INCREMENTAL BAYESIAN ADAPTATION

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. A canonical model, representing "pure" speech variability and a set of transforms representing unwanted acoustic variabilities are trained. It is necessary to have transforms in order to deal with the testing acoustic conditions. One problem here is to robustly estimate the transforms parameters where there is limited or even no adaptation data. Recently, Lower bound based Bayesian approaches have been used to solve this problem in batch adaptation mode, of which point estimates, MAP or ML, and variational Bayes are two main approximation forms. This paper extends the Bayesian adaptation framework to incremental mode. Strict Bayesian inference and various approximated information propagation strategies during adaptation are discussed in detail. The techniques are examined for both ML and discriminative systems. The experiments on a large vocabulary speech recognition task showed that the incremental Bayesian adaptation can lead to robust performance with limited data at the start and gradually improve with more data available.

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. Though adaptive training is usually derived from a maximum likelihood perspective, it may be described within a Bayesian framework [2]. With sufficient training data, the standard point estimate adaptive training can be justified within the Bayesian framework [2]. However during recognition there is usually no control over the amount of data available. It is therefore preferable to use a full Bayesian approach to obtain robust estimate of transform parameters. Lower bound based Bayesian approaches have been investigated in batch adaptation mode and applied to adaptively trained systems with MLLR transforms [2]. The standard point estimates, Maximum Likelihood (ML) [3] or Maximum a Posteriori (MAP) [4], and variational Bayes (VB) [5, 6, 2] with real transform distributions are two main forms of approximation approaches. Using a strict Bayesian inference process in *batch* mode, the VB approach was shown to significantly outperform both standard point estimates when the adaptation data is very limited [2]. The strict Bayesian inference [1] is further investigated in this paper, where the inference evidence is calculated for *every* hypothesis candidate of the same observation sequence. This yields a different adaptation process from the standard iterative self adaptation.

Another important area is incremental, or on-line, adaptation. In on-line adaptation, the data often becomes available gradually rather than in a batch. This paper investigates the lower bound based Bayesian adaptation in incremental mode. Various Bayesian information propagation strategies are investigated. An efficient incremental Bayesian adaptation framework with recursive transform (distribution) update formulae is established.

Discriminative adaptive training is used in all state-of-the-art speech recognition systems [7, 8]. Due to the lack of correct transcription in unsupervised adaptation, transforms in adaptation is hard to be discriminatively estimated. To keep consistent criterion in transform estimation in testing adaptation, a simplified discriminative adaptive training is normally used where only the canonical model parameters are discriminatively updated with the ML-estimated transforms fixed [8]. Bayesian adaptation can also be applied to the discriminative canonical model with the prior transform distribution estimated from the ML training transforms. In this paper, ML and discriminative adaptively trained systems with MLLR transforms are used as a particular application of incremental Bayesian adaptation. The discriminative system in this paper is implemented using the Minimum Phone Error (MPE) criterion [9] rather than the Maximum Mutual Information criterion in [7].

2. ADAPTATION USING BAYESIAN INFERENCE

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

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

where $P(\mathcal{H})$ is the language model, for example N-gram, $p(\mathbf{O}|\mathcal{H})$ is the acoustic marginal likelihood of interest in Bayesian adaptation. Normally HMMs, with Gaussian mixture model (GMM) as the state output distributions, are used as the underlying acoustic model to calculate $p(\mathbf{O}|\mathcal{H}, \mathcal{T})$. The standard HMMs training regards the training data as a whole block. Hence, the resultant HMMs include both speech and non-speech variabilities and can be directly used for inference. In comparison, the adaptive training separates the speech and non-speech variabilities on each real homogeneous block. The resultant canonical model only represents

This work was supported by DARPA grant ????????. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

[1] The general term *inference* in this paper has the same meaning with *recognition* or *decoding* in speech community.

speech variability and requires transforms to represent acoustic conditions in testing adaptation. Though the point estimate of the canonical model can be justified with the sufficient data assumption during training [2], it may not be robust to employ a point estimate of transform during adaptation when there is limited, or even no, adaptation data. It is then useful to extract the prior distribution of the transform parameters from the training data and employ a full Bayesian approach in testing adaptation. Hence,

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

where $\mathbf{O}$ is assumed to belong to a single homogeneous block, $p(\mathcal{T})$ is the prior transform distribution. The prior may be updated to the posterior distribution given some supervision data, in which case, the Bayesian adaptation is referred to as *posterior adaptation* [10]. In this paper, the update of the prior is not concerned. As direct calculation of 2 is infeasible, a lower bound approximation is used. Applying Jensen's inequality yields

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

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 component sequence $\boldsymbol{\theta}$ and transform parameters $\mathcal{T}$. The above becomes equality when

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

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

1. Point Estimates (Standard MAP or ML) [4, 2]

With sufficient adaptation data assumption, the transform distribution can be approximated by a Dirac delta function

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

This is equivalent to the standard Maximum a Posteriori (MAP) adaptation. The equivalent lower bound of 3 can be updated using the standard EM algorithm, resulting in a MAP estimate $\hat{\mathcal{T}}_K$ after K iterations. The lower bound can then be calculated by

$$\mathcal{L}(\hat{\mathcal{T}}_K) = \log p(\mathbf{O}|\mathcal{H},\hat{\mathcal{T}}_K) + \log p(\hat{\mathcal{T}}_K) \tag{6}$$

If a non-informative prior is used, the point estimate degrades to the Maximum Likelihood (ML) estimate. The advantage of point estimates is the 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) [5, 6, 2]

The variational component sequence and transform distributions are assumed to be conditionally independent

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

Then VBEM algorithm [5, 2] was proposed to optimise the lower bound 3 with respect to the two variational distributions rather than particular parameters. This is an iterative process resulting in an optimal transform distribution. After K iterations' VBEM algorithm, the resultant lower bound can be expressed as

$$\mathcal{L}\left(q_K(\mathcal{T})\right) = \log \mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H}) - \int_{\mathcal{T}} q_K(\mathcal{T}) \log \frac{q_K(\mathcal{T})}{p(\mathcal{T})}\,d\mathcal{T} \tag{8}$$

where $q_K(\mathcal{T})$ is the brief notation for $q_K(\mathcal{T}|\mathbf{O},\mathcal{H})$, $\mathcal{Z}_{\boldsymbol{\Theta}}(\mathbf{O},\mathcal{H})$ can be calculated as the standard $p(\mathbf{O}|\mathcal{H},\hat{\mathcal{T}}_K)$ except that $q_K(\mathcal{T})$ is applied to the Gaussian components instead of $\hat{\mathcal{T}}_K$ [2]. As VB employs real distributions, it has been shown to perform more robustly than the point estimate with very limited adaptation data [2].

Given the form of $q(\boldsymbol{\theta},\mathcal{T})$, the tightness of the lower bound is dependent on the iteration number. At each iteration, the lower bound is guaranteed to increase to the marginal likelihood. The final lower bound, equation 8 or 6, is used for inference instead of 2. The assumption here is that the ordering of the real likelihood is similar to the ordering of the lower bound if it is tight enough. It is worth emphasising that, to do strict inference with 2, the lower bound of *every* possible hypothesis needs to be calculated. This means one distinct transform (distribution) is required for each hypothesis of the same observation sequence. It is interesting to compare this to the standard iterative adaptation such as iterative MLLR. In iterative MLLR, one transform is estimated for each observation sequence using the 1-best hypothesis as supervision. This transform is then used to do inference on all possible hypothesis and the process repeated if necessary. As the estimated transform is biased to the particular supervision, the lower bound of other hypothesis candidates should be looser than the lower bound calculated using 3. This may significantly affect the performance especially for complex transforms or short sentences. This hypothesis-bias problem for unsupervised self adaptation has been discussed in detail in [10].

Though calculating lower bound for *every* possible hypothesis may be trivial for simple tasks, such as isolated word recognition, it is hard to be directly used for tasks of Large Vocabulary Continuous Speech Recognition (LVCSR) because the number of hypothesis combinations is too large to explore. One possible solution is to generate candidate hypothesis sequences with a reasonably small number, stored as *N-Best List*. Strict inference based on 3 may then be done on those hypothesis candidates, which is referred to as *N-Best rescoring*. Given sufficient hypothesis candidates generated, this rescoring process is reasonable. This approach will be adopted in the experiments of this paper. Another widely used alternative for inference on LVCSR tasks is the *Viterbi* algorithm, in which the likelihood of one optimal state sequence is used to approximate the whole likelihood. As the Viterbi algorithm is hard to be used for the strict inference, it is not concerned in this paper.

3. INCREMENTAL BAYESIAN ADAPTATION

The Bayesian adaptation discussed in section 2 runs in a *batch* mode where all test data are assumed to be available before adaptation. However, in the real world, test data often become available gradually. To deal with this issue, *incremental* adaptation is often used, where information from the previously inferred data may be propagated to do adaptation and inference of the subsequent data. This section will investigate the incremental adaptation within the Bayesian framework. The key problem here is what information to propagate and how to use it. Different information propagation strategies are discussed with the variational Bayes approximation. The point estimate version may be easily analogised.

Each homogeneous data block is assumed to be split into U utterances, for example, observation $\mathbf{O} \equiv \mathbf{O}_{1:U} \equiv \{\mathbf{O}_1, \cdots, \mathbf{O}_U\}$. Then, the information that may be propagated to the U^{th} utterance include the previously inferred hypothesis sequence $\hat{\mathcal{H}}_{1:U-1}$ and the optimal variational distributions $q_K(\mathcal{T}|\mathbf{O}_{1:U},\mathcal{H}_{1:U})$ and $q_K(\boldsymbol{\theta}_{1:U}|\mathbf{O}_{1:U},\mathcal{H}_{1:U})$ where K is the iteration number in VBEM

learning. These information may be propagated individually or together, resulting in different adaptation/inference process.

1. No information

In this case, the lower bound of the *whole* U utterances has to be re-optimised and the inference has to be redone from scratch. The incremental adaptation degrades to the batch mode.

2. Inferred hypothesis sequence $\hat{\mathcal{H}}_{1:U-1}$ and posterior transform distribution $q_K(\mathcal{T}|\mathbf{O}_{1:U-1}, \hat{\mathcal{H}}_{1:U-1})$

In this case, $\mathcal{H}_U$ is the only free partial hypothesis sequence during the adaptation, while the hypothesis sequence from the 1^{st} to the $(U-1)^{th}$ is fixed as $\hat{\mathcal{H}}_{1:U-1}$. Meanwhile, the previous variational transform distribution $q_K(\mathcal{T}|\mathbf{O}_{1:U-1}, \hat{\mathcal{H}}_{1:U-1})$ is used as the initial transform distribution instead of the prior $p(\mathcal{T})$ to get better alignment. Then the variational distributions in 7 becomes

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

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

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

As a result, the inference only concerns all possible hypotheses of the U^{th} utterance. The VBEM algorithm remains unchanged except that $\mathbf{O}_{1:U-1}$ only needs to be re-aligned against $\hat{\mathcal{H}}_{1:U-1}$.

3. Posterior component sequence distribution $q_K(\boldsymbol{\theta}_{1:U-1}|\mathbf{O}_{1:U-1}, \hat{\mathcal{H}}_{1:U-1})$

The posterior component sequence distribution is closely related to the sufficient statistics for updating the transform distribution. With this information further propagated, 9 becomes

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

From 12, the previous $U-1$ utterances do not need to be re-aligned at all, only $q(\boldsymbol{\theta}_U|\mathbf{O}_U, \mathcal{H}_U)$ needs to be calculated, i.e., only the sufficient statistics of the U^{th} utterance need to be accumulated. With the information propagation strategy 3, an efficient VBEM algorithm can be derived as below:

1. Initialisation:

set $k = 0$, the initial transform distribution is given by 11. For the first utterance, set $q_0(\mathcal{T}) = p(\mathcal{T})$.

2. VBE step:

In this 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 [2].

3. VBM step:

The optimal transform distribution can be shown as

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

From 13, the total sufficient statistics is 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 inference on the U^{th} utterance is done using the VB lower bound 8. It can be shown the normalisation term in 8 can also be calculated recursively

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

Note that here one normalisation term is calculated for *each* possible hypothesis $\mathcal{H}_U$. $\hat{\mathcal{H}}_U$ is worked out by strict inference. Then the process can go ahead to the next utterance. With point estimate approximation, such as MAP, similar incremental EM algorithm and strict inference process can be derived. The main difference is that here 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. In this paper, the incremental Bayesian adaptation is applied to MLLR transform. The exact formulae are similar to those in [2].

4. EXPERIMENTAL RESULTS

The performance of the incremental Bayesian adaptation was evaluated on a large vocabulary speech recognition system, conversational telephone speech task. The training data set consists of 5446 speakers (2747 female, 2699 male), about 295 hours of data. The performance was evaluated on the 2003 evaluation test dataset, `eval03`, consisting of 144 speakers (77 female, 67 male), about 6 hours of data. All systems used a 12-dimensional PLP front-end with log energy and its first, second and third derivatives with Cepstral mean and variance normalisation and VTLN. An HLDA transform was then applied to reduce the feature dimension to 39. A standard decision-tree state-clustered triphones with an average of 16 Gaussian components per state was constructed as the starting point for the adaptive training. This is the baseline ML speaker-independent (SI) model. After 4 iterations of standard MPE training [9], the baseline MPE-SI model was obtained.

Two adaptively trained systems were built. The first was a ML-SAT system constructed using MLLR, where a single component Gaussian prior was estimated from the training data. The second was a simplified MPE-SAT system with the ML-estimated transforms fixed during MPE training of the canonical model. Hence, the transform prior from the ML-SAT training is also applicable to the MPE-SAT system. To evaluate the effect of adaptive training, a transform prior for the non-adaptively trained ML-SI system was extracted from the training transforms estimated based on the ML-SI model. For the SAT systems in this paper, separate speech and silences transforms were used, the priors for which were independently estimated. As indicated in section 2, N-best rescoring is employed for inference. Two 150-best lists were generated for ML and MPE systems from corresponding SI models respectively. All results shown are based on the two 150-best lists. During adaptation, 1 iteration is employed for updating the transform (distribution). The baseline performance is shown in table 1. The

Incremental Adaptation	ML-SI	MPE-SI
—	32.83	29.20
ML + Threshold	31.23	27.81

Table 1. `eval03` WER (%) of baseline SI systems

second row shows the performance of standard adaptation on SI systems. A threshold was used to determine the minimum posterior occupancy to estimate a robust ML transform. On the contrary, Bayesian adaptation approaches did not employ any threshold because the prior information is considered [2].

[2]ML approach in table 2 was viewed as an approximation of Bayesian

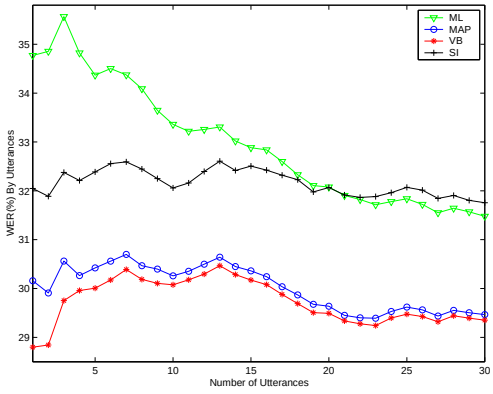

Fig. 1. Incremental adaptation WER (%) of different number of utterances on ML-SAT system

To investigate the effect of the data amount on different Bayesian approaches, WERs of the first 30 utterances of the ML-SAT system were plotted in figure 1. The SI line in figure 1 refers to the non-adapted SI performance. At the beginning of adaptation, the ML adaptation performed significantly worse than SI due to insufficient adaptation data. With more data available, the ML estimate became more reasonable, hence the performance gradually improved and outperformed SI. The insufficient data problem was well solved by MAP and VB, they both had lower WERs than ML and SI all the time. Comparing the two approaches shows that VB significantly outperformed MAP at the beginning, consistent with [2]. With more data available, the two gradually became close. This is expected as VB is more robust than MAP only when the adaptation data is limited. Given sufficient adaptation data, point estimate is feasible and the variance of the VB transform distribution is small, hence the two became close to each other.

Approx.	ML-SI	ML-SAT	MPE-SAT
ML	32.23	31.84	28.72
MAP	30.92	30.40	27.46
VB	30.88	30.31	27.42

Table 2. `eval03` WER (%) of incremental adaptation with single Gaussian transform distribution

The final incremental Bayesian adaptation results are shown in table 2. Observing the performance of ML adaptation on ML-SAT, it is about 1% absolute better than the ML-SI performance. This is the effect of data accumulation in incremental adaptation, which may result in a robust transform estimate. However, this result is still 0.6% worse than the standard ML adaptation with threshold, which shows the ML approximation can not take full advantage of the ML-SAT system. Employing the prior information, MAP and VB both significantly outperformed the standard ML adaptation. Though VB is 0.1% better than MAP, the difference is small due to the gradually increased data amount. Comparing the performance of ML-SAT system to ML-SI system shows that the adaptively trained system consistently and significantly outperformed non-adaptively trained system by over 0.4%. The simplified MPE-SAT system yielded significant gains (about 3%) over ML-SAT systems with all Bayesian adaptation techniques. The relative gain between different Bayesian adaptation techniques of ML-SI and MPE-SAT systems were also similar to that of ML-SAT system. This shows that the gain from Bayesian adaptation is additive to the gain of adaptive training and discriminative training.

adaptation, hence, no threshold was set, either.

5. CONCLUSION

This paper describes an incremental Bayesian adaptation framework. Strict Bayesian inference for adaptation is discussed. Lower bound based approaches are employed to approximate the true inference evidence. The point estimate only works when adaptation data is sufficient while the variational Bayes approach uses real distributions over parameters and may obtain more robust performance than the point estimate. The Bayesian adaptation approaches are then extended to incremental mode. Information gathered from previously inferred utterances can be propagated do to adaptation on subsequent utterances. By appropriately propagating the information, efficient recursive adaptation formulae can be derived. Besides ML adaptively trained systems, the incremental Bayesian adaptation can also be applied to simplified discriminative adaptively trained systems where only the model parameters are discriminatively updated with the ML-estimated transforms fixed. The incremental adaptation approaches are evaluated on a conversational telephone speech task. Experiments showed that VB approach can obtain robust performance at the start and overall good performance at the end. The gain of Bayesian adaptation is additive to adaptive training and discriminative training.

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

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

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

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

[6] 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.

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

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

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

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