INCREMENTAL ADAPTATION USING BAYESIAN INFERENCE

K. Yu and M.J.F. Gales

Engineering Department, Cambridge University
Trumpington St. Cambridge, CB2 1PZ, U.K.
{ky219,mjfg}@eng.cam.ac.uk

ABSTRACT

Adaptive training is a powerful technique to build system on non-homogeneous training data. Here, a canonical model, representing "pure" speech variability and a set of transforms representing unwanted acoustic variabilities are both trained. To use the canonical model for recognition, a transform for the test acoustic condition is required. For some situations a robust estimate of the transform parameters may not be possible due to limited, or no, adaptation data. One solution to this problem is to view adaptive training in a Bayesian framework and marginalise out the transform parameters. Exact implementation of this Bayesian inference is intractable. Recently, lower bound approximations based on variational Bayes have been used to solve this problem for batch adaptation with limited data. This paper extends this Bayesian adaptation framework to incremental adaptation. Various lower-bound approximations and options for propagating information within this incremental framework are discussed. Experiments using adaptive models trained with both maximum likelihood and minimum phone error training are described. Using incremental Bayesian adaptation gains were obtained over the standard approaches, especially for limited data.

1. INTRODUCTION

Adaptive training is a powerful approach to build speech recognition systems on *non-homogeneous* data [1]. During training, two sets of parameters are extracted. The first set is the *canonical* model parameters, which represent the "pure" speech variability. The second set, the *transform* parameters, 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 was originally derived for maximum likelihood training. However discriminative training[1] has also been examined within this framework [3, 4]. For some situations, such as conversational telephone speech, no supervised adaptation data is available, thus the correct transcript to discriminatively train the test set transformation is not possible. To maintain a consistent criterion in transform estimation in testing adaptation, simplified discriminative adaptive training is normally used where only the canonical model parameters are discriminatively updated with the ML-estimated transforms fixed [4]. This is the approach adopted for the discriminative adaptive training discussed in this paper.

Adaptive training may be described within a Bayesian framework [5, 6]. Distributions over the canonical model parameters and transform parameters are now estimated. The likelihood of the observation sequence is then obtained by marginalising out over these canonical model and transform distributions. Using standard techniques to control the complexity of the canonical model and number of transforms, the usual point estimate adaptive training can be justified within this Bayesian framework [6]. However during recognition, or inference, there is usually no control over the amount of data available. It is therefore preferable to use a full Bayesian approach for inference. This is the scenario considered in this paper. Point estimates will be used for the canonical model parameters and distributions for the transform parameters. As exact inference using this framework is intractable, lower bound approximations have previously been investigated for batch adaptation and adaptively trained systems [6]. Two classes of approximation have been examined. The first is based on point estimates, using either Maximum Likelihood (ML) [7] or Maximum a Posteriori (MAP) [8]. The second is based on variational Bayes (VB) [9, 10, 6] with transform distributions.

For some situations, rather than all the data being available in one block, as in batch adaptation, the data becomes available causally. Using *incremental*, or on-line, adaptation the transform may be updated as each utterance becomes available and recognition results produced causally. This paper investigates lower bound based Bayesian techniques for incremental adaptation. Various information propagation strategies between utterances are described and their effect on computational cost discussed. An efficient incremental Bayesian adaptation framework with recursive transform distribution update formulae is established. This is then applied to both ML and Minimum Phone Error (MPE) trained models [2]. Results are presented on a Conversational Telephone Speech (CTS) task.

2. ADAPTATION USING BAYESIAN INFERENCE

The aim of inference is to find the hypothesis, $\hat{\mathcal{H}}$, satisfying

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} p(\mathbf{O}|\mathcal{H})P(\mathcal{H}) \tag{1}$$

for observation sequence $\mathbf{O}$ where $P(\mathcal{H})$ is the language model and $p(\mathbf{O}|\mathcal{H})$ is the acoustic likelihood. HMMs with Gaussian mixture model (GMM) as the state output distributions are commonly used as the underlying acoustic model to calculate $p(\mathbf{O}|\mathcal{H})$. Alternatively in adaptive training HMMs are used to model the observation given the transform for that homogeneous data block,

This work was in part supported by DARPA under the EARS programme and under the GALE programme via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

[1]The adaptive discriminative training using MLLR in this paper is implemented using the MPE criterion [2] rather than the Maximum Mutual Information criterion in [3].

$p(\mathbf{O}|\mathcal{H},\mathcal{T})$. If there is no adaptation data available then the likelihood is computed as [6]

$$p(\mathbf{O}|\mathcal{H}) = \int_{\mathcal{T}} p(\mathbf{O}|\mathcal{H},\mathcal{T})p(\mathcal{T})\,d\mathcal{T} \quad (2)$$

where $\mathbf{O}$ belongs to a single homogeneous block and $p(\mathcal{T})$ is the prior transform distribution. If adaptation data is available then the posterior transform distribution given the adaptation data is used in equation 2 rather than the prior, this is sometimes referred to as *posterior adaptation* [5]. As the amount of adaptation data increases the posterior distribution may be approximated by a point estimate based on either the ML or MAP transform estimate.

Direct calculation of equation 2 is intractable with HMMs, so various forms of approximation are used. In this work a lower bound approximation, based on Jensen's inequality, is used.

$$\log p(\mathbf{O}|\mathcal{H}) \geq \left\langle \log \frac{p(\mathbf{O},\boldsymbol{\theta}|\mathcal{T},\mathcal{H})p(\mathcal{T})}{q(\boldsymbol{\theta},\mathcal{T})} \right\rangle_{q(\boldsymbol{\theta},\mathcal{T})} \quad (3)$$

where $< f(x) >_{q(x)}$ denotes the expectation of function $f(x)$ with respect to the distribution of $q(x)$ and $q(\boldsymbol{\theta},\mathcal{T})$ is a joint distribution over the Gaussian component sequence $\boldsymbol{\theta}$ and transform parameters $\mathcal{T}$. The above becomes equality when

$$q(\boldsymbol{\theta},\mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O},\mathcal{H},\mathcal{T})p(\mathcal{T}|\mathbf{O},\mathcal{H}) \quad (4)$$

Using equation 4 is impractical, so alternative approximate forms of $q(\boldsymbol{\theta},\mathcal{T})$ are required. The tightness of the bound is dependent on the precise form of the approximation used. There are two forms commonly used:

1. Point Estimates [8, 6]: With sufficient adaptation data the transform distribution can be approximated by a Dirac delta function

$$q(\boldsymbol{\theta},\mathcal{T}) = P(\boldsymbol{\theta}|\mathbf{O},\mathcal{H},\hat{\mathcal{T}})\delta(\mathcal{T}-\hat{\mathcal{T}}) \quad (5)$$

The point estimate value, $\hat{\mathcal{T}}$, may be obtained using ML or MAP estimates. Substituting this point estimate into equation 3 yields a lower bound involving the entropy of the delta function.

$$\log p(\mathbf{O}|\mathcal{H}) \geq \log p(\mathbf{O}|\mathcal{H},\hat{\mathcal{T}}_k) + \log p(\hat{\mathcal{T}}_k) + H(\delta(\mathcal{T}-\hat{\mathcal{T}}_k)) \quad (6)$$

where $H()$ is the entropy. k is used to indicate the iteration number as the lower bound can be made tighter by iteratively refining the component sequence posterior distribution (equivalent to the standard EM training). As the entropy of a delta function is $-\infty$, this yields a very loose bound. However since only the rank ordering is of interest in inference and the entropy of the delta function is the same for all values, the entropy term may be ignored. If a non-informative prior is used, the MAP estimate becomes the Maximum Likelihood (ML) estimate. The advantage of point estimates is low computational cost and the compatibility with standard training/decoding algorithms. However, it may not be robust, even for MAP, for very limited adaptation data case.

2. Variational Bayes (VB) [9, 10, 6]: Rather than using a point estimate a distribution over the transform parameters may be used. In the VB approximation the component sequence and transform distributions are assumed to be conditionally independent

$$q(\boldsymbol{\theta},\mathcal{T}) \quad = \quad q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H})q(\mathcal{T}|\mathbf{O},\mathcal{H}) \quad (7)$$

The VBEM algorithm [9, 6] can be used to optimise the lower bound, equation 3 with respect to the two variational distributions rather than particular parameter values. This is an iterative process resulting in an optimal transform distribution on which the lower-bound ordering can be based. The resultant lower bound after K iterations can be expressed as

$$\log p(\mathbf{O}|\mathcal{H}) \geq \log \mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H}) - \int_{\mathcal{T}} q_K(\mathcal{T}) \log \frac{q_K(\mathcal{T})}{p(\mathcal{T})}\,d\mathcal{T} \quad (8)$$

where $q_K(\mathcal{T})$ is the compact notation for $q_K(\mathcal{T}|\mathbf{O},\mathcal{H})$. $\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H})$ is the normalisation term for $q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H})$ and can be calculated in a similar fashion to $p(\mathbf{O}|\mathcal{H},\hat{\mathcal{T}}_K)$ in equation 6 except that the pseudo distribution is used rather than the standard Gaussian distribution [6]. As VB employs real distributions, it performs more robustly than the point estimate with very limited adaptation data [6].

For both the point and the VB approximations, the lower bound rank ordering, equation 8 or 6, is assumed to give the same ordering as the actual likelihood. These lower approximations will be used for inference. For the point estimates, either MAP or ML, there is an interesting difference between using this lower bound approximation and standard unsupervised adaptation. For the lower bound approximation a transform, and resultant lower bound value, is estimated for *each* hypothesis. The tightness of the bound is then increased for each of the hypothesis. In standard unsupervised adaptation only the 1-best hypothesis is used to estimate the transform, which is then used to rescore the data. This introduces an inherent bias towards the 1-best solution.

An alternative form of approximation to equation 2 is the Frame-Independent (FI) [5], also referred to as the Bayesian predictive distribution [11]. Here the prior distribution is directly applied to marginalise each component distribution independently. This effectively alters the form of dynamic Bayesian network being used [5]. This has the advantage over the lower-bound approaches that the standard decoding schemes may be used.

3. INCREMENTAL BAYESIAN ADAPTATION

The Bayesian adaptation discussed in section 2 describes decoding in a *batch* mode. All test data are assumed to be available for decoding in a single block. However, in some real world applications test data often become available gradually. To deal with this issue *incremental* adaptation is often used. Here information from the previous utterances are propagated to the current utterance. The current utterance is then decoded and the result output. This section will discuss incremental adaptation within a Bayesian framework. The key issue is what information to propagate and how to use it. For incremental adaptation, each homogeneous data block is assumed to be split into U utterances, $\mathbf{O} \equiv \mathbf{O}_{1:U} \equiv \{\mathbf{O}_1,\cdots,\mathbf{O}_U\}$. Information is propagated to the U^{th} utterance from the previous $U-1$ utterances. The hypothesis for all the data, $\mathcal{H}$ consists of a set of hypotheses for utterances within it, $\mathcal{H}_{1:U} \equiv \{\mathcal{H}_1,\ldots,\mathcal{H}_U\}$. Various levels of information can be propagated.

1. No information: The lower bound for all U utterances is optimised. This involves rescoring all U blocks, obtaining a new $\hat{\mathcal{H}}_{1:U}$. Thus the U^{th} utterance may change the "best" hypothesis for the preceding utterances. This approach breaks the standard causal aspects of incremental adaptation and is highly computationally expensive.

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 U^{th} block.

$$q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H}) = q(\boldsymbol{\theta}|\mathbf{O},\hat{\mathcal{H}}_{1:U-1},\mathcal{H}_U) \quad (9)$$

$$q(\mathcal{T}|\mathbf{O},\mathcal{H}) = q(\mathcal{T}|\mathbf{O},\hat{\mathcal{H}}_{1:U-1},\mathcal{H}_U) \quad (10)$$

In this configuration there is a choice of 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

$$q_0(\mathcal{T}|\mathbf{O},\hat{\mathcal{H}}_{1:U-1},\mathcal{H}_U) = q_K(\mathcal{T}|\mathbf{O}_{1:U-1},\hat{\mathcal{H}}_{1:U-1}) \quad (11)$$

where K is the number of VBEM iterations used. Inference only involves possible hypotheses for the U^{th} utterance. The VBEM algorithm remains unchanged except that $\mathbf{O}_{1:U-1}$ only needs to be re-aligned against $\hat{\mathcal{H}}_{1:U-1}$, rather than all possibilities.

3. Posterior sequence distribution and hypotheses: Just propagating the hypotheses still requires the 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 9 becomes

$$q(\boldsymbol{\theta}|\mathbf{O},\mathcal{H}) = q(\boldsymbol{\theta}_U|\mathbf{O}_U,\mathcal{H}_U)\prod_{u=1}^{U-1} q_K(\boldsymbol{\theta}_u|\mathbf{O}_u,\hat{\mathcal{H}}_u) \quad (12)$$

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., only the sufficient statistics of the U^{th} utterance need to be accumulated. This is the most efficient form and the one used in this paper.

Using the information propagation strategy 3, an efficient, modified version of the VBEM algorithm described in [6] can be derived. Initially only the VB approximation is considered.

1. Initialisation: set $k = 1$, the initial transform distribution is given by 11. For the first utterance, set $q_0(\mathcal{T}) = p(\mathcal{T})$.
2. VBE step: $q_k(\boldsymbol{\theta}_U|\mathbf{O}_U,\mathcal{H}_U)$, and corresponding statistics, are calculated using the forward backward algorithm with Gaussian components adapted by the transform distribution of the previous iteration, $q_{k-1}(\mathcal{T}|\mathbf{O}_{1:U},\hat{\mathcal{H}}_{1:U-1},\mathcal{H}_U)$, similar to [6].
3. VBM step: The optimal transform distribution can be shown as

$$\begin{aligned}\log q_k(\mathcal{T}|\mathbf{O}_{1:U},\hat{\mathcal{H}}_{1:U-1},\mathcal{H}_U) \propto \\ \log p(\mathcal{T}) + \langle \log p(\mathbf{O}_U,\boldsymbol{\theta}_U|\mathcal{T},\mathcal{H}_U)\rangle_{q_k(\boldsymbol{\theta}_U|\mathbf{O}_U,\mathcal{H}_U)} \\ + \textstyle\sum_{u=1}^{U-1}\left\langle \log p(\mathbf{O}_u,\boldsymbol{\theta}_u|\mathcal{T},\hat{\mathcal{H}}_u)\right\rangle_{q_K(\boldsymbol{\theta}_u|\mathbf{O}_u,\hat{\mathcal{H}}_u)}\end{aligned} \quad (13)$$

The sufficient statistics are a summation of those of the current utterance and those of the previous $U-1$ utterances, which are propagated and do not need to be re-calculated. This recursive formulae significantly reduces the computation cost.
4. $k = k+1$**. Goto 3** until $k = K$.

Having obtained the optimal transform distribution with the above incremental VBEM algorithm, the ranking for inference of the U^{th} utterance can be done using the VB lower bound in equation 8. The normalisation term can also be efficiently calculated using

$$\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H}) = \mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O}_U,\mathcal{H}_U)\prod_{u=1}^{U-1}\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O}_u,\hat{\mathcal{H}}_u) \quad (14)$$

Note that that a normalisation term must be calculated for each possible hypothesis $\mathcal{H}_U$.

With the point estimate approximations, a similar incremental EM algorithm and inference process can be derived. The main difference is that the transform estimate, rather than the distribution, is propagated. The initial estimate for the first utterance in this case can be set to the mean of $p(\mathcal{T})$ for MAP and an identity transform for ML.

4. EXPERIMENTAL RESULTS

The performance of the incremental Bayesian adaptation was evaluated on a large vocabulary speech recognition CTS task. The training data set consists of 5446 speakers, about 295 hours of data. The performance was evaluated on the 2003 evaluation test dataset, `eval03`, consisting of 144 speakers, about 6 hours of data. Standard Speaker Independent (SI) ML and MPE trained decision-tree state-clustered triphone models with 16 components per state were built using a 39-dimensional PLP-based frontend with HLDA and VTLN. For more details of the training configuration see [6]. Speaker adaptively trained (SAT), both ML and MPE based, systems were also built using speech and silence MLLR transforms. The priors for these transforms were estimated separately. As previously mentioned, for the MPE-SAT system a simplified training framework using the ML-transforms was run. Since the inference in sections 2 and 3 requires the whole utterance hypothesis to be used, N-best rescoring was employed. Two 150-best lists were generated for ML and MPE systems from corresponding SI models. All results shown are based on the two 150-best lists, though there was little performance difference when spot-checks were run with a 300-best list. During adaptation, 1 iteration ($K = 2$) is employed for updating the transform estimate or distribution, which is then used to compute the lower bound. For these experiments a homogeneous block was a conversation side, average length 153.75 seconds, with an average utterance length of 3.13 seconds.

Bayesian Approx.	ML Train		MPE Train	
	SI	SAT	SI	SAT
—	32.83	—	29.20	—
FI	—-	32.90	—	29.74
ML	35.54	35.16	—	32.27
MAP	32.16	31.76	—	28.80
VB	31.77	31.50	—	28.63

Table 1. Utterance level adaptation of ML-SI, ML-SAT and MPE-SAT systems with single Gaussian prior transform distribution

As a baseline for the incremental adaptation, table 1 shows the performance of ML and MPE trained SI and SAT model sets where the adaptation was performed on a per-utterance basis (similar to the ML results in [6]). The FI approximation [5] shows similar performance for the ML SAT system as the ML SI, since only a single component transform prior is being used. However for the MPE system the use of FI approximation is about 0.6% absolute worse than the SI MPE system. This is felt to be due calculating the transform prior using ML. Thus though the canonical model parameters are discriminatively trained, the discriminative power is reduced by applying an ML based transform prior. For this reason the MPE-SI system was not evaluated for configurations using transform priors. This issue also reduces the gains of the MPE-SAT VB system over the baseline MPE-SI system, about 0.5%

absolute, compared to 1.3% absolute of the ML-SAT VB system over the baseline ML-SI system. As an additional contrast, standard MAP adaptation (using the 1-best hypothesis) gave an error rate of 32.0% WER. This is about 0.2% worse than the lower-bound Bayesian approximation here. This illustrates the bias that results from using a single hypothesis to estimate the transforms for all the N-best rescoring, as discussed in section 2.

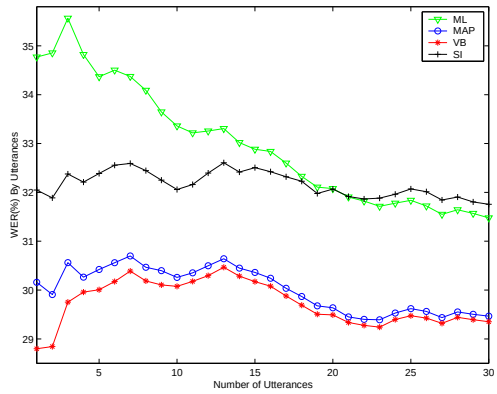

Fig. 1. Incremental adaptation cumulative WER (%) of different number of utterances on ML-SAT system

To investigate performance of different Bayesian approximations for incremental adaptation, cumulative WERs of the first 30 utterances of the ML-SAT system are plotted in figure 1. The SI line in figure 1 refers to the non-adapted SI performance. As expected, for a limited number of utterances the order of performance is similar to that shown in table 1. The VB approximation shows the best performance. As the number of utterances increases the difference between the VB and MAP approximations become far smaller.

Bayesian Approx	ML Train		MPE Train	
	SI	SAT	SI	SAT
—	32.83	—	29.20	—
ML+thresh	31.23	—	27.81	—
ML	32.23	31.84	—	28.72
MAP	30.92	30.40	—	27.47
VB	30.88	30.31	—	27.44

Table 2. Incremental adaptation of ML-SI, ML-SAT and MPE-SAT systems with single Gaussian transform distribution

As a baseline for Bayesian incremental adaptation, standard incremental adaptation using an occupancy threshold to decide whether a transform can be robustly estimated is shown in table 2 (`ML+thresh`). As expected incremental adaptation shows gains over standard SI decoding. The final results after a whole conversation side for various Bayesian approximations are also shown in table 2. For the ML trained systems, the best performance was obtained using the VB approximation with a SAT model set. This gave an gain of 0.9% absolute over the baseline system. For MPE training the MPE-SAT system also gave gains over the standard adapted MPE-SI system, though the gain was less, 0.4%, than for the ML systems. The ML prior will again have affected the performance. As expected for both ML and MPE trained systems MAP and VB performed about the same, as after a few utterances the MAP point estimate is a reasonable approximation to the transform distribution.

5. CONCLUSION

This paper has described an incremental Bayesian adaptation framework. A lower bound approximation is used to make the inference practical and both point estimates and variational Bayes approximations are discussed. Various forms of incremental adaptation are described, where different levels of information are propagated from one utterance to another. Due to its efficiency a scheme where both the hypotheses and component posterior distributions from the previous utterances are propagated was implemented. Using this framework it is possible to use adaptively trained systems for incremental adaptation, even when the utterance lengths are very short. These adaptively trained systems showed gains over standard SI systems for a CTS incremental adaptation task. In addition to ML trained systems, MPE trained systems were examined. Again adaptively trained systems showed gains over standard SI systems, though the gains were smaller. This is felt to be due to the use of an ML estimated transform prior. Discriminative forms of transform prior will be investigated in future work.

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

[3] T. Schaaf J. McDonough and A. Waibel, "On maximum mutual information speaker-adapted training," in *Proc. ICASSP*, Florida, USA, May, 2002.

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

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

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

[7] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous sensity HMMs using multivariate linear regression," *ICSLP*, pp. 451–454, 1994.

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

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

[10] 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.

[11] 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.