BAYESIAN DISCRIMINATIVE ADAPTATION FOR SPEECH RECOGNITION

C. K. Raut and M. J. F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{ckr21,mjfg}@eng.cam.ac.uk

ABSTRACT

Linear transform-based speaker adaptation is a standard part of many speech recognition systems. For unsupervised adaptation maximum likelihood estimation is typically used, as discriminative transforms are more heavily biased towards the supervision hypothesis which may contain errors. In this work a Bayesian framework for discriminative adaptation is investigated. This reduces the hypothesis bias and allows robust estimates even with a limited amount of data. Various forms of discriminative maximum-a-posteriori estimation, and associated issues, are detailed. To address these problems, the use of discriminative mapping transforms is also described. The proposed framework is evaluated on an English conversational speech task.

Index Terms— speech recognition, model adaptation, discriminative transforms, maximum-a-posteriori estimation

1. INTRODUCTION

Speaker or environmental adaptation is an important stage for automatic speech recognition systems. Linear transforms are widely used for adapting model parameters in HMM-based systems. For example, the mean $\boldsymbol{\mu}$ of the model parameters is transformed to obtain the speaker-adapted mean $\hat{\boldsymbol{\mu}}^{(s)}$ as

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}^{(s)}\boldsymbol{\mu} + \mathbf{b}^{(s)} = \mathbf{W}^{(s)}\boldsymbol{\xi} \tag{1}$$

where $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)} \ \mathbf{b}^{(s)}]$ is the linear transform for speaker s and $\boldsymbol{\xi} = [\boldsymbol{\mu}^T \ 1]^T$ is the extended mean vector. These transforms are usually estimated by maximising the likelihood of adaptation data, maximum-likelihood linear regression (MLLR) [1].

Discriminative criteria such as minimum phone error (MPE) [2] are commonly used to train HMMs in state-of-the-art systems. Training models with discriminative criteria has been found to reduce word error rate (WER) significantly. Hence, the use of discriminative criteria like MPE has been investigated for transform estimation as well [3]. Though discriminative transforms can give performance gains for supervised adaptation, they are seldom used for unsupervised adaptation for which the correct transcript is not known. This is because discriminative transforms are highly sensitive to errors in the supervision hypothesis and are biased towards it. Though confidence score and lattice based approaches [3, 4] have been investigated to deal with these problems, only limited, if any, gains are obtained. Recently, discriminative mapping transforms (DMTs) [4] have been successfully applied in these situations giving improved performance.

This work was supported in part under the GALE program of the Defence Advanced Research Projects Agency (DARPA), Contract No. HR0011-06-C-0022. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

Another issue with the unsupervised instantaneous adaptation is that it is not normally possible to start adapting the models straightaway. Adaptation must be delayed until robust parameter estimation is achieved. This prevents any adaptation gains for single utterances where there is a limited amount of data. A maximum-a-posteriori (MAP) estimation has been proposed in [5] for robustly estimating MLLR transforms even with a small amount of adaptation data. Similarly, an N-best list based instantaneous unsupervised adaptation scheme has been proposed in [6] that uses MAP estimates of mean bias. The N-best list based scheme can also deal with the errors in the supervision hypothesis. An N-best list based Bayesian framework for MLLR affine transforms has been investigated in [7] for the unsupervised instantaneous adaptation .

In this work a Bayesian approach is investigated for discriminative adaptation. The Bayesian framework can reduce the hypothesis bias and makes the discriminative adaptation less sensitive to supervision hypothesis errors. Moreover, this Bayesian approach allows robust estimation of discriminative transforms even with a limited amount of data. This makes it possible to use them for instantaneously adapting model parameters. After describing maximum-likelihood Bayesian adaptation in the next section, various forms of maximum-a-posteriori (MAP) estimation of discriminative transforms are described. This is followed by a description of the use of discriminative mapping transforms for the Bayesian adaptation.

2. MAXIMUM-LIKELIHOOD BAYESIAN ADAPTATION

The standard approach to unsupervised speaker adaptation is a multi-stage scheme: an initial hypothesis is obtained; transform parameters estimated; and the data re-recognised. An alternative approach to achieve "instantaneous" adaptation is to embed the adaptation transform into the acoustic model, an adaptive HMM [7]. Here, a Bayesian approach is adopted that considers the transform as a random variable and uses *a priori* information for it. In such a system, the best hypothesis $\hat{\mathcal{H}}$ for observation $\mathbf{O}$ is obtained as

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} p(\mathcal{H}|\mathbf{O}) = \arg\max_{\mathcal{H}} \Big\{ p(\mathbf{O}|\mathcal{H}) P(\mathcal{H}) \Big\} \tag{2}$$

where the acoustic score is marginal likelihood given as

$$p(\mathbf{O}|\mathcal{H}) \quad = \quad \int p(\mathbf{O}|\mathcal{H}, \mathbf{W}) p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{ml}}) \, d\mathbf{W}. \tag{3}$$

The transform prior $p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{ml}})$ is assumed to be a Gaussian for mean MLLR transforms. The hyper-parameters $\boldsymbol{\phi}_{\mathtt{ml}}$ of the prior are obtained through an empirical Bayes approach from the point estimates of training transforms (assuming enough data for their robust estimates). With S speakers in the training data set, this is given as

$$\hat{\boldsymbol{\phi}}_{\mathtt{ml}} = \arg\max_{\boldsymbol{\phi}} \sum_{s=1}^{S} \log p(\hat{\mathbf{W}}_{\mathtt{ml}}^{(s)}|\boldsymbol{\phi}). \tag{4}$$

The inference resulting from Equations 2 and 3 is called Bayesian adaptive inference. In standard HMMs, the Viterbi algorithm is used to compute the likelihood of the observation sequence. This relies on the conditional independence assumptions used in HMMs. These are not valid for the adaptive HMM due to additional dependence on transforms. An N-best rescoring framework is therefore used in [7] for the Bayesian adaptive inference. In this N-best rescoring framework the marginal likelihood $p(\mathbf{O}|\mathcal{H})$ is separately computed for each possible hypothesis $\mathcal{H}$ in the N-best list. However, the marginal likelihood as given in Equation 3 is intractable. Different forms of approximations, including variational Bayes lower-bound, can be used. In the maximum-a-posteriori (MAP) approximation [7] the point estimates rather than distributions for the transforms are used and the computation of the marginal likelihood is tractable. The MAP point estimates of ML transforms are obtained as

$$\hat{\mathbf{W}}_{\mathtt{map}}^{(\mathcal{H})} \quad = \quad \arg\max_{\mathbf{W}} \Big\{ \; p(\mathbf{O}|\mathcal{H},\mathbf{W}) p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{ml}}) \; \Big\}. \qquad (5)$$

With the MAP point estimates of the transforms for each possible hypothesis in the N-best list, the best hypothesis is selected as

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}} \Big\{ p(\mathbf{O}|\mathcal{H},\mathbf{W}_{\mathtt{map}}^{(\mathcal{H})}) p(\mathbf{W}_{\mathtt{map}}^{(\mathcal{H})}|\boldsymbol{\phi}_{\mathtt{ml}}) \; P(\mathcal{H}) \Big\}. \qquad (6)$$

This Bayesian approach has been found to yield robust estimates of ML-based transforms for instantaneous adaptation and has led to reductions in WER. Though this framework has been used with discriminatively trained acoustic models, discriminative transform estimates have not been previously investigated. The next section investigates Bayesian discriminative adaptation framework.

3. BAYESIAN DISCRIMINATIVE ADAPTATION

Discriminative adaptation uses discriminative criteria such as maximum mutual information (MMI) or minimum phone error (MPE) to estimate the transform parameters. When used in a supervised adaptation mode, discriminative linear transforms (DLTs) can be robustly estimated and reductions in the error rate obtained [3]. However, this is not the case for unsupervised adaptation, as the supervision hypothesis may contain errors. Discriminatively-estimated transforms are very sensitive to such errors in the supervision hypothesis and are more biased towards the supervision. To deal with these problems, a Bayesian approach is proposed for discriminatively-estimated transforms. The maximum-a-posteriori (MAP) Bayesian estimation of discriminative transforms is described below.

3.1. Discriminative MAP Adaptation

Though the MPE criterion will be used in the experiments, the maximum mutual information (MMI) criterion will be used in this section as it simplifies the description of discriminative MAP adaptation. The MAP estimate of discriminative transforms using the MMI criterion is given as

$$\begin{aligned} \hat{\mathbf{W}}_{\mathtt{dmap}}^{(\mathcal{H})} &= \arg\max_{\mathbf{W}} \Big\{ P(\mathcal{H}|\mathbf{O},\mathbf{W}) p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{d}}) \Big\} \\ &= \arg\max_{\mathbf{W}} \left\{ \frac{p(\mathbf{O}|\mathcal{H},\mathbf{W})P(\mathcal{H})}{\sum_{\breve{\mathcal{H}}} p(\mathbf{O}|\breve{\mathcal{H}},\mathbf{W})P(\breve{\mathcal{H}})} p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{d}}) \right\} \qquad (7) \end{aligned}$$

where $\breve{\mathcal{H}}$ is drawn from all possible hypotheses corresponding to observation $\mathbf{O}$. The estimation of hyper-parameter $\boldsymbol{\phi}_{\mathtt{d}}$ for the discriminative transform prior $p(\mathbf{W}|\boldsymbol{\phi}_{\mathtt{d}})$ turns out to be same as in Equation 4, though using point estimates of training speaker-set discriminative transforms.

The optimisation of the MAP objective function for ML in Equation 5 is straightforward, as a strict-lower bound can be obtained and the EM algorithm used. However, the same is not true for the discriminative MAP objective function in Equation 7. Discriminative objective functions can be optimised using a weak-sense auxiliary function [2] (related to extended Baum-Welch) both for training of HMMs and estimating discriminative transforms. The same approach is investigated to optimise the discriminative MAP objective function. In addition, a lower-bound approach is described.

3.1.1. Weak-Sense Auxiliary Function

The discriminative objective function is usually optimised by defining a weak-sense auxiliary function that has the same gradient at the current parameters as the criterion. The auxiliary function for the discriminative MAP objective function in Equation 7 can be expressed as

$$\begin{aligned} \mathcal{Q}(\boldsymbol{W},\hat{\boldsymbol{W}}) = \; & \mathcal{Q}^{\mathtt{num}}(\boldsymbol{W},\hat{\boldsymbol{W}}) - \mathcal{Q}^{\mathtt{den}}(\boldsymbol{W},\hat{\boldsymbol{W}}) + \mathcal{Q}^{\mathtt{sm}}(\boldsymbol{W},\hat{\boldsymbol{W}}) \\ & + \mathcal{Q}^{\mathtt{p}}(\boldsymbol{W},\hat{\boldsymbol{W}}) \qquad (8) \end{aligned}$$

where $\hat{\mathbf{W}}$ is the current estimate of the transform. The rows of the transforms are assumed to be independent, and the numerator (`num`), the denominator (`den`) and the smoothing (`sm`) terms are expressed in terms of row-wise sufficient statistics $\{\boldsymbol{G}_i^{\mathtt{num/den/sm}}, \boldsymbol{k}_i^{\mathtt{num/den/sm}}\}$ for the ith row of transforms, as given in [3]. The row-wise sufficient statistics corresponding to the prior term $\mathcal{Q}^{\mathtt{p}}(\boldsymbol{W},\hat{\boldsymbol{W}})$ are $\{\boldsymbol{\Sigma}_{w_i}^{-1}, \boldsymbol{\Sigma}_{w_i}^{-1}\boldsymbol{\mu}_{w_i}\}$, where $\boldsymbol{\mu}_{w_i}$ is the mean and $\boldsymbol{\Sigma}_{w_i}$ is the covariance of the prior distribution for the ith row of the transform, $\boldsymbol{w}_i$. With the overall sufficient statistics $\{\boldsymbol{G}_i, \boldsymbol{k}_i\}$ (summation of sufficient statistics of all terms), the MAP estimate of the ith row of the discriminative transform is given as $\boldsymbol{G}_i^{-1}\boldsymbol{k}_i$.

In the weak-sense auxiliary function given in Equation 8, $\mathcal{Q}^{\mathtt{num}}(\boldsymbol{W},\hat{\boldsymbol{W}})$ and $\mathcal{Q}^{\mathtt{den}}(\boldsymbol{W},\hat{\boldsymbol{W}})$ are effectively the lower bounds (LB) of the numerator (excluding the prior term) and the denominator parts of the discriminative objective function. As the LB of the denominator term is subtracted, the resulting expression is not guaranteed to be a lower bound to the discriminative objective function. This implies that maximising the auxiliary function is not guaranteed to maximise the objective function. As the resulting expression may not even be concave, a smoothing term $\mathcal{Q}^{\mathtt{sm}}(\boldsymbol{W},\hat{\boldsymbol{W}})$ is added, which is tunable by a smoothing factor D_m for each component m. With small smoothing factors, the optimisation may diverge, whereas very high values of smoothing factors may not give sufficient update to the parameters. Note that even after adding this smoothing term, the weak-sense auxiliary function is not a lower-bound. This is true when adding the prior term as well.

The weak-sense auxiliary function described above was used to estimate the discriminative MAP transforms for the utterance level adaptation (experimental setup described in Section 4). With the normally used values of smoothing factors, the discriminative MAP objective function was found to generally oscillate with the iterations leading to quite unreliable estimates for the transforms. An ML I-smoothing "prior" and a scale to the transform prior term were also used in the experiments. Though the weak-sense auxiliary function with the I-smoothing "prior" has been found to generally work for discriminative transforms estimation, addition of a discriminative transform prior makes the scenario different. This is because the transform prior term represents the *likelihood of a transform* given the prior distribution, and its nature and dynamic range are different from those of the I-smoothing prior and other terms, specially when

the transform prior is very informative (small variance). On the other hand, the I-smoothing term represents the *likelihood of certain observation points*, and its nature and dynamic range are similar to the numerator term.

It should be noted that other gradient and Hessian based optimisation schemes can be used for discriminative MAP estimation. Like the weak-sense auxiliary function, they are not guaranteed to converge and fine tuning of learning parameters is required. Furthermore, they are generally not elegant and efficient for a large speech recognition system with high dimensionality transform matrices.

3.1.2. Jensen and Reverse-Jensen Inequalities based Lower Bound

Rather than using the weak-sense auxiliary function in the previous section, a strict lower-bound should yield similar attributes to the lower-bounds successfully used with the ML-criterion [7]. Maximising such a lower-bound is guaranteed not to decrease the value of the objective function. However, finding a lower-bound has been problematic for a discriminative objective function due to the denominator term. To obtain an overall lower-bound, the numerator term should be lower-bounded, whereas the denominator term requires an upper-bound.

Obtaining an upper-bound directly on the complete transform denominator term is highly complicated. Instead the reverse-Jensen inequality described in [8] can be used. The complete upper-bound is found by computing an upper-bound for each Gaussian component. This component-specific bound is obtained by exploiting the convexity of the cumulant function of the Gaussian component [8]. With these bounds in place, the auxiliary function can be expressed in the same form as the weak-sense auxiliary function (Equation 8) including the smoothing term [9]. The upper-bound to the denominator term requires computing the appropriate values of smoothing factors. The values of the smoothing factor can be shown to be [8, 9]

$$\begin{aligned} D_m = & \sum_t \gamma_{mt}^{\text{den}} \\ & + \sum_t \max\left[\gamma_{mt}^{\text{den}}\Big(\boldsymbol{o}_t^{\mathsf{T}}(\boldsymbol{\mu}_m\boldsymbol{\mu}_m^{\mathsf{T}} + \boldsymbol{\Sigma}_m)^{-1}\boldsymbol{o}_t - 1\Big), 0\right] \\ & + 4\sum_t f(\gamma_{mt}^{\text{den}}/2)\big(\boldsymbol{o}_t - \boldsymbol{\mu}_m\big)^{\mathsf{T}}\boldsymbol{\Sigma}_m^{-1}\big(\boldsymbol{o}_t - \boldsymbol{\mu}_m\big) \\ & + 4\sum_t f(\gamma_{mt}^{\text{den}}/2)\Big(\big(\boldsymbol{o}_t - \boldsymbol{\mu}_m\big)^{\mathsf{T}}\boldsymbol{\Sigma}_m^{-1}\big(\boldsymbol{o}_t - \boldsymbol{\mu}_m\big) - 1\Big)^2 \end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean and the covariance for mixture m, γ_{mt}^{den} is the denominator occupancy of mixture m at time t, and

$$f(\gamma) = \begin{cases} \gamma + \frac{1}{4\log(6)} + \frac{25/36}{\log(6)^2} - 1/6 & \gamma \geq 1/6 \\ \frac{1}{4\log(1/\gamma)} + \frac{(\gamma-1)^2}{\log(1/\gamma)^2} & \gamma \leq 1/6 \end{cases}. \tag{10}$$

By using the values of smoothing factor as given in Equation 9, an auxiliary function that is a lower-bound to the objective function can be obtained. Using such an auxiliary function, the discriminative objective function can be optimised iteratively by EM algorithm. Any increase in this auxiliary function is guaranteed not to decrease the objective function.

The MAP estimation of discriminative transforms using a lower-bound was evaluated for the utterance-level adaptation experiment described in Section 4. Unfortunately, the values of smoothing factors given by Equation 9 to form a strict lower-bound turn out to be very high. The majority of them are larger by a factor of 10^6 or more than the normally used values of smoothing factors in the weak-sense auxiliary function. This leads to minimal changes in the transform parameters, consequently not altering the rank ordering of the hypotheses. This may be due to the very loose lower-bound obtained for the discriminative objective function. It is known that the bounds obtained with reverse-Jensen's inequality are very loose [8]. Moreover, the transform estimation requires objective function maximisation involving computation of statistics summed over several components. In this case, the cumulative effect of the loose lower bounds may be even more severe than for acoustic model updates. Note in [9] further approximations were used so that the Jensen's reverse inequality based approach gave similar results to a weak-sense auxiliary function for model estimation. However, these approximations are not suitable for this work as a strict lower bound is required. A strict lower-bound of the discriminative objective function obtained by tightly upper-bounding the whole denominator term could possibly improve the optimisation, but is not investigated.

3.2. DMT-based Bayesian Discriminative Adaptation

As seen above, it is difficult to obtain useful MAP estimates of discriminative transforms. However, *robust* estimates of discriminative transforms are crucial for instantaneous unsupervised adaptation. Discriminative mapping transforms (DMTs) [4] can be used for this purpose. Instead of directly estimating discriminative transforms, a DMT maps speaker-specific ML transforms into discriminative ones. The mapping itself is speaker-independent. In this work, the form of DMT used allows the final adapted mean obtained using MLLR-based DMT adaptation (MLLR+DMT) to be expressed as

$$\hat{\boldsymbol{\mu}}^{(s)} = \mathbf{A}_{\mathtt{d}}\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} + \mathbf{b}_{\mathtt{d}} = \mathbf{W}_{\mathtt{d}}\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(s)} \tag{11}$$

where $\hat{\boldsymbol{\xi}}_{\mathtt{ml}}^{(s)} = [\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)T}\ 1]^T$, $\hat{\boldsymbol{\mu}}_{\mathtt{ml}}^{(s)} = \mathbf{W}_{\mathtt{ml}}^{(s)}\boldsymbol{\xi}$ is the MLLR adapted mean, and $\mathbf{W}_{\mathtt{d}} = [\mathbf{A}_{\mathtt{d}}\ \mathbf{b}_{\mathtt{d}}]$ is the DMT transform. As DMTs are speaker-independent, they are estimated from the training data and the same transforms are used while testing. As they are not re-estimated on the test data, they are not sensitive to the supervision hypothesis errors or limited amount of data. The parameters of DMTs are estimated in the same way as discriminative linear transforms (DLTs), however using data from all speakers [4]. Using the MPE criterion, DMT estimation can be expressed as

$$\mathbf{W}_{\mathtt{d}} = \arg\min_{\mathbf{W}}\Big\{\sum_s\sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(s)};\mathbf{W},\mathbf{W}_{\mathtt{ml}}^{(s)},\boldsymbol{\lambda})\mathcal{L}(\mathcal{H},\mathcal{H}_{\mathtt{r}})\Big\} \tag{12}$$

where $\mathcal{L}(\mathcal{H},\mathcal{H}_{\mathtt{r}})$ is the phone-level loss function of the hypothesis $\mathcal{H}$ for the given supervision $\mathcal{H}_{\mathtt{r}}$.

In the Bayesian discriminative adaptation framework, the speaker-independent DMT can be applied to the MAP estimates of MLLR transforms. The best hypothesis is thus selected by using MAP estimates of ML transforms along with the DMT as

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{H}}\Big\{p(\mathbf{O}|\mathcal{H},\mathbf{W}_{\mathtt{map}}^{(\mathcal{H})},\mathbf{W}_{\mathtt{d}})p(\mathbf{W}_{\mathtt{map}}^{(\mathcal{H})}|\boldsymbol{\phi}_{\mathtt{ml}})\ P(\mathcal{H})\Big\}. \tag{13}$$

If the standard form in equation 12 is used then the DMTs are estimated based on speaker-level MAP-transforms. The transform priors in this case will make little difference as the amount of available data will typically be quite large. In contrast for inference using equation 13 MAP estimates are found at the utterance level, where the impact of the priors will be large. This mismatch between DMT training and use in recognition may impact performance. It is possible to estimate DMTs on a per-utterance MAP-estimates (thus removing this mismatch) but this is not done in this work.

DMTs in this section have been described as acting on the MAP transform estimates. They may also be applied to, for example, the

variational Bayes (VB) approximation in [7]. For limited data the VB approximation has been found to yield slightly better performance than MAP-estimates.

4. EXPERIMENTAL RESULTS

The evaluation experiments were conducted on a large vocabulary English conversational telephone speech (CTS) task. The acoustic training data consisted of about 296 hours of speech from 5446 speakers. The test-set eval03 consists of about 6 hours of data from 144 speakers, taken from Swbd and Fisher corpora. The speech data was parameterised using 12 PLP Cepstral coefficients plus the 0th order (C0) coefficient. First, second and third derivatives were also appended. An heteroscedastic linear discriminant analysis transform was used to project this 52-dimensional feature-vector down to 39 dimensions. Cepstral mean and variance as well as vocal tract length normalisation was applied to the features. All systems were based on state-clustered triphone HMMs having 6k distinct states with an average of 16 Gaussian components. A trigram language model trained on 1044M words and a 58k words multiple pronunciation dictionary were used for decoding.

Speaker independent (SI) and speaker adaptive training (SAT) model sets were obtained using ML and MPE criteria. MPE-SAT uses MLLR-transforms estimated using the ML-SAT system, and only model parameters are updated during training. A *mean* transform was used in all experiments for adaptation. All speaker-specific transforms used two base classes: one for speech and another for silence. For the DMT, 1000 regression base classes were used. DMTs are based on the MPE criterion. The supervision hypothesis for adaptation was obtained from the corresponding SI model for ML and MPE systems. All adaptation is done at the utterance level reflecting the scenario of the instantaneous adaptation. The N-best list size was 150 for the rescoring experiments.

System	Adaptation		WER%	
	Training	Testing	ML	MPE
SI	-	-	32.8	29.2
SI	-	MLLR	35.5	32.4
		MAPMLLR	32.2	29.0
		MAPMLLR+DMT	30.8	28.4
SAT	MLLR	MLLR	35.2	32.3
		MAPMLLR	31.8	28.8
		MAPMLLR+DMT	30.9	28.6

Table 1. The WER% with different utterance level N-best adaptation

The experimental results for the utterance-level N-best adaptation are given in Table 1. The MAP estimates of ML transforms, MAPMLLR, reduced the WER significantly. Using DMT with the MAPMLLR N-best adaptation gives a further improvement of 1.4% and 0.9% absolute on the ML SI and SAT systems, respectively, compared to using MAPMLLR alone. Similarly, for the MPE systems, the gains obtained with DMT over MAPMLLR for the utterance-level adaptation is 0.6% and 0.2% absolute on SI and SAT models, respectively. These gains are less than those obtained with DMTs for *speaker level* adaptation. For example, DMTs gave a gain of 0.8% absolute for speaker-level adaptation on MPE systems. The reduction in gains compared to the speaker-level adaptation is felt, in part, to be due to mismatch in applying a DMT estimated from speaker-level ML transforms to the utterance-level MAP estimates of ML transforms. The SAT systems are more affected than the SI systems, as they are more sensitive to any mismatch in the training and the testing transforms than SI systems.

Adaptation	Supervision	
	1-best	N-best
MAPMLLR	32.0	31.8
MAPMLLR+DMT	31.6	30.9

Table 2. A typical performance comparison for the 1-best and the N-best utterance-level adaptation on the ML-SAT system

A comparison of Bayesian N-best adaptation with the 1-best adaptation is given in Table 2, typically for the ML-SAT system. As it can be observed, the N-best adaptation is giving better performance than the 1-best adaptation. Furthermore, the N-best MAPMLLR+DMT adaptation gives a gain of 0.9% absolute compared to using MAPMLLR alone, and a gain of 0.7% absolute compared to the 1-best adaptation using MAPMLLR+DMT.

5. CONCLUSION

This paper has investigated a Bayesian framework for instantaneous unsupervised discriminative adaptation. Discriminative transforms are often biased towards the supervision hypothesis and are very sensitive to errors in the supervision. To handle these problems Bayesian discriminative adaptation is investigated. In contrast to ML-based Bayesian adaptation, lower-bound approximations are not straightforward to define for the discriminative case. Two forms of discriminative MAP approximations are described, though neither were found to yield useful approaches. An alternative scheme based on discriminative mapping transform (DMT) was then described. The N-best rescoring framework for Bayesian discriminative adaptation framework was detailed. Here the speaker-independent DMT is applied over separate MAP estimates of ML transforms for each hypothesis. The technique was evaluated on a large vocabulary English conversational telephone speech task. It was found to outperform standard approaches for the instantaneous adaptation.

6. REFERENCES

[1] 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.

[2] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2003.

[3] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Computer Speech and Language*, vol. 22, no. 3, pp. 256–272, 2008.

[4] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised discriminative adaptation using discriminative mapping transforms," in *Proc. ICASSP*, 2008, pp. 4273–4276.

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

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

[7] K. Yu and M. J. F. Gales, "Bayesian adaptive inference and adaptive training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1932–1943, 2007.

[8] T. Jebara, *Discriminative, Generative and Imitative Learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[9] M. Afify, "Extended Baum-Welch reestimation of Gaussian mixture models based on reverse Jensen inequality," in *Proc. Interspeech*, 2005, pp. 1113–1116.