Bayesian Adaptive Inference and Adaptive Training

Kai Yu and Mark J. F. Gales

Abstract— Large vocabulary speech recognition systems are often built using found data, such as broadcast news. In contrast to carefully collected data, found data normally contains multiple acoustic conditions, such as speaker or environmental noise. Adaptive training is a powerful approach to build systems on such data. Here transforms are used to represent the different acoustic conditions and then a canonical model is trained given this set of transforms. This paper describes a Bayesian framework for adaptive training and inference. This framework addresses some limitations of standard ML approaches. In contrast to the standard approach, the adaptively trained system can be directly used in unsupervised inference, rather than having to rely on initial hypotheses being present. In addition, for limited adaptation data, robust recognition performance can be obtained. The limited data problem often occurs in testing as there is no control over the amount of the adaptation data available. In contrast, for adaptive training, it is possible to control the system complexity to reflect the available data. Thus, the standard point estimates may be used. As the integral associated with Bayesian adaptive inference is intractable, various marginalisation approximations are described, including a variational Bayes approximation. Both batch and incremental modes of adaptive inference are discussed. These approaches are applied to adaptive training of maximum likelihood linear regression and evaluated on a large vocabulary speech recognition task. Bayesian adaptive inference is shown to significantly outperform standard approaches.

Index Terms— adaptive training, Bayesian inference, Bayesian adaptation, incremental, variational Bayes

I. INTRODUCTION

ADAPTIVE training [1], [2] has become increasingly popular as greater use has been made of *found* data, such as broadcast news. For these forms of data, it is not possible to control the "non-speech" acoustic conditions, such as speaker or environmental noise, which affect the acoustic signals. These changes in acoustic conditions lead to variabilities in the signal that are not associated with the words uttered. Found training data is thus highly *non-homogeneous* with multiple acoustic conditions being present in the training corpus. One approach for building systems on non-homogeneous data is *multi-style* training [3]. Here all training data are treated as a single block to train the HMMs, for example, speaker-independent training. These multi-style systems model both speech and non-speech variabilities. Alternatively, the non-homogeneity of the training data may be handled by first training a set of transforms, one for each of the acoustic conditions (or homogeneous block). Then a *canonical model* is trained given this set of transforms. This is *adaptive training*.

Adaptive training is usually derived from a maximum likelihood (ML) perspective [1]. However, there are a number of issues associated with using adaptively trained systems for speech recognition, or inference. One problem is that the adaptively trained system cannot be directly used in unsupervised inference. To use the canonical model for inference, a target domain transform is required. For unsupervised inference, the hypothesis to generate this transform is not available. One approach to handle this problem is to use a multi-style model, e.g. a speaker independent model, to generate an initial hypothesis of the test data. Target domain transforms are then estimated using the ML criterion with this initial hypothesis. Another problem with the traditional framework is that if there is only limited adaptation data, ML estimates of transforms are not reliable and may be overly "tuned" to the initial hypothesis.

These problems may be addressed by interpreting adaptive training and inference in a Bayesian framework [4]. Here, the parameters of the system are treated as random variables. The likelihood of the observation sequence is then obtained by marginalising out over the parameter distributions. Though this approach may be applied to both transform and model parameters, in this paper only transform parameters are considered as random variables. This is because by 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 "sufficient data" assumption is good given the appropriate complexity. With this assumption, the standard point estimates used in adaptive training can be justified [4]. In contrast to standard adaptive training, a transform *prior distribution* is obtained during Bayesian adaptive training in addition to the standard canonical model estimate. During adaptive inference, as it is often not possible to control the amount of the adaptation data, the "sufficient data" assumption may be poor. Hence, the standard adaptation scheme with point estimate of transforms may not work in the limited data case. Rather than using the standard "adaptation-recognition" process, in Bayesian adaptive inference an integrated scheme is adopted. The task is to calculate the marginal likelihood of each possible hypothesis by integrating out over the transform distribution associated with each distinct hypothesis. This allows the canonical model to be directly used in unsupervised mode inference and avoids the over-tuning to the initial hypothesis. Furthermore, the use of Bayesian approaches in inference effectively handles the limited adaptation data problem due to the incorporation of the transform distribution. Note, in this work, the point estimates are used for the canonical model. Though discussed from an ML perspective, this Bayesian adaptive inference framework can also be extended to discriminative criteria.

The marginalisation integral over the transform distribution is intractable due to the presence of the latent variables associated with HMM. Two classes of approximations for this integral are investigated in this paper. The first class uses a *lower bound* to approximate the intractable marginal likelihood in inference. An iterative process is used to make this lower bound as tight as possible to the marginal likelihood. Point estimates of transforms, such as maximum a posteriori

(MAP) [5] and ML [6], sit within this class. Variational Bayes (VB) [7] is another lower bound based Bayesian approximation approach. In VB a distribution over the parameters, rather than a point estimate is used. This should lead to more robust recognition performance than the point estimates. VB has previously been applied to train distributions over HMM model parameters [8]. As an application to simple adaptation, VB was also used in [9] to train distributions of a mean bias vector and a scaling factor in supervised adaptation on an isolated words recognition task. However, in contrast to this work, the VB approaches in [8] and [9] were not consistent between training and inference. Instead, an approximate approach, the *frame-independent* (FI) assumption, was used in inference. This approach belongs to the second class of approximation approaches discussed in this paper. Approaches in this class do not involve an iterative process and approximate the marginal likelihood directly. Hence they are referred to as *direct* approximations. Sampling approaches are one form of direct approximations [10]. The FI assumption has previously been investigated for adaptation and also referred to as Bayesian predictive adaptation [11], [12], [13]. Though a distribution over the transform parameters, rather than a point estimate, is used, the transform is allowed to effectively change from frame to frame, possibly limiting performance gains. This paper examines both *lower bound* and *direct* approaches. Both incremental [14] and batch modes [4] Bayesian adaptive inference are discussed. These general Bayesian approximations are then applied to a specific transform: maximum likelihood linear regression (MLLR) [6].

This paper is arranged as follows. Section II describes adaptive training and inference within a Bayesian framework. Section III discusses various approximation approaches to calculate the intractable marginal likelihood. Incremental inference is then described in section IV. Section V applies the approximations to MLLR. Experiments on a conversational telephone speech task, for both ML and discriminative models are shown in section VI.

II. A BAYESIAN FRAMEWORK FOR ADAPTIVE TRAINING AND ADAPTIVE INFERENCE

Adaptive training has become a popular technique to build systems on non-homogeneous training data. It is normally described in a ML framework. This section describes adaptive training and inference from a Bayesian perspective.

A. Bayesian Adaptive Training

In adaptive training, two sets of parameters are used to model the audio signal variabilities. A set of transforms is used to represent non-speech variabilities for each homogeneous data block and a canonical model is used to represent the speech variability. First the training data is partitioned into S blocks, $\mathcal{O} = \{\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(S)}\}$, where $\mathbf{O}^{(s)}$ represents a homogeneous block associated with a particular acoustic condition s. Treating the two sets of parameters as random variables, the marginal likelihood can be expressed as

$$p(\mathcal{O}|\mathbf{H}) = \int_{\mathcal{M}} p(\mathcal{O}|\mathbf{H}, \mathcal{M}) p(\mathcal{M}|\Phi) \, d\mathcal{M} \qquad (1)$$

where

$$p(\mathcal{O}|\mathbf{H}, \mathcal{M}) = \prod_{s=1}^{S} \int_{\mathcal{T}} p(\mathbf{O}^{(s)}|\mathcal{H}^{(s)}, \mathcal{M}, \mathcal{T}) p(\mathcal{T}|\phi) \, d\mathcal{T} \qquad (2)$$

$p(\mathcal{M}|\Phi)$ and $p(\mathcal{T}|\phi)$[1] are the prior distributions for the canonical model $\mathcal{M}$ and transform parameters $\mathcal{T}$ respectively, Φ and ϕ are hyper-parameters of the prior distributions, $\mathbf{H} = \{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(S)}\}$ is the transcription sequence, where $\mathcal{H}^{(s)}$ is the transcription for homogeneous block s. Hidden Markov models (HMMs), with Gaussian mixture model (GMM) as the state output distributions, are used as the underlying acoustic model. Thus

$$p(\mathbf{O}^{(s)}|\mathcal{H}^{(s)}, \mathcal{M}, \mathcal{T}) = \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathcal{M}, \mathcal{H}^{(s)}) \prod_{t} p(\mathbf{o}_t|\mathcal{M}, \mathcal{T}, \theta_t) \quad (3)$$

where $\boldsymbol{\theta}$ is the hidden Gaussian component sequence for $\mathcal{H}^{(s)}$, $P(\boldsymbol{\theta}|\mathcal{M})$ is the distribution of a particular sequence $\boldsymbol{\theta}$, $p(\mathbf{o}_t|\mathcal{M}, \mathcal{T}, \theta_t)$ is the Gaussian distribution at component θ_t, $\mathbf{o}_t$ is the observation vector at time t.

Adaptive training may be viewed as modifying the dynamic Bayesian network (DBN) associated with the acoustic model. Figure 1 shows the comparison between a standard HMM and an adaptive HMM. For HMMs, figure 1(a), the
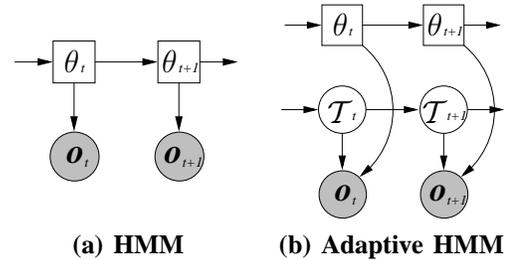

(a) HMM   **(b) Adaptive HMM**

Fig. 1. Dynamic Bayesian network comparison between standard HMM and adaptive HMM

observations are conditionally independent given the hidden variables. In contrast, figure 1(b) shows the DBN for an adaptive HMM. Here an additional level of dependency is introduced, observations are also dependent on a transform. Within a homogeneous data block, the transform is assumed to be unchanged, thus, $\mathcal{T}_t = \mathcal{T}_{t+1}$. The DBNs given in figure 1 can be used in various ways for training and inference. Standard multi-style training and decoding is an example of using the HMM DBN in both stages. It is also possible to use the HMM DBN in training and the adaptive HMM DBN in inference. This is similar to performing adaptation on multi-style trained models. If the adaptive HMM DBN is used in training, a canonical model representing the speech variability is estimated given a set of transforms. Thus, the adaptive HMM DBN must be used during inference. The effect of different ways of using the DBNs for training and inference will be illustrated in the experiments.

There is normally no prior model or transform information available before training. Therefore, the prior distributions

---

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

of the two sets of parameters must be estimated using the training data. Two issues need to be considered. First is the form of the prior distribution. A preferable choice is to use a *conjugate prior* to the likelihood of the complete data set when performing expectation maximisation (EM) algorithm [7]. This may result in tractable mathematical formulae. For example, for mean-based transform such as MLLR [6], a Gaussian distribution over the transform parameters is the conjugate prior to the complete data set [15][2]. The second issue is the estimation of the hyper-parameters, once the prior form is determined. They may be estimated using the *empirical Bayes* approach [17], [18]. The basic idea is to maximise the marginal likelihood in equations (1) and (2) with respect to the hyper-parameters of both priors. Directly optimising these equations is highly complex due to the existence of hidden variables. Lower bounds may be introduced to make the optimisation feasible. For the canonical model prior, introducing a variational distribution $q(\mathcal{M})$ and applying Jensen's inequality yields a lower bound of equation (1)

$$\log p(\mathcal{O}|\mathbf{H}) \geq \left\langle \log \frac{p(\mathcal{O}|\mathbf{H}, \mathcal{M})p(\mathcal{M}|\Phi)}{q(\mathcal{M})} \right\rangle_{q(\mathcal{M})} \quad (4)$$

$$= \langle \log p(\mathcal{O}|\mathbf{H}, \mathcal{M}) \rangle_{q(\mathcal{M})} - \texttt{KL}\left(q(\mathcal{M})||p(\mathcal{M}|\Phi)\right) \quad (5)$$

where $< f(x) >_{q(x)} = \int_x f(x)q(x)\, dx$ denotes the expectation of function $f(x)$ with respect to the distribution of $q(x)$, $\texttt{KL}(q(x)||p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)}\, dx$ is the Kullback Leibler (KL) distance of two distributions. The above becomes equality when

$$q(\mathcal{M}) = p(\mathcal{M}|\mathcal{O}, \mathbf{H}, \Phi) \quad (6)$$

The $\texttt{KL}$ distance is always positive unless the two distributions are the same, in which case the distance is zero. Therefore, from equations (5) and (6), the optimal canonical model prior is obtained by choosing it to have the same functional form and hyper-parameters as the posterior, as shown below

$$p(\mathcal{M}|\Phi) = p(\mathcal{M}|\mathcal{O}, \mathbf{H}, \Phi) \quad (7)$$

Note that equation (7) is only possible if a conjugate prior to the likelihood $p(\mathcal{O}|\mathbf{H})$ exists [3]. Calculating the canonical model posterior $p(\mathcal{M}|\mathcal{O}, \mathbf{H}, \Phi)$ is still complex. This issue will be addressed later.

The estimation of the transform prior is complicated due to the homogeneity constraint. A separate variational transform distribution is required for *each* homogeneous block s. Applying Jensen's inequality to equation (2) yields

$$\log p(\mathcal{O}|\mathbf{H}, \mathcal{M}) \geq \sum_{s=1}^{S} \left\langle \log p(\mathbf{O}^{(s)}|\mathcal{H}^{(s)}, \mathcal{M}, \mathcal{T}) \right\rangle_{q^{(s)}(\mathcal{T})}$$
$$- \sum_{s=1}^{S} \texttt{KL}\left(q^{(s)}(\mathcal{T})||p(\mathcal{T}|\phi)\right) \quad (8)$$

where equality is achieved when for each block

$$q^{(s)}(\mathcal{T}) = p(\mathcal{T}|\mathbf{O}^{(s)}, \mathcal{H}^{(s)}, \mathcal{M}, \phi) \quad (9)$$

---

[2]For discussion about mixture priors, refer to [16].

[3]In the general case, where a conjugate prior does not exist, it is not possible to set the KL divergence to zero in the lower bound (5). Optimizing the bound is still valid, however the optimum will not satisfy equation (7).

As there are S transform posterior distributions, the KL distance in equation (8) can not be simply minimised by setting $p(\mathcal{T}|\phi)$ equal to the posterior distributions as in equation (7).

When building speech recognition systems, it is possible to control the complexity of the system being trained so that each Gaussian component and transform have "sufficient data". For example minimum occupancies may be used during the construction of decision tree to ensure robust canonical model estimates and transforms may be shared among groups of Gaussian components. With these complexity control schemes, it is reasonable to assume that the variances of the parameter posterior distributions are sufficiently small that they can be approximated by a Dirac delta function. Hence

$$p(\mathcal{M}|\mathcal{O}, \mathbf{H}, \Phi) \approx \delta(\mathcal{M} - \hat{\mathcal{M}}) \quad (10)$$

$$p(\mathcal{T}|\mathbf{O}^{(s)}, \mathcal{H}^{(s)}, \mathcal{M}, \phi) \approx \delta(\mathcal{T} - \hat{\mathcal{T}}^{(s)}) \quad (11)$$

where $\hat{\mathcal{M}}$ and $\hat{\mathcal{T}}^{(s)}$ are point estimates of the two sets of parameters. Considering equations (7) and (10) and using them in equation (5), $\hat{\mathcal{M}}$ is the ML estimate given the sufficient data assumption. Similarly, $\hat{\mathcal{T}}^{(s)}$ is also the ML estimate. Hence, the canonical model prior is a Dirac delta function with the ML estimate as the mode. Using equation (11) in equation (8), it can be shown that the hyper-parameters of the transform prior can be estimated by [16]

$$\hat{\phi} = \arg\max_{\phi} \sum_{s=1}^{S} \log p(\hat{\mathcal{T}}^{(s)}|\phi) \quad (12)$$

To summarise, *given sufficient training data*, Bayesian adaptive training yields an ML estimate of canonical model and a non-point transform prior distribution. The training involves the following steps:

1) Interleave ML update of the canonical model and the transforms for each homogeneous block. This is the same procedure as the standard ML adaptive training [1], [3].
2) Treat each transform as a sample in the parametric space and find an ML estimate of the hyper-parameters ϕ of the transform prior distribution using equation (12).

By interpreting adaptive training from the Bayesian perspective, the standard ML estimate of canonical model may be justified. In addition, a non-point transform prior distribution is motivated, which is important for Bayesian adaptive inference. It is worth emphasising that the transform prior distribution is dependent on the particular canonical model set used.

B. Bayesian Adaptive Inference

Once the canonical model and the transform prior distributions are estimated during training, they can be used together for inference. For adaptively trained systems, due to the homogeneity constraint, the inference must be performed at the homogeneous block level. For each block,

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} p(\mathcal{H}|\mathbf{O}) = \arg\max_{\mathcal{H}} p(\mathbf{O}|\mathcal{H})P(\mathcal{H}) \quad (13)$$

where $\hat{\mathcal{H}}$ is the inferred hypothesis, $\mathbf{O}$ is the observation sequence of a particular homogeneous block, $p(\mathbf{O}|\mathcal{H})$ and $P(\mathcal{H})$ are acoustic and language model scores of each hypothesis

respectively. $P(\mathcal{H})$ may be obtained from an N-gram language model. The key problem here is to calculate the marginal likelihood

$$p(\mathbf{O}|\mathcal{H}) = \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H}, \hat{\mathcal{M}}, \mathcal{T}) p(\mathcal{T}|\phi) \, d\mathcal{T} \qquad (14)$$

This process is referred to as *Bayesian adaptive inference*. The point estimate of the canonical model, $\hat{\mathcal{M}}$, is used for inference because marginalisation over a Dirac delta function will result in a likelihood given the mode of that Dirac delta function. In unsupervised inference, where no supervision data is available, equation (14) allows the canonical model to be directly used for inference. In supervised mode, $p(\mathcal{T}|\phi)$ may be updated to posterior distribution for inference, which is referred to as *posterior adaptation* [3]. In this paper, supervised mode will not be further discussed as there is no supervision data available for the tasks considered.

In recognition with standard HMMs, the *Viterbi* algorithm [19] is usually used to efficiently calculate the likelihood of observation sequence. This relies on the conditional independence assumption of HMMs to make the inference efficient. However, this conditional independence assumption is not valid for adaptive HMMs due to the additional dependence on the transform. Hence, the Viterbi algorithm is not suitable for Bayesian adaptive inference. Instead, *N-Best rescoring* [20] is used in this work to reflect the nature of adaptive HMM. Though the N-Best rescoring may limit the performance gain, and loss, due to the limited number of candidate hypothesis sequences, given sufficient hypothesis candidates, this N-Best list is likely to contain the "best" hypothesis. In N-Best rescoring, marginal likelihood of every possible hypothesis, $p(\mathbf{O}|\mathcal{H})$ is separately calculated. Due to the coupling of transform parameters and hidden state/component sequence, the Bayesian integral in equation (14) is intractable. Approximations are required to calculate the marginal likelihood $p(\mathbf{O}|\mathcal{H})$. Various approaches will be discussed in section III.

Note that the Bayesian adaptive inference process is an integrated process. There is no distinct of "adaptation" and "recognition" stage as in standard decoding process. The standard process is a special case of the integrated Bayesian inference process. This is discussed in section III-A. In contrast to some previously investigated *Bayesian predictive adaptation* (BPA) approaches [11], [21], Bayesian adaptive inference strictly deals with the Bayesian integral over the *whole* observation sequence, while the BPA approaches implicitly assume the Bayesian integral is performed at every time instance. This will be discussed in detail in section III-B.

The Bayesian framework described before is based on the likelihood criterion. To obtain state-of-the-art performance, the discriminative criterion is often used [22]. Discriminative adaptive training and inference can also be interpreted from the Bayesian perspective [16]. In this paper, the training procedure adopted is to only discriminatively update the canonical model given the ML estimated transforms. Minimum phone error (MPE) is used as the discriminative criterion to train the canonical models [22]. Hyper-parameters of the transform

prior distribution are estimated from the ML transforms for the discriminative canonical model. This transform prior distribution is used in Bayesian inference as discussed before. It is worth noting that the transform prior is calculated from ML transforms and is applied in a non-discriminative way in inference. This may limit the possible gains of adaptive training when using the discriminative criterion.

III. APPROXIMATE INFERENCE SCHEMES

The marginal likelihood calculation in equation (14) is generally intractable, hence approximations are required. Bayesian adaptive inference procedure will become:

1) Calculate the approximate value $\mathcal{L}(\mathbf{O}|\mathcal{H})$ for $p(\mathbf{O}|\mathcal{H})$ in equation (14).
2) Use $\mathcal{L}(\mathbf{O}|\mathcal{H})$ instead of $p(\mathbf{O}|\mathcal{H})$ in equation (13) to find the best hypothesis.

In this section, two main categories of approximation approaches are described [4]. One set of approaches iteratively tighten a lower bound to the real integral. These are referred to as *lower bound* approximations. The second set directly approximates the integral, referred to as *direct* approximations.

A. Lower Bound Approximations

As described in section II, a lower bound may be constructed to approximate the marginal likelihood in equations (1) and (2). The same approach may be used for inference. Introducing a joint distribution, $q(\boldsymbol{\theta}, \mathcal{T})$, over the component sequence, $\boldsymbol{\theta}$, and transform parameters, $\mathcal{T}$ and applying Jensen's inequality yields a lower bound $\mathcal{L}(\mathbf{O}|\mathcal{H})$ as below [4]

$$\log p(\mathbf{O}|\mathcal{H}) \geq \mathcal{L}(\mathbf{O}|\mathcal{H}) = \left\langle \log \frac{p(\mathbf{O}, \boldsymbol{\theta}|\mathcal{T}, \mathcal{H})p(\mathcal{T})}{q(\boldsymbol{\theta}, \mathcal{T})} \right\rangle_{q(\boldsymbol{\theta}, \mathcal{T})} \quad (15)$$

where $p(\mathcal{T})$ is the brief notation for the transform prior distribution, $p(\mathcal{T}|\phi)$, and will be used in the rest of this paper. The above becomes an equality when

$$q(\boldsymbol{\theta}, \mathcal{T}) = p(\boldsymbol{\theta}, \mathcal{T}|\mathbf{O}, \mathcal{H}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \mathcal{T})p(\mathcal{T}|\mathbf{O}, \mathcal{H}) \quad (16)$$

Using equation (16) is impractical because the calculation of the transform posterior $p(\mathcal{T}|\mathbf{O}, \mathcal{H})$ requires the marginal likelihood to be calculated. Tractable variational distributions for $P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \mathcal{T})$ and $p(\mathcal{T}|\mathbf{O}, \mathcal{H})$ are described in this section. An iterative *learning* process is then used to update these variational distributions to make the lower bound as tight as possible. The tightness of the bound is dependent on the form of the variational distributions, point estimate or variational Bayes, and the number of iterations.

When using lower bound approximations for inference, there is an assumption that the rank ordering of the real inference evidence in equation (13) is similar to the ordering of the evidence in which the lower bound value is used instead of the log likelihood, i.e.

$$\mathcal{L}(\mathbf{O}|\mathcal{H}_1) + \log P(\mathcal{H}_1) > \mathcal{L}(\mathbf{O}|\mathcal{H}_2) + \log P(\mathcal{H}_2)$$
$$\Rightarrow \log p(\mathbf{O}|\mathcal{H}_1) + \log P(\mathcal{H}_1) > \log p(\mathbf{O}|\mathcal{H}_2) + \log P(\mathcal{H}_2)$$

---

[4]For clarity of notation, the block index s and the notation of the canonical model set $\mathcal{M}$ are dropped.

How good this assumption is will depend on the forms of the lower bound. Generally, it is important to get a tight lower bound for $p(\mathbf{O}|\mathcal{H})$. In order to achieve this, it is necessary to optimise the lower bound with respect to *every possible* hypothesis respectively, which is similar to N-Best supervision [23]. In contrast to the work in [23] where no theoretical justification was proposed, the work here motivates it from a viewpoint of tightening the lower bound during adaptive inference. It is also interesting to compare N-Best supervision to the standard 1-Best supervision adaptation approaches such as iterative MLLR [24]. In iterative MLLR, a transform is estimated using the 1-Best hypothesis of the test data as supervision. This transform is then used to calculate inference evidence for *all* possible hypothesis and the process is repeated if necessary. 1-Best supervision will lead to a "tight" lower bound for the "best" hypothesis. However for the other competing hypotheses, the lower bounds will not be as tight as they could be. This biases those hypotheses to the 1-Best hypothesis and may significantly affect the performance, especially for complex transforms or short sentences as shown in section VI. A number of other schemes have previously been proposed to address the 1-Best bias problem. Two such schemes are lattice MLLR [25] and confidence MLLR [26]. In contrast to the N-best supervision framework, these schemes do not directly address the problem, but rather use some form of measure of the confidence of a particular transcription. The disadvantage of these approaches is that some form of sentence posterior, or confidence score, is required. These scores are hard to reliably obtain from a speech recognition system and require the use of techniques such as acoustic deweighting [25]. These confidence-based schemes are computationally efficient compared to the N-best supervision framework. However, it is felt that the strict mathematical framework of the Bayesian adaptive inference approach offers a more flexible scheme for future development. Furthermore, it is worth emphasising that the estimate of lattice MLLR or confidence MLLR may still be unreliable when there is only very limited data because the ML criterion is still used in transform estimation.

Two forms of lower bound approximations are described in this paper.

1) Point Estimates: In the same fashion as ML adaptive training, given sufficient data, a Dirac delta function may be used as the transform posterior resulting in a point version of equation (16)

$$q(\boldsymbol{\theta}, \mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H}, \mathcal{T})\delta(\mathcal{T} - \hat{\mathcal{T}}) \tag{17}$$

where $\hat{\mathcal{T}}$ is a point estimate of transform for the target domain. Equation (15) may then be re-expressed as

$$\log p(\mathbf{O}|\mathcal{H}) \geq \log p(\mathbf{O}|\mathcal{H}, \hat{\mathcal{T}}) + \log p(\hat{\mathcal{T}}) + \mathtt{H}(\delta(\mathcal{T} - \hat{\mathcal{T}})) \tag{18}$$

where $\mathtt{H}(p(x))$ is the entropy of $p(x)$. For all point estimates of $\hat{\mathcal{T}}$, the entropy of the Dirac delta function is the same $-\infty$ [27]. As $\mathtt{H}(\delta(\mathcal{T} - \hat{\mathcal{T}}))$ is a negative constant with infinite value, it can be ignored without affecting the rank ordering of the lower bound. The rank ordering of the lower bound is then determined by

$$\mathcal{K}_{\mathtt{MAP}}(\hat{\mathcal{T}}) = \log p(\mathbf{O}|\mathcal{H}, \hat{\mathcal{T}}) + \log p(\hat{\mathcal{T}}) \tag{19}$$

Equation (19) yields a maximum a posteriori (MAP) estimate. In contrast to the standard MAP linear regression (MAPLR) [5] approach, in the N-Best supervision framework, a distinct MAP estimate is required for every possible hypothesis and the transform prior term $\log p(\hat{\mathcal{T}})$ must be considered in inference. The EM algorithm may be used to optimise $\mathcal{K}_{\mathtt{MAP}}(\hat{\mathcal{T}})$. If a single component prior distribution is used, the transform update formulae are similar to the MAPLR [5]. A mixture prior can also be used as discussed in [4]. The MAP estimate is the same as the standard ML estimate if a non-informative prior is used. In this case, the prior term in equation (19) disappears and the likelihood of the observation sequence given the ML estimate can be directly used in inference. Therefore, the standard ML estimate of transforms is one case of the lower bound approximations within the Bayesian framework. Note that, the ML estimate described here naturally requires N-Best supervision to tighten the lower bound as discussed before. In contrast, the widely used standard ML adaptation approach not only uses an ML estimate of transform, but also adopts a 1-Best supervision paradigm when estimating the ML transform. Hence, the standard adaptation approach has two levels of approximations and is a special case of Bayesian adaptive inference.

2) Variational Bayes: The use of Dirac delta distribution is only reasonable given sufficient adaptation data. For limited data, this assumption will be poor, possibly affecting the approximation quality. In order to make the lower bound tighter, another form of approximation approach, *variational Bayes* (VB), may be used [16]. Here, the distributions of the component sequence posterior $q(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H})$ and the transform posterior $q(\mathcal{T}|\mathbf{O}, \mathcal{H})$ are assumed to be conditionally independent. Thus

$$q(\boldsymbol{\theta}, \mathcal{T}) = q(\boldsymbol{\theta}|\mathbf{O}, \mathcal{H})q(\mathcal{T}|\mathbf{O}, \mathcal{H}) \tag{20}$$

This assumption is necessary to obtain a tractable mathematical form. For simplicity of notation the two posteriors will be denoted as $q(\boldsymbol{\theta})$ and $q(\mathcal{T})$. The lower bound in equation (15) can be re-written as an auxiliary function. At the $(k+1)^{th}$ iteration, this may be expressed as

$$\mathcal{Q}\left(q_{k+1}(\boldsymbol{\theta}), q_k(\mathcal{T})\right) = \langle \log p(\mathbf{O}, \boldsymbol{\theta}|\mathcal{T}, \mathcal{H}) \rangle_{q_{k+1}(\boldsymbol{\theta})q_k(\mathcal{T})}$$
$$+ \mathtt{H}(\log q_{k+1}(\boldsymbol{\theta})) - \mathtt{KL}(q_k(\mathcal{T})||p(\mathcal{T})) \quad (21)$$

where $q_k(\boldsymbol{\theta})$ and $q_k(\mathcal{T})$ are the variational component sequence and transform posterior distributions at the kth iteration respectively. The aim is now to obtain forms of $q(\boldsymbol{\theta})$ and $q(\mathcal{T})$ that maximise this auxiliary function, thus making the lower bound as tight as possible.

Taking the functional derivatives of the auxiliary function in equation (21) with respect to $q(\boldsymbol{\theta})$ and $q(\mathcal{T})$ respectively, an EM-like algorithm can be obtained, referred to as *Variational Bayesian EM* (VBEM) [7]. VBEM is guaranteed not to decrease the bound at each iteration. The process is:

1) Initialise: $q_0(\mathcal{T}) = p(\mathcal{T})$, $k = 1$.
2) VB Expectation (VBE): The optimal variational posterior component sequence distribution can be shown as

$$q_k(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\theta}(\mathbf{O}, \mathcal{H})} \exp\left(\langle \log p(\mathbf{O}, \boldsymbol{\theta}|\mathcal{T}, \mathcal{H}) \rangle_{q_{k-1}(\mathcal{T})}\right) \quad (22)$$

where $\mathcal{Z}_\theta(\mathbf{O}, \mathcal{H})$ is the normalisation term to make $q_k(\boldsymbol{\theta})$ a valid distribution. As $\log p(\mathbf{O}, \boldsymbol{\theta} | \mathcal{T}, \mathcal{H})$ can be factorised at the frame-level, the expectation with respect to $q_k(\mathcal{T})$ can be performed at the 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 [5]

$$\tilde{p}(\mathbf{o}_t | \theta_t) = \exp\left(< \log p(\mathbf{o}_t | \mathcal{T}, \theta_t) >_{q_{k-1}(\mathcal{T})}\right) \quad (23)$$

$\tilde{p}(\mathbf{o} | \theta)$ is referred to as a *pseudo-distribution* [4] because it is not necessarily normalised to be a valid distribution. $\mathcal{Z}_\theta(\mathbf{O}, \mathcal{H})$ can be simply calculated using the forward algorithm with $\tilde{p}(\mathbf{o}_t | \theta_t)$,

$$\mathcal{Z}_\theta(\mathbf{O}, \mathcal{H}) = \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta} | \mathcal{M}) \prod_t \tilde{p}(\mathbf{o}_t | \theta_t) \quad (24)$$

3) VB Maximisation (VBM): Given the variational component sequence posterior, the optimal $q_k(\mathcal{T})$ can be found

$$q_k(\mathcal{T}) = \frac{1}{\mathcal{Z}_\mathcal{T}(\mathbf{O}, \mathcal{H})} p(\mathcal{T}) \exp\left(\langle \log p(\mathbf{O}, \boldsymbol{\theta} | \mathcal{T}, \mathcal{H}) \rangle_{q_k(\boldsymbol{\theta})}\right) \quad (25)$$

where $\mathcal{Z}_\mathcal{T}(\mathbf{O}, \mathcal{H})$ is the normalisation used to make $q_k(\mathcal{T})$ a valid distribution. When using a conjugate prior $p(\mathcal{T})$, the estimation of $q(\mathcal{T})$ only requires updating the hyper-parameters of the prior $p(\mathcal{T})$. The exact form will be discussed in section V.

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

Having obtained the final transform distribution after K iterations, $q(\mathcal{T}) = q_K(\mathcal{T})$, the value of the lower bound in equation (15) is required for inference. By calculating $q_{K+1}(\boldsymbol{\theta})$ based on $q_K(\mathcal{T})$ using equation (22) and using it in equation (21), the lower bound can be re-expressed as [16]

$$\mathcal{L}(\mathbf{O} | \mathcal{H}) = \log \mathcal{Z}_\theta(\mathbf{O}, \mathcal{H}) - \text{KL}\left(q_K(\mathcal{T}) || p(\mathcal{T})\right) \quad (26)$$

where $\mathcal{Z}_\theta(\mathbf{O}, \mathcal{H})$ is given in equation (24), which can be regarded as the "likelihood" based on the pseudo-distribution. The KL distance will have a closed-form solution if the form of transform distribution is appropriately chosen as discussed in section V.

The above derivations are based on a single transform for all Gaussian components. It can be extended to multiple base-class case, where an independent transform is used for a group of Gaussian components. The resultant VBEM algorithm is similar to the global case except that the sufficient statistics for each variational transform distribution are accumulated based on the corresponding group of Gaussians [16].

The steps for lower bound based inference are summarised as below:

1) Initialisation. Set initial transform $\hat{\mathcal{T}}_0^{\text{ML}}$, $\hat{\mathcal{T}}_0^{\text{MAP}}$ or transform distribution $q_0(\mathcal{T})$.
2) Iteratively update $\hat{\mathcal{T}}$ or $q(\mathcal{T})$ to tighten the lower bound. In the ML approximation, $\hat{\mathcal{T}}_K^{\text{ML}}$ is obtained by

---

maximising $\log p(\mathbf{O} | \mathcal{H}, \mathcal{T})$, where K is the number of iterations. In the MAP approximation, $\hat{\mathcal{T}}_K^{\text{MAP}}$ is obtained by maximising equation (19). In the VB approximation, the variational distribution $q_K(\mathcal{T})$ is obtained by maximising equation (21). Note that the transforms (distributions) are specifically estimated for each possible hypothesis.

3) Calculate the lower bound value for each hypothesis using the final transform distribution respectively. The ML lower bound value is $\log p(\mathbf{O} | \mathcal{H}, \hat{\mathcal{T}}_K^{\text{ML}})$, The MAP lower bound is equation (19) with $\hat{\mathcal{T}}_K^{\text{MAP}}$. The VB lower bound is calculated using equation (26) with $q_K(\mathcal{T})$.
4) The lower bound value is then used instead of $\log p(\mathbf{O} | \mathcal{H})$ in equation (14) for inference.

B. Direct Approximations

There are a number of approaches to approximate the likelihood integral, which do not require an iterative process to tighten the lower bound. These forms of approximation will be referred to as *direct approximations*. In contrast to the lower bound approximations, direct approximations may be greater or less than the likelihood.

Sampling approaches are a standard method for directly approximating intractable probabilistic integrals. The basic idea is to draw samples from the distribution and use the average integral function value to approximate the real probabilistic expectation [10]. As the number of transform parameters increases the number of samples required to obtain good estimates dramatically increases. As it is hard to efficiently control the computational cost, this approach is only applicable to systems with small number of adaptation parameters, for example, cluster adaptive training [4].

An alternative approach is to modify the DBN of the adaptive HMM associated with the inference process. One simple approach is to allow the transforms to change at each time instance. Figure 2(a) shows the DBN of the adaptive

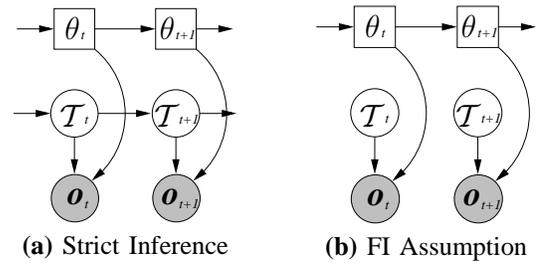**(a)** Strict Inference   **(b)** FI Assumption

Fig. 2. Dynamic Bayesian network comparison between strict inference and the frame-independent assumption

HMM, where the transform parameters are constrained to be constant over all frames within one homogeneous block. This yields the integral in equation (14).

If the constraint on transform transitions is relaxed, the DBN in figure 2(b) is obtained. This allows the transform to vary from one time instance to another and will be referred to as the *frame-independent* assumption. This assumption has been implicitly used in the Bayesian prediction approaches for HMM parameters, where the resultant distribution is called

---

[5] The transform $\mathcal{T}$ is assumed to only affect the Gaussian mixture parameters.

Bayesian predictive distribution [28]. In [8] and [9], this approach was used as the inference scheme for parameter distribution trained using VB approach. The assumption has been also investigated for Bayesian adaptation [15], [12], [11]. Using this approximation in equation (3) yields

$$p(\mathbf{O}|\mathcal{H}) \approx \mathcal{L}(\mathbf{O}|\mathcal{H}) = \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathcal{M},\mathcal{H}) \prod_t \bar{p}(\mathbf{o}_t|\theta_t) \quad (27)$$

where

$$\bar{p}(\mathbf{o}_t|\theta_t) = \int_{\mathcal{T}} p(\mathbf{o}_t|\mathcal{T},\theta_t) p(\mathcal{T}) \, d\mathcal{T} \quad (28)$$

is the Bayesian predictive distribution at θ_t. With an appropriate form of $p(\mathcal{T})$, this frame-level integral is tractable. For example, in MLLR adaptation, a single Gaussian distribution or GMM may be used as the transform prior to obtain a tractable predictive distribution[15], [12]. In strict adaptive inference, maintaining the constraint on transform transition within homogeneous blocks will be exponentially expensive with the increase of the size of the blocks. One advantage of using the FI approximation is that the additional computational cost compared to decoding with standard HMMs is small. With this approximation, no iterative estimation scheme is required and Viterbi decoding may be used. However, it breaks the homogeneity causality of the adaptive HMM. When using a single Gaussian prior distribution, the FI approximation is similar to the multi-style training approach, where the acoustic condition can usually change from frame to frame (the standard HMM assumption) [3]. Unless the posterior distributions of each homogeneous block or a multiple component prior are used, the results with FI approximation will be similar to the multi-style system performance.

IV. INCREMENTAL BAYESIAN ADAPTIVE INFERENCE

Bayesian adaptive inference has been described in a *batch* mode where all test data are available for decoding in a single block. However, in some applications, test data becomes available gradually. *Incremental* inference is often used. This section discusses incremental adaptive inference within a Bayesian framework based on lower bound approximations [14]. Only variational Bayes is discussed here, the treatment of point estimates is similar.

For incremental adaptive inference, the homogeneous data block comprises multiple utterances which become available causally. $\mathbf{O} = \mathbf{O}_{1:u} = \{\mathbf{O}_1, \cdots, \mathbf{O}_u\}$ denotes the 1st to the uth utterances. Similarly, the hypothesis for all u utterances, $\mathcal{H}$, consists of a set of hypotheses, $\mathcal{H}_{1:u} = \{\mathcal{H}_1, \ldots, \mathcal{H}_u\}$. Information can be propagated to the uth utterance from the preceding $u-1$ utterances. The key questions are what information should be propagated between utterances and how to use this propagated information. Various forms of information propagation are discussed in the context of the VB approximation.

1. No information: The lower bound for all u utterances is optimised. This involves rescoring all u blocks, obtaining a new hypothesis $\hat{\mathcal{H}}_{1:u}$. The uth utterance may change the "best" hypothesis for the preceding utterances. This approach breaks the standard causal aspects of incremental adaptive inference.

As the transform is kept constant within each homogeneous block in strict adaptive inference, new data will cause a re-computation for all utterances. The computational cost then increases exponentially.

2. Inferred hypothesis sequence: If the causal constraint is enforced, then the best hypothesis for the previous $u-1$ utterances is fixed as $\hat{\mathcal{H}}_{1:u-1}$. The optimisation of the bound is then only based on possible hypotheses for the uth block. The variational distributions in equation (20) become

$$q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H}) = q(\boldsymbol{\theta}|\mathbf{O},\hat{\mathcal{H}}_{1:u-1},\mathcal{H}_u) \quad (29)$$

$$q(\mathcal{T}|\mathbf{O},\mathcal{H}) = q(\mathcal{T}|\mathbf{O},\hat{\mathcal{H}}_{1:u-1},\mathcal{H}_u) \quad (30)$$

In this configuration there is a choice of the initial transform distribution to use. The transform prior, $p(\mathcal{T})$, can be used to initialise the VBEM process. Alternatively, the distribution from the previous utterances may be used. Thus

$$q_0(\mathcal{T}|\mathbf{O},\hat{\mathcal{H}}_{1:u-1},\mathcal{H}_u) = q_K(\mathcal{T}|\mathbf{O}_{1:u-1},\hat{\mathcal{H}}_{1:u-1}) \quad (31)$$

where K is the number of VBEM iterations used. Inference only involves finding the hypothesis for the uth utterance.

3. Posterior sequence distribution and hypotheses: Propagating the inferred hypotheses still requires the corresponding posterior component sequence distribution for all u utterances to be computed. This posterior may also be fixed and propagated to the next utterance. Thus equation (29) becomes

$$q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H}) = q(\boldsymbol{\theta}_u|\mathbf{O}_u,\mathcal{H}_u) \prod_{i=1}^{u-1} q_K(\boldsymbol{\theta}_i|\mathbf{O}_i,\hat{\mathcal{H}}_i) \quad (32)$$

The previous $u-1$ utterances do not need to be re-aligned. Only $q(\boldsymbol{\theta}_u|\mathbf{O}_u,\mathcal{H}_u)$ needs to be computed, i.e., the sufficient statistics of the uth utterance need to be accumulated. This is the most efficient form. The standard incremental adaptation scheme uses a similar strategy, where the alignments of the previous utterances are fixed and the statistics propagated [29]. However, in the standard approach, only one transform is estimated for decoding the current utterance. In a Bayesian inference framework, a distinct transform is estimated for each possible hypothesis of the current utterance.

Using the information propagation strategy 3, an efficient, modified version of the VBEM algorithm can be derived [14]. With the point estimate approximations, a similar incremental EM algorithm and inference process can be derived [16]. The main difference is that point estimates of the transforms, rather than the distributions, are propagated.

V. APPLICATION TO MLLR

Maximum Likelihood Linear Regression (MLLR) is a widely used linear transform based approach in adaptive training, referred to as speaker adaptive training (SAT) [6]. In MLLR, the mean vectors of the Gaussian components are adapted by a linear transform. The adapted mean vector $\hat{\boldsymbol{\mu}}^{(m)}$ is expressed as

$$\hat{\boldsymbol{\mu}}^{(m)} = \mathbf{A}\boldsymbol{\mu}^{(m)} + \mathbf{b} = \mathbf{W}\boldsymbol{\xi}^{(m)} \quad (33)$$

where $\boldsymbol{\xi}^{(m)} = [\boldsymbol{\mu}^{(m)T} \quad 1]^T$ is the extended mean vector and $\mathbf{W} = [\mathbf{A} \ \mathbf{b}]$ is the extended linear transform.

A. Bayesian adaptive training for MLLR

Standard SAT is first performed resulting in a canonical HMM model and a set of transforms. This relies on the use of standard complexity control schemes to ensure the sufficiency of the training data. A transform prior distribution $p(\mathcal{T})$ is then estimated from these transforms using equation (12). For MLLR, a Gaussian distribution may be used as the conjugate prior to the likelihood of the complete dataset. In this case, each row of the transform is assumed to be independent given the prior component [15], [4]. Thus

$$p(\mathcal{T}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d; \boldsymbol{\mu}_{\mathbf{w}_d}, \boldsymbol{\Sigma}_{\mathbf{w}_d}) \qquad (34)$$

where the transform $\mathcal{T} = \mathbf{W}$, D is the size of the original mean vector, $\mathbf{w}_d^T$ is the dth row of $\mathbf{W}$. This row-independent assumption is consistent with the diagonal covariance matrices commonly used for HMM systems [15].

B. Bayesian adaptive inference for MLLR

Given the canonical model and the transform prior distribution, unsupervised Bayesian adaptive inference can be performed. The key problem is to calculate the approximate value for each possible hypothesis marginal likelihood $p(\mathbf{O}|\mathcal{H})$.

The first form discussed is a direct approximation. MLLR has too many parameters to use the sampling approach. Hence, only the frame-independent assumption approach is considered. For MLLR, the resultant predictive distribution in equation (28) is also a Gaussian distribution as derived in [15] and [12]. For the dth element, the mean and the variance values of the predictive distributions are:

$$\begin{aligned}
\bar{\mu}_d^{(m)} &= \boldsymbol{\mu}_{\mathbf{w}_d}^T \boldsymbol{\xi}^{(m)} \\
\bar{\sigma}_{dd}^{(m)} &= \sigma_{dd}^{(m)} + \boldsymbol{\xi}^{(m)T} \boldsymbol{\Sigma}_{\mathbf{w}_d} \boldsymbol{\xi}^{(m)}
\end{aligned}$$

where $\boldsymbol{\Sigma}^{(m)}$ is the diagonal covariance matrix of the canonical model, of which $\sigma_{dd}^{(m)}$ is the dth diagonal element. $\boldsymbol{\mu}_{\mathbf{w}_d}$ and $\boldsymbol{\Sigma}_{\mathbf{w}_d}$ are the mean and covariance of the dth row of the transform prior distribution in equation (34). With the predictive distribution, the approximate value for $p(\mathbf{O}|\mathcal{H})$ can be calculated using equation (27) and used for inference.

The second form considered are the lower bound approximations. A distinct transform or transform distribution is estimated for each possible hypothesis. The hypothesis itself is used as the supervision (the N-Best supervision scheme). The final transform or transform distribution is then used to calculate the lower bound value for inference as described in section II-B. The estimation formulae of transform or transform distribution are given below.

The ML estimate of transform is the standard MLLR, which was described in [6] and is not reproduced here. The final ML transform $\hat{\mathbf{W}}_K^{\text{ML}}$ (K is the iteration number) is used to calculate $\log p(\mathbf{O}|\mathcal{H}, \hat{\mathbf{W}}_K^{\text{ML}})$. MAP Linear Regression (MAPLR) with Gaussian prior was originally presented in [30]. Given

sufficient statistics

$$\begin{aligned}
\mathbf{G}_d &= \sum_t \sum_m \frac{\gamma_m(t)}{\sigma_{dd}^{(m)}} \boldsymbol{\xi}^{(m)} \boldsymbol{\xi}^{(m)T} \qquad (35) \\
\mathbf{k}_d &= \sum_t \sum_m \frac{\gamma_m(t) o_{t,d}}{\sigma_{dd}^{(m)}} \boldsymbol{\xi}^{(m)} \qquad (36)
\end{aligned}$$

where $\gamma_m(t)$ is the posterior occupancy of Gaussian component m at time t calculated using the forward-backward algorithm given the current hypothesis and transform estimate. The dth row of transform $\hat{\mathbf{W}}^{\text{MAP}}$ is estimated by

$$\hat{\mathbf{w}}_d^{\text{MAP}} = \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1} + \mathbf{G}_d\right)^{-1} \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1} \boldsymbol{\mu}_{\mathbf{w}_d} + \mathbf{k}_d\right) \qquad (37)$$

This estimate is iteratively updated. After K iterations, the final MAP transform $\hat{\mathbf{W}}_K^{\text{MAP}}$ is used to calculate $\log p(\mathbf{O}|\mathcal{H}, \hat{\mathbf{W}}_K^{\text{MAP}}) + \log p(\hat{\mathbf{W}}_K^{\text{MAP}})$ as the approximated value.

For the VB approximation, the pseudo-distribution is first required. This can be shown to be an unnormalised distribution, where component m has the form [16]

$$\begin{aligned}
\log \tilde{p}(\mathbf{o}|m) = & \log \mathcal{N}(\mathbf{o}; \tilde{\mathbf{W}}_\mu \boldsymbol{\xi}^{(m)}, \boldsymbol{\Sigma}^{(m)}) \\
& -\frac{1}{2} \sum_{d=1}^{D} \frac{\boldsymbol{\xi}^{(m)T} \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d} \boldsymbol{\xi}^{(m)}}{\sigma_{dd}^{(m)}} \qquad (38)
\end{aligned}$$

where $\tilde{\mathbf{W}}_\mu = [\tilde{\boldsymbol{\mu}}_{\mathbf{w}_1}, \cdots, \tilde{\boldsymbol{\mu}}_{\mathbf{w}_D}]^T$ is the mean of the variational transform posterior $q(\mathcal{T})$. This has the same functional form as the prior $p(\mathcal{T})$ in equation (34). Given the statistics calculated using the above pseudo-distribution, $q(\mathcal{T})$ can be updated. The mean and covariance matrix of the dth row of the variational transform posterior distribution can be shown to be

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d} &= \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1} + \mathbf{G}_d\right)^{-1} \\
\tilde{\boldsymbol{\mu}}_{\mathbf{w}_d} &= \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d} \left(\boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1} \boldsymbol{\mu}_{\mathbf{w}_d} + \mathbf{k}_d\right) \qquad (39)
\end{aligned}$$

where $\boldsymbol{\mu}_{\mathbf{w}_d}$ and $\boldsymbol{\Sigma}_{\mathbf{w}_d}$ are the parameters of the prior distribution, $\mathbf{G}_d$ and $\mathbf{k}_d$ have the same form as the standard statistics in equations (35) and (36) except that the component posterior, $\gamma_m(t)$, is calculated based on the pseudo-distribution with the current variational transform distribution. Once the final transform distribution has been estimated after K iteration, it can be used in equation (26) to calculate the VB lower bound for inference. As both $p(\mathcal{T})$ and $q_K(\mathcal{T})$ are Gaussian distributions, the KL distance in equation (26) has a closed-form solution given by

$$\text{KL}\left(q_K(\mathcal{T}) || p(\mathcal{T})\right) = -\frac{1}{2} \sum_{d=1}^{D} \Big(\log |\tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d} \boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1}|$$

$$+ \text{tr}\left(\mathbf{I} - (\tilde{\boldsymbol{\Sigma}}_{\mathbf{w}_d} + (\tilde{\boldsymbol{\mu}}_{\mathbf{w}_d} - \boldsymbol{\mu}_{\mathbf{w}_d})(\tilde{\boldsymbol{\mu}}_{\mathbf{w}_d} - \boldsymbol{\mu}_{\mathbf{w}_d})^T) \boldsymbol{\Sigma}_{\mathbf{w}_d}^{-1}\right) \Big) (40)$$

where $\text{tr}(\cdot)$ is the trace of a square matrix, $\mathbf{I}$ is an identity matrix.

VI. EXPERIMENTS

A. System setup

The performance of various forms of Bayesian inference approximations was evaluated on a large vocabulary conversational telephone speech task using MLLR to represent non-speech variabilities. The training dataset consists of 3

corpora recorded with slightly different acoustic conditions and collection framework. They are the LDC distributed Callhome English, Switchboard and Switchboard-Cellular corpora, consisting of 5446 speakers (2747 female, 2699 male), about 295 hours of data. The test dataset, `eval03` was taken from the NIST RT-03 Spring Evaluation. It has 144 speakers (77 female, 67 male), about 6 hours of data. All systems used a 12-dimensional PLP front-end with log energy and first, second and third derivatives. Cepstral mean and variance normalisation and VTLN were used. An HLDA transform was then applied to reduce the feature dimension to 39. A decision-tree state-clustered tri-phone model set with an average of 16 Gaussian components per state was constructed as the starting point for adaptive training. This is the baseline speaker-independent (SI) model. Initially ML training was performed to yield the *ML-SI* system. This was used as the starting point for all the other systems. The *MPE-SI* system was obtained using four iterations of MPE training [22]. The ML adaptively trained system, *ML-SAT*, was built using separate speech and silence MLLR transforms. Separate single Gaussian priors for these speech and silence transforms were independently estimated. For the discriminative adaptively trained system, *MPE-SAT*, the final transforms for the ML-SAT system were used, and 4 iterations of MPE training applied. Having trained the MPE-SAT model, transforms for each training speaker were again obtained using the ML criterion, and used to estimate the transform priors for the MPE-SAT model. Transform priors for the non-adaptively trained systems, ML-SI or MPE-SI, were obtained using a similar fashion.

As discussed previously, the Viterbi algorithm is not appropriate for Bayesian inference. In these experiments, N-Best rescoring was used for inference. 150-Best lists were generated for ML and MPE systems using the corresponding SI models. Though the use of N-best lists can limit performance difference, using spot-checks on the best VB configuration on the ML-SAT system with a 300-Best list showed little difference in performance.

B. Utterance Level Bayesian Adaptive Inference

To illustrate the effects of the Bayesian approximation approaches, homogeneous blocks were initially based on a *single* utterance, not as in the standard case on a side basis. For the `eval03` test set the average utterance length is 3.13 seconds, compared to the average side length of 153.75 seconds. This dramatically limits the available data and illustrates the issue of poor transform estimation with limited data. Table I shows the performance of Bayesian adaptive inference on the SI and the SAT systems. The baseline unadapted error rates of the ML-SI and MPE-SI systems are shown in the first line of the table and are 32.8% and 29.2% respectively.

For the FI approximation in table I, the performance of the ML-SAT system is similar to the baseline ML-SI system, which is expected as the FI approximation is similar to the multi-style training. However, the MPE-SAT system is about 0.5% worse than MPE-SI system. This degradation is because the transform prior for MPE-SAT system was estimated and applied for inference in a non-discriminative fashion. This

Bayesian	ML Train		MPE Train	
Approx	SI	SAT	SI	SAT
—	32.8	—	29.2	—
FI	—	32.9	—	29.7
ML	35.5	35.2	32.4	32.3
MAP	32.2	31.8	29.0	28.8
VB	31.8	31.5	28.8	28.6

TABLE I

UTTERANCE LEVEL BAYESIAN ADAPTIVE INFERENCE PERFORMANCE

problem may be solved if the prior distribution is discriminatively estimated and applied in Bayesian inference. However, this issue is not addressed in this paper.

The last three lines show results for different forms of lower bound approximations. The ML approximation uses a point estimate of the transform with no prior distribution. MAP uses a point estimate that takes into account the prior. VB integrates over the transform prior distribution to calculate the marginal likelihood. All three approximations were used within the N-Best supervision framework, i.e., adaptive inference was performed separately for each possible hypothesis. As these lower bound approximations use an iterative learning process, they must be appropriately initialised. Depending on the form used, the learning process used different initialisations of the transform (distribution) at the zeroth iteration. An identity transform was used for the ML approximation. The MAP approach used the mean of the prior transform distribution. The prior distribution was used in the zeroth iteration of the VB approximation. A single iteration was used in these experiments to estimate the transform distribution used for final inference. Additional iteration gave only small differences in performance [16].

Comparing the VB approximation performance of the ML-SAT system to the unadapted ML-SI baseline, there is a significant gain of 1.3%[6]. The performance of the ML-SI system may be viewed as using standard HMM assumptions in both training and inference. In contrast, using the VB approximation with the ML-SAT system corresponds to using the adaptive HMM DBN in both stages. This significant performance gain illustrates the importance of using the adaptive HMM DBN in both stages. Using the ML approximation with the ML-SAT system, which is the standard ML adaptation scheme but with N-Best supervision rather than 1-best supervision, is about 2.4% absolute worse than that of the ML-SI baseline. This is expected as the transform parameters were estimated using an average of only 300 frames. This problem is reduced by using the MAP estimation, a 1% absolute gain over the ML-SI baseline is obtained. This shows the importance of using prior information when estimating transforms with little data. Note, the VB approximation is 0.3% absolute better than the MAP approach, which is a relatively small gain but has been shown

---

[6]Wherever the term "*significant*" is used, a pair-wise significance test has been done using NIST provided software `sctk-1.2`, which uses a standard approach to conduct significance tests with the significance level of 5% [31].

to be statistically significant [7]. Bayesian adaptive inference was also performed on the ML-SI system. Comparing these results to the performance of the ML-SAT system, the ML-SAT system significantly outperforms the ML-SI system by over 0.3% for all the approximate adaptive inference schemes. This shows the importance of using adaptive HMM in the training stage. For MPE trained systems, similar trends can be observed. However, the gains of the MPE-SAT system over the MPE-SI system are greatly reduced compared to the ML case. For example, the gain of using the VB approximation for the MPE-SAT system over the MPE-SI system is only about 0.6%, which is smaller than the 1.3% gain of the ML-SAT systems. This again shows the effect of using ML based transform prior distributions in a non-discriminative way in inference.

The above experiments on lower bound approximations were all based on the N-Best supervision framework, where one transform distribution was generated for each possible hypothesis. As discussed before, using the 1-Best hypothesis as the supervision may lead to a loose lower bound for the other competing hypotheses and consequently degrade the performance. This effect was investigated using the ML-SAT system. Note that using the ML approximation with 1-Best supervision is the standard unsupervised adaptation approach, which is the most widely used adaptation approach. The results are shown in table II.

Bayesian	Supervision	
Approx.	N-Best	1-Best
ML	35.2	34.4
MAP	31.8	32.0
VB	31.5	32.0

TABLE II

COMPARISON BETWEEN 1-BEST AND N-BEST SUPERVISION (N=150).

Comparing the standard adaptation baseline, i.e. ML approximation with 1-best supervision, to the VB approximation with N-Best supervision, which is the strict Bayesian adaptive inference performance, there is a statistically significant difference of about 3% absolute. For both the MAP and the VB approximations, the 1-Best supervision is significantly worse that the N-Best supervision. One of the reasons for this is that, though the 1-Best supervision, may lead to a tight lower bound for the 1-best hypothesis used as supervision hypothesis, for all the other hypotheses, the transform distribution will have a looser lower bound than using the the N-Best supervision. This biases the inference process to the 1-best supervision hypothesis, The results illustrate the impact of this on WER. It is also interesting to note that the degradation for the VB approximation (0.5%) is larger than MAP (0.2%). This is felt to be because the VB approximation creates a tighter lower bound and is more likely to be tuned to the 1-Best supervision.

---

[7]The MAP approximation in table I was performed with the N-Best supervision, which is not the standard MAP. The standard MAP with 1-Best supervision is shown in II.

C. Incremental Bayesian Adaptive Inference

In the previous section, the homogeneous blocks were assumed to be based on individual utterances and the adaptive inference was performed in a batch mode on all the data. This section gives results using Bayesian adaptive inference in an incremental mode with side-based homogeneous blocks. Only lower bound approximations were examined. The data was incrementally added in the order that it appears in each side.
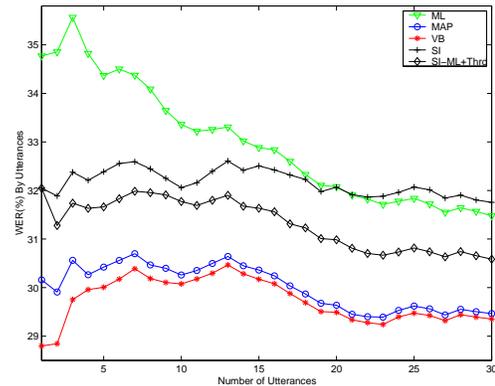

Fig. 3.   Utterance cumulative WER (%) of the ML-SAT system

To investigate the performance of different Bayesian approximations in detail, cumulative WERs of the first 30 utterances of the ML-SAT system are shown in figure 3. The SI line in figure 3 corresponds to the unadapted ML-SI baseline. As an additional baseline for incremental adaptive inference, the ML-SI model was also adapted using the standard robust ML adaptation technique [32]. Here, a threshold was used to determine the minimum posterior occupancy to estimate a robust ML transform. This is the `SI-ML+Thrd` line in figure 3. From figure 3, the `SI-ML+Thrd` line always shows better performance than the unadapted SI system and gradually improves with more data available. This shows that the simple use of a threshold can achieve robustness. When comparing different adaptation approaches on the ML-SAT system, for a limited number of utterances the order of performance is similar to that shown for the ML-SAT system in table I. The VB approximation has the best performance. As the number of utterances increases the difference between the VB and MAP approximations becomes smaller[8]. Given sufficient adaptation data, the point transform estimates are reasonably good approximations. Hence, the VB and MAP approximations show similar performance. The ML approximation is significantly worse than all the others at the beginning because of insufficient adaptation data. From figure 3, the performance of the ML approximation gradually improves as more data comes and outperforms the unadapted SI system after 20 utterances. However, due to the poor performance at the

---

[8]The WER curves in figure 3 are not monotonically decreasing due to the order of the utterances. As shown in table I, the average performance of all utterances for VB approximation is 31.5%. However, the average WER for the *first* utterances of all speakers is below 29% as shown in figure 3. This means that, on average, the first utterances of the speakers happened to be "easy" to recognise. Some "difficult" utterances came later and led to the fluctuations in figure 3.

beginning, the cumulative WER is still significantly worse than `SI-ML+Thrd`, MAP and VB after 30 utterances.

Bayesian Approx	ML Train		MPE Train	
	SI	SAT	SI	SAT
SI-ML+Thrd	31.2	—	27.8	—
ML	32.2	31.8	28.9	28.7
MAP	30.9	30.4	27.7	27.5
VB	30.9	30.3	27.7	27.4

TABLE III

INCREMENTAL BAYESIAN ADAPTIVE INFERENCE PERFORMANCE ON THE COMPLETE DATA SET

Table III shows the overall performance on the complete test data. The `SI-ML+Thrd` in table III is the standard robust ML adaptation on top of the SI models[9]. As expected, the performance of `SI-ML+Thrd` approximation is significantly better than both the ML-SI and the MPE-SI systems in table I. The performance of ML approximation is 0.6% worse than `SI-ML+Thrd`, illustrating the lack of robustness of the ML approximation. Using prior information, the MAP and the VB approximations both significantly outperform the ML approximation and the standard `SI-ML+Thrd` approach. Both give about the same performance. Comparing the performance of the ML-SAT system to the ML-SI system shows that the adaptively trained system consistently and significantly outperforms the non-adaptively trained system by over 0.4% for all approximations. For MPE training, there are similar trends as in the ML case. However, the gains of adaptively trained system are again reduced due to the use of the ML based transform prior distribution.

VII. CONCLUSION

The use of adaptive training has become increasingly popular as more use is made of found data, where there is little control over the acoustic conditions and speaker changes. However there are a number of issues associated with adaptive training that limit how system may currently be applied. These include how to handle limited target domain data, and how to perform "unsupervised" inference. This paper has presented a Bayesian framework for adaptive training and inference that resolves these limitations. In this framework, the model parameters are treated as random variables. For adaptive training, there are two distinct sets of parameters, the canonical model and the transform parameters. Though both of these may be treated as random variables, only the transform parameters are treated in this way in this paper. The canonical model parameters are treated as point-estimates, as standard complexity control techniques can be used during training to ensure robust parameter estimate. Bayesian adaptive inference is then presented as an appropriate way to perform inference with this form of system. As the marginalisation integral associated with this process is intractable, two forms of approximations were

---

[9]In contrast to the standard ML approach, Bayesian approximation does not use any threshold because prior information is considered in the Bayesian adaptive inference. The ML approach in the second row of table III is viewed as an *Bayesian approximation* approach, hence no threshold was set.

---

described. Lower bound approximations, which includes both point estimates (MAP or ML) and variational Bayesian (VB) approach, use an iterative process to tighten a lower bound to the marginal likelihood. In contrast, direct approximations, such as the frame-independent assumption, do not use an iterative process. The marginal likelihood is approximated directly.

The performance of these approximate Bayesian adaptive inference schemes was evaluated on a large vocabulary conversational telephone speech recognition task. MLLR was used as the form of transform to represent each homogeneous block. Both batch and incremental mode inference were investigated. Experiments show that adaptively trained systems can obtain significant gains over multi-style systems, even with very limited data. Variational Bayes is shown to significantly outperform the other approximation approaches with limited data, though compared to the MAP approximation, the absolute gain was not large. In incremental inference, as more data become available, the performance of the MAP approximation gradually approaches the performance of the VB approximation. In addition to ML adaptive training, MPE adaptive training was also examined. Similar trends are observed when using Bayesian adaptive inference. However, the gains of MPE systems are all reduced compared to the ML case because the transform prior is estimated on ML transforms and used in a non-discriminative way during inference.

This paper has only discussed Bayesian adaptive inference within the strict N-Best supervision framework. Empirically, additional approximations, such as Viterbi-like dynamic programming, are required to reduce the computation cost of the N-Best supervision framework. This will be a future research direction. Another possible research direction is to investigate using non-point Bayesian approximations in both adaptive training and inference. This is useful for the scenario where the model complexity can not be controlled to reflect the amount of training data.

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] ——, "Adaptive training for robust ASR," in *Proc. ASRU*, 2001.
[4] K. Yu and M. J. F. Gales, "Bayesian adaptation and adaptively trained systems," in *Proc. ASRU*, 2005.
[5] W. Chou, "Maximum a-posterior linear regression with elliptical symmetric matrix variate priors," in *Proc. ICASSP*, 1999, pp. 1–4.
[6] 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.
[7] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, 2003.
[8] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational Bayesian approach to speech recognition," in *NIPS 15*, 2003.

[9] 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.

[10] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

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

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

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

[14] K. Yu and M. J. F. Gales, "Incremental adaptation using Bayesian inference," in *Proc. ICASSP*, 2006.

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

[16] K. Yu and M. J. F. Gales, "Bayesian adaptation and adaptive training," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR542, 2006.

[17] H. Robbins, "An empirical Bayes approach to statistics," in *Proc. Third Berkeley Symposium on Math. Statist. and Prob.*, 1955, pp. 157–164.

[18] ——, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Stastist.*, vol. 35, pp. 1–20, 1964.

[19] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 1967.

[20] R. Schwartz and Y. L. Chow, "The N-Best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. ICASSP*, 1990, pp. 81–84.

[21] J. Chien and G. Liao, "Transformation-based Bayesian predictive classification using online prior evolution," *IEEE transactions on speech and audio processing*, vol. 9, pp. 399–410, 2001.

[22] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, orlando.

[23] T. Matsui and S. Furui, "N-Best-based unsupervised speaker adaptation for speech recognition," *Computer Speech and Language*, vol. 12, pp. 41–50, 1998.

[24] P. C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proc. ICSLP*, 1996, pp. 1133–1136.

[25] L. F. Uebel and P.C.Woodland, "Speaker adaptation using lattice-based MLLR," *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.

[26] T. Anastasakos and S. V. Balakrishnan, "The use of confidence measures in unsupervised adaptation of speech recognisers," in *Proc. ICSLP*, vol. 6, 1998, pp. 2303–2306.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, Inc, 2001.

[28] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on a Bayesian prediction approach," *IEEE transactions on speech and audio processing*, vol. 7, pp. 426–440, 1999.

[29] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Language Technology Workshop*, 1995, pp. 104–109.

[30] 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.

[31] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition," in *Proc. ICASSP*, 1989, pp. 532–535.

[32] S. J. Young, D. Kershaw, J. J. Odell, D. Ollason, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK version 3.0)*. Cambridge University Engineering Department, 2000.

Kai Yu Biography text here.

Mark Gales Biography text here.