolute gain over the SI system, showing the importance of the use of prior information when estimating transforms with little data.

For CAT the simple sample approximation to the Bayesian integral could be used. Here, 200 samples were drawn from the CAT prior distribution and used to rescore the N-best lists. This may be viewed as a bound on the performance. This gave an error rate the same as the MAP system, about 0.7% absolute better than the SI system. The frame-independent approximation could be used for both the CAT and the SAT systems. For the CAT system as a 2-component prior was used the decoding involved an average of 32-components per-state, in addition to each component having a full-covariance matrix. The performance of the FI system was slightly disappointing. For the CAT system the gain over the SI system was only 0.35% and for the SAT system the performance was marginally worse. This shows the effect of not constraining the transform to be consistent for each homogeneous block.

The final approximation considered was the VB approximation. This should yield more robust estimates as a distribution over the transform parameters is used rather than a point estimate. Again, as a VBEM approach is used, the form of the variational approximation must be considered. Here the zeroth iteration, $q_0(\mathcal{T})$, where the transform prior was used, and the first iteration, $q_1(\mathcal{T})$, using the distribution from the zeroth iteration, were examined. Again the importance of the tight bound was illustrated. For the looser bound of the zeroth iteration the performance was actually worse than that of the SI system. On the first iteration, using $q_1(\mathcal{T})$, the performance for CAT was about the same as that of the sampling and MAP approaches. For the SAT system the VB approximation was 0.3% absolute better than the MAP system, this is statistically significant using the pair-wise significance test.

6. CONCLUSION

This paper has described adaptive training within a Bayesian framework. This motivates the use of adaptive training to give both the canonical model set to be adapted during recognition and the prior distribution for the transform parameters. As the complexity can be simply controlled during training to reflect the amount of data, the use of Bayesian schemes for training are not considered. Instead this paper examines the application of Bayesian approaches to applying a transform distribution during recognition. Three non-point estimate approaches are described, sampling, the frame-independent predictive distribution, and a variational Bayes approximation. These are compared to point estimate schemes based on both ML and MAP estimation. These approaches are examined on a conversational telephone speech task. In order to restrict the amount of data, the homogeneous blocks for this task were considered at the utterance, rather than the side, basis. Using this set-up the use of transform distributions estimated and applied using a variational Bayes approach significantly outperformed all other approaches. Though the task considered is slightly artificial in the sense that the homogeneous blocks are only considered at the utterance level, this is equivalent to the start of any adaptive recognition process. Furthermore the framework described can be simply extended to employing a posterior transform distribution if adaptation data is available, which is referred to as *posterior adaptation* [15].

7. REFERENCES

[1] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[2] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.

[3] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational Bayesian approach to speech recognition," in *NIPS 15*, 2003.

[4] W. Chou, "Maximum a-posterior linear regression with elliptical symmetric matrix variate priors," *Proc. ICASSP*, pp. 1–4, 1999.

[5] A. C. Surendran and Chin-Hui Lee, "Transformation based Bayesian prediction for adaptation of HMMs," *Speech Communication*, vol. 34, pp. 159–174, 2001.

[6] Jen-Tzung Chien, "Linear regression based Bayesian predictive classification for speech recognition," *IEEE transactions on speech and audio processing*, vol. 11, pp. 70–79, 2003.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Bayesian adaptation revisited," *Proc. ISCAITRW ASR2000*, pp. 112–119, 2000.

[8] M. J. F. Gales, "Acoustic factorization," in *Proc. ASRU*, 2001.

[9] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on Viterbi Bayesian predictive classification," in *Proc. ICASSP*, 1997.

[10] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University College London, 2003.

[11] S. Watanabe and A. Nakamura, "Acoustic model adaptation based on coarse/fine training of transfer vectors and its application to a speaker adaptation task," in *Proc. ISLP*, 2004.

[12] M. J. F. Gales, "Cluster adaptive training for speech recognition," in *Proc. ICSLP*, 1998, pp. 1783–1786.

[13] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[14] C. Chesta, O. Siohan, and C. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," *Proc. EuroSpeech*, vol. 1, pp. 211–214, 1999.

[15] M. J. F. Gales, "Adaptive training for robust ASR," in *Proc. ASRU*, 2001.