MPE-BASED DISCRIMINATIVE LINEAR TRANSFORM FOR SPEAKER ADAPTATION

L. Wang and P.C. Woodland

Machine Intelligence Laboratory,
Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {lw256, pcw}@eng.cam.ac.uk

ABSTRACT

In this paper, we present a discriminative method for speaker adaptation, where the minimum phone error (MPE) criterion is used to estimate the discriminative linear transforms (DLTs), including both mean and diagonal variance transforms. The I-smoothing technique is essential to improve the generalization of DLTs. Experiments on supervised adaptation for non-native speakers on the North American Business (NAB) Spoke 3 task show that MPE-based DLT outperforms both MLLR and a previously proposed discriminative method for transform estimation. Preliminary experiments on unsupervised DLT estimation are also reported for conversational telephone speech transcription.

1. INTRODUCTION

Speaker adaptation is crucial to producing a speaker-specific system from a speaker-independent HMM set, given only a small amount of adaptation data. For a variety of adaptation tasks, maximum likelihood linear regression (MLLR) [4] for model-space transformation is an effective and efficient approach. MLLR can use a so-called regression-class tree to adjust the number of generated transforms, according to the amount of adaptation data available. Using the maximum likelihood (ML) criterion to estimate the transform parameters, MLLR can be used to estimate mean transforms, diagonal variance transforms, or full variance transforms for the HMM parameters. MLLR can operate in either supervised or unsupervised mode.

Since discriminative training criteria [11], such as maximum mutual information (MMI) and minimum phone error (MPE) [6, 8] have been successfully used to train HMM-based acoustic models, it is then expected that the same discriminative criteria can benefit the estimation of the linear transforms for both adaptive training [10] and adaptation. In related work [9], the H-criterion was used to estimate discriminative linear transform (DLT), where the H-criterion is an interpolation the of ML and MMI training crteria. The conditional maximum likelihood linear regression [2], which is equivalent to the MMI criterion, has been used for transform generation. Furthermore, the use of minimum classification error (MCE) to estimate the diagonal variance transform has been explored [3]. Apart from the use of unconstrained transforms (different transforms for the means and variances), we have previously also investigated the use of the MPE criterion for constrained (same transform for means and variances) DLT estimation, which can be applied in the feature-specae for discriminative speaker adaptive training [10].

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

Experiments on conversational telephone speech (CTS) transcription for the direct estimation of HMM parameters [11, 6, 8] have demonstrated that MPE training reliably outperforms MMI training on test data, since error driven criteria such as MPE, focus on correctable errors in the training data rather than outliers which may reduce the effectiveness of MMI training. Therefore, in this paper, we concentrate on using the MPE criterion for linear transform estimation, and investigate the use of MPE-based DLT for both supervised and unsupervised speaker adaptation. The derivation for MPE DLT estimation relies on the use of weak-sense auxiliary functions [8]. Furthermore it is necessary to smooth the discriminative statistics with those used for ML estimation. This I-smoothing [6] can also be used to improve the generalization of MPE-based DLT.

The effectiveness of MPE-based DLT estimation is first evaluated in the context of supervised adaptation to non-native speakers from a HMM set trained on native speakers and uses the North American Business News Spoke 3 task. We have then investigated the use of discriminative estimation for unsupervised adaptation, which by the nature of discriminative training techniques is a diffcicult problem.

The rest of this paper is organized as below. In Section 2, we describe the MPE criterion for DLT estimation, including the use of a weak-sense auxiliary function and statistics smoothing technique. The supervised adaptation results on the NAB Spoke 3 task are presented in Section 3. For unsupervised mode adaptation, the experiments are carried out on CTS transcription. Finally, some issues concerning MPE-based DLT for unsupervised adaptation are discussed in the last section.

2. THE MPE CRITERION FOR DISCRIMINATIVE LINEAR TRANSFORM

The MPE criterion was recently proposed for continuous speech recognition training. It aims to maximise an approprixation to the training set phone accuracy, evaluated in a word recognition context. The MPE objective function is defined in [6, 8] as

$$\mathcal{F}_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_{\hat{w}} P_\lambda(\mathcal{O}_r|\mathcal{M}^{\hat{w}})^\kappa P(\hat{w}) RawAccuracy(\hat{w})}{\sum_{w} P_\lambda(\mathcal{O}_r|\mathcal{M}^{w})^\kappa P(w)}, \tag{1}$$

where $\mathcal{M}^w$ is the composite model corresponding to the word sequence w, $P(w)$ is the probability of the word sequence w and κ is

the acoustic scale. The $RawAccuracy(\hat{w})$ measures the number of phones correctly recognized in the sentence $\hat{w}$.

In the implementation of MPE training, lattices marked with time information at HMM level are used to represent both the correct transcriptions and confusable hypotheses from recognition, as in the case of MMI training [11]. Using lattices for discriminative training in this way can significantly reduce the computational load for generating the statistics needed for parameter estimation.

2.1. The Optimization of Discriminative Objective Functions

For the optimization of discriminative criteria, the weak-sense auxiliary function was proposed [7, 8], in contrast to the use of the standard or "strong-sense" auxiliary function used in standard ML training. Given the objective function $\mathcal{F}(\lambda)$, the weak-sense auxiliary function is defined to satisfy the following condition:

$$\frac{\partial}{\partial\hat{\lambda}}\mathcal{G}(\hat{\lambda},\lambda)\Big|_{\hat{\lambda}=\lambda}=\frac{\partial}{\partial\hat{\lambda}}\mathcal{F}(\hat{\lambda})\Big|_{\hat{\lambda}=\lambda}$$

where λ refers to the original parameter set and $\hat{\lambda}$ represents the newly estimated one. This equation implies that if there is a local maximum in the objective function, it must also be a local maximum of the auxiliary function. Although optimizing the weak-sense auxiliary function doesn't guarantee an increase in the objective function, it can still offer the minimum condition for the optimization of $\mathcal{F}(\lambda)$. For discriminative training, the weak-sense auxiliary function provides a feasible approach to optimize the objective functions with negative terms.

For the MPE objective function, the auxiliary function proposed in [7, 8] is then based on the log likelihood of phone arc q, $\log\ p(q)$:

$$\mathcal{G}_{MPE}(\lambda,\hat{\lambda})=\sum_{r=1}^{R}\sum_{q=1}^{Q_r}\frac{\partial\mathcal{F}_{MPE}}{\partial\log\ p(q)}\Big|_{(\lambda=\hat{\lambda})}\log\ p(q). \quad (2)$$

Here each sentence r contains a set of phone arcs $q=1,...,Q_r$, and $p(q)$ represents the likelihood of arc q calculated from the corresponding starting to ending times.

Eq. (2) can be separated into two parts in terms of the positive and negative values of $\frac{\partial\mathcal{F}_{MPE}}{\partial\log\ p(q)}\Big|_{(\lambda=\hat{\lambda})}$, which are analogous to the numerator and denominator terms in the MMI auxiliary function. More important, it is proven that the model parameter updating formulations have the similar forms as those used in MMI training, provided that the numerator/denominator statistics have modified definitions[8].

When using the MPE criterion to estimate linear transforms, a weak-sense auxiliary function is used to derive the optimisation procedure. As in standard MLLR [1], an MPE-based DLT is used to transform the Gaussian means with a matrix $\mathbf{A}$ and a bias $\mathbf{b}$,

$$\tilde{\mu}_m=\mathbf{A}\mu_m+\mathbf{b}=W\xi_m,$$

where $W=[\ \mathbf{b}\ \ \mathbf{A}]$, $\xi_m=[\ 1\ \ {\mu_m}^T]^T$. With the quantity defined for MPE training, $\gamma_q^{MPE}=\frac{1}{\kappa}\frac{\partial\mathcal{F}_{MPE}}{\partial\log\ p(q)}$, the auxiliary function consists of three individual parts, each of which has a Gaussian expression,

$$\begin{aligned}&\mathcal{G}_{MPE}(W,\hat{W})\\&=\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(\gamma_q^{MPE})\log\mathcal{N}(\mathbf{o}(t),\hat{W}\xi_m,\Sigma_m)\\&-\sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(-\gamma_q^{MPE})\log\mathcal{N}(\mathbf{o}(t),\hat{W}\xi_m,\Sigma_m)\\&+\mathcal{G}_{sm}(W,\hat{W})\end{aligned} \quad (3)$$

where $\gamma_{qm}(t)$ is the posterior probability over time t, at state j, mixture component m on condition of arc q. The function $f(\gamma_q^{MPE})$ defined as below determines that the arcs with positive γ_q^{MPE} will be used to accumulate the numerator statistics, while those with negative values will be used to get denominator statistics.

$$\begin{cases} f(\gamma_q^{MPE}) &= \max(0,\ \gamma_q^{MPE})\\ f(-\gamma_q^{MPE}) &= \max(0,\ -\gamma_q^{MPE})\end{cases}$$

The smoothing function in Eq. (3) associates with the initial adapted model parameters to improve the stability of training,

$$\begin{aligned}&\mathcal{G}_{sm}(W,\hat{W})=\sum_m D_m\Big[-\frac{1}{2}\Big(\log|\hat{\Sigma}_m|\\&+(W\xi_m-\hat{W}\xi_m)^T\hat{\Sigma}_m^{-1}(W\xi_m-\hat{W}\xi_m)+\Sigma_m\hat{\Sigma}_m^{-1}\Big)\Big]\end{aligned}$$

Obviously the differential of this smoothing function at $\hat{W}=W$ is zero, so that adding this function ensures that the whole auxiliary function still satisfies the condition for weak-sense definition. And D_m is defined as the smoothing factor with a constant E,

$$D_m=E\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(-\gamma_q^{MPE}).$$

2.2. MPE-based discriminative linear transformation

Calculating the partial differential of Eq. (3) with respect to each row of the linear transform $\hat{\mathbf{w}}^{(i)}$ yields a close solution:

$$\begin{aligned}\hat{\mathbf{w}}^{(i)}&=\mathbf{G}^{(i)^{-1}}\mathbf{k}^{(i)}\\\mathbf{G}^{(i)}&=\sum_m\frac{1}{\sigma_{m(i)}^2}\Big((\gamma_m^{num}-\gamma_m^{den})+D_m\Big)\xi_m^T\xi_m\\\mathbf{k}^{(i)}&=\sum_m\frac{1}{\sigma_{m(i)}^2}\Big(\theta_m^{num}(\mathcal{O}_{(i)})-\theta_m^{den}(\mathcal{O}_{(i)})+D_m\tilde{\mu}_{m(i)}\Big)\xi_m^T\end{aligned} \quad (4)$$

Here, the numerator/denominator statistics to estimate MPE-based DLT have the following forms,

$$\begin{aligned}\gamma_m^{num}&=\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f({\gamma_q}^{MPE})\\\theta_m^{num}(\mathcal{O})&=\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f({\gamma_q}^{MPE})\mathbf{o}(t)\\\theta_m^{num}(\mathcal{O}^2)&=\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f({\gamma_q}^{MPE})\mathbf{o}^2(t),\end{aligned} \quad (5)$$

$$\gamma_m^{den} = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE})$$
$$\theta_m^{den}(\mathcal{O}) = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) \mathbf{o}(t)$$
$$\theta_m^{den}(\mathcal{O}^2) = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) \mathbf{o}^2(t). \quad (6)$$

The smoothing factor $D_m = E\gamma_m^{den}$ is given as in [5], E is a constant between 1 to 2 (selected by empirical results), and $\tilde{\mu}_m$ is the adapted mean vectors with the initial MLLR transform W.

Moreover, we can also derive the diagonal variance transform $\hat{\mathbf{H}}$ under MPE criterion after applying the newly estimated mean transforms $\hat{W}$. As ML-based diagonal variance transform, the MPE-based diagonal variance transform is used to transform the diagonal variance,

$$\hat{\Sigma}_m = \mathbf{H}^T \Sigma_m \mathbf{H}, \qquad \hat{\Sigma}_m^{-1} = \mathbf{L}_m^T \mathbf{H}^{-1} \mathbf{L}_m$$

where $\mathbf{L}_m$ is the Choleski factor of Σ_m^{-1}. Given that

$$\begin{aligned}
\hat{\mathbf{o}}(t) &= \mathbf{o}(t) - \hat{\mu}_m = \mathbf{o}(t) - \hat{W}\xi_m \\
\theta_m^{num}(\hat{\mathcal{O}}^2) &= \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(\gamma_q^{MPE}) \hat{\mathbf{o}}^2(t) \\
\theta_m^{den}(\hat{\mathcal{O}}^2) &= \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) \hat{\mathbf{o}}^2(t)
\end{aligned}$$

the auxiliary function in Eq. (3) could be rewritten as below:

$$\begin{aligned}
&\mathcal{G}(\mathbf{H}, \hat{\mathbf{H}}) = \\
&\sum_{r}^{R} \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(\gamma_q^{MPE}) \log \mathcal{N}(\mathbf{o}(t), \hat{\mu}_m, \hat{\mathbf{H}}^T \Sigma_m \hat{\mathbf{H}}) \\
&- \sum_{r}^{R} \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) \log \mathcal{N}(\mathbf{o}(t), \hat{\mu}_m, \hat{\mathbf{H}}^T \Sigma_m \hat{\mathbf{H}}) \\
&+ \sum_m D_m \Big[-\frac{1}{2} \Big(\log |\hat{\mathbf{H}}|^2 + \tilde{\Sigma}_m \mathbf{L}_m^T \hat{\mathbf{H}}^{-1} \mathbf{L}_m \\
&+ (\tilde{\mu}_m - \hat{\mu}_m)^T \mathbf{L}_m^T \hat{\mathbf{H}}^{-1} \mathbf{L}_m (\tilde{\mu}_m - \hat{\mu}_m) \Big) \Big] \quad (7)
\end{aligned}$$

Solving the partial differential with respect to $\hat{\mathbf{H}}$, we can obtain the MPE-based diagonal variance transform for each element $\hat{\mathbf{h}}_{(i)}$:

$$\begin{aligned}
\hat{\mathbf{h}}_{(i)} &= \frac{\sum_m \Big\{ \frac{1}{\sigma_{m(i)}^2} \Big[\theta_m^{num}(\hat{\mathcal{O}}_{(i)}^2) - \theta_m^{den}(\hat{\mathcal{O}}_{(i)}^2) + D_m Z_{m(i)} \Big] \Big\}}{\sum_m \gamma_m^{num} - \gamma_m^{den} + D_m} \\
Z_{m(i)} &= \tilde{\sigma}_{m(i)}^2 + (\tilde{\mu}_{m(i)} - \hat{\mu}_{m(i)})^2 \quad (8)
\end{aligned}$$

where D_m has the same definition as used in Eq. (3), $\tilde{\mu}_m$ and $\tilde{\sigma}_m^2$ are transformed mean and diagonal variance by the initial MLLR.

2.3. The smoothing technique for MPE-based DLT

The I-smoothing technique was introduced [8] to prevent MPE training from over-training and improve its generalization. The basic idea behind I-smoothing is to incorporate the information from ML statistics as a "prior", so as to smooth the discriminative statistics over each Gaussian component. In our implementation, an extra term $\log P(\hat{W})$ is appended to the auxiliary function in Eq. (3), which is given by ignoring the terms independent of $\hat{W}$:

$$\begin{aligned}
\log P(\hat{W}) = \sum_m \Big[-\frac{1}{2} \Big(&\tau \log |\Sigma_m| \\
&+ \frac{\tau}{\gamma_m^{ml}} \sum_t \gamma_m^{ml}(t) (\mathbf{o}(t) - \hat{W}\xi_m)^T \Sigma_m^{-1} (\mathbf{o}(t) - \hat{W}\xi_m) \Big) \Big] \quad (9)
\end{aligned}$$

where τ points of statistics are coming from ML training. Hence, the numerator statistics to estimate MPE-based DLT will be modified with the occupancy count τ:

$$\begin{aligned}
\gamma_m^{num'} &= \gamma_m^{num} + \tau, \\
\theta_m^{num}(\mathcal{O})' &= \theta_m^{num}(\mathcal{O}) + \frac{\tau}{\gamma_m^{ml}} \theta_m^{ml}(\mathcal{O}), \\
\theta_m^{num}(\mathcal{O}^2)' &= \theta_m^{num}(\mathcal{O}^2) + \frac{\tau}{\gamma_m^{ml}} \theta_m^{ml}(\mathcal{O}^2). \quad (10)
\end{aligned}$$

3. EXPERIMENTS

3.1. Supervised adaptation on WSJ

In our experiments on WSJ, the acoustic models were constructed with ML training on the SI-284 WSJ0+1 training set. The front end used MF-PLP analysis to get the 39-dimensional features, including static cepstra with 1st and 2nd order derivatives. Thus the gender independent cross-word triphone HMMs consist of 6399 tied-states with 12 Gaussians per state. The testing adaptation is performed on 1994 NAB Spoke 3 (s3-dev and s3-eval) task with an enrollment set (40 utterances) and a testing set (about 20 utterances) for each speaker.

The lattice-based framework as used in MPE training is also employed here to estimate MPE-based DLT on the enrollment set. Initially, word lattices are generated by fast decoding on the adapted models (using MLLR after 3 iterations), with a 20K WSJ bigram language model (LM). Then the denominator and numerator phone-level lattices are created by aligning the recognized word lattices and correct transcriptions separately with a unigram LM. The appropriate statistics for MPE-based DLT are accumulated via a forward-backward pass through the lattices marked with the phone starting/ending time. Thus using the regression-class tree with 16 baseclasses for speech and 1 baseclass for silence, mean and diagonal variance transforms are estimated under both MLLR and MPE-based DLT schemes.

To evaluate MPE-based DLT for supervised adaptation, the full decoding with 5K word vocabulary and a bigram LM is operated and then the generated lattices are expanded with a trigram LM for further rescoring. We list the lattice rescoring results after adaptation in Table 1, where H-cri DLT refers to the use of H-criterion for DLT estimation [9]. And M2 means that the word lattices for DLT estimation are re-generated by decoding on the adapted models (using DLT after 3 iterations).

It is observed that MPE-based DLT can improve the supervised adaptation by reducing WER absolute 1.0%, in comparison

Test sets	iterations	MLLR	H-cri DLT	MPE-DLT
s3-dev	1 ite	13.2	12.4	12.2
s3-eval	1 ite	11.1	10.3	10.1
s3-dev	3 ite	12.4	11.9	11.8
s3-eval	3 ite	10.4	10.6	10.1
s3-dev	M2	-	11.9	11.8
s3-eval	M2	-	10.1	10.0

Table 1. The WER(%) on NAB Spoke 3 after MLLR, H-criterion DLT and MPE-based DLT adaptation.

with standard MLLR after 1 iteration. After multiple iterations, MPE-based DLT still outperforms MLLR by decreasing WER absolute 0.6% for s3-dev and 0.3% for s3-eval. It is worth noting that the convergence can be improved when comparing MLLR and DLT after 3 ite. vs. 1 ite.. Moreover, MPE-based DLT performs better than H-criterion DLT, since the smoothing term in the auxiliary function ensures its convergence.

3.2. Unsupervised adaptation on CTS transcription

The acoustic models were built with MPE training on CTS transcription (76 hours training set). Trained with a HLDA front-end, the speaker-independent triphone HMMs contain 5920 tied-states with 12 Gaussian components per state. For testing , half of the official development set for the 2001 NIST evaluation *dev01sub* are used with approximate 3 hours of speech. And lattice rescoring rather than full decoding is operated [10] to evaluate the linear transforms estimated under different criteria.

The 1-best Viterbi output after lattice-MLLR adaptation (5 mean transforms and global full variance transform) and confusion network (CN) decoding is then as the hypothesis to generate the reference phone marked lattices for DLT estimation. The denominator word lattices are created with the bigram LM, which are then aligned to phone level lattices. Therefore, 2 mean and diagonal variance transforms are estimated under MLLR scheme, MMI-based DLT (the estimation formulations are identical to that in [2]) and MPE-based DLT scheme individually. The cheating results, where the true transcriptions are used to generate phone level reference lattices for DLT estimation, are also summarized in the following table.

Adaptation	hypothesis		true trans
		+CN (27.0)	
MLLR	27.7	27.0	26.1
MMI-DLT	27.5	26.8	24.8
MPE-DLT	27.3	26.9	23.2

Table 2. The WER(%) on $dev01sub$ for MPE system, after MLLR, MMI-based DLT and MPE-based DLT adaptation.

It can be seen that MPE-based DLT yields absolute 0.4% WER gain over MLLR, and 0.2% gain over MMI-based DLT. After CN decoding, MPE-based DLT gives 0.1% decrease in WER compared with that of the hypothesis. Although MMI-based DLT yields more 0.1% gain over MPE-based DLT after CN decoding, we believe that MPE-based DLT could perform better than MLLR and MMI-based DLT which is proven from the cheating column. This aspect implies that MPE-based DLT can be developed for unsupervised adaptation with further work.

4. DISCUSSIONS AND CONCLUSIONS

This paper has investigated using the MPE criterion for DLT estimation, which can be applied to both supervised and unsupervised adaptations. With the presented weak-sense auxiliary function, the estimation formulations for MPE-based DLT have been derived, where I-smoothing is used to prevent over-training. The experimental results on WSJ task have shown that MPE-based DLT can considerably improve the supervised adaptation performance.

Our ongoing investigations focus on using MPE-based DLT in the unsupervised style. To effectively use the hypothesis for MPE-based DLT estimation, the confidence scores from CN outputs can be used to accumulate the statistics with high confidence. Alternatively, with the numerator lattices, the most likely lattices can be weighted by posterior probabilities to estimate DLT.

5. REFERENCES

[1] M.J.F. Gales (1998) "Maximum Likelihood Linear Transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98.

[2] A. Gunawardana & W. Byrne (2001). "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression ," in *Proc. Eurospeech'01*, Scandinavia.

[3] X.D. He & W. Chou (2003), "Minimum Classfication Error (MCE) Model Adaptation of Continuous Density HMMs, " *Proc. Eurospeech'03*, Geneva.

[4] C.J. Leggetter & P.C Woodland (1995). "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *Computer Speech & Language*, Vol. 9, pp. 171-185.

[5] J. McDonough, T. Schaaf & A. Waibel (2002). "On Maximum Mutual Information Speaker-Adaptation Training, " *Proc. ICASSP'02*, Orlando.

[6] D. Povey & P.C. Woodland (2002). "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," *Proc. ICASSP'02*, Orlando.

[7] D. Povey, M.J.F. Gales, D.Y. Kim, & P.C. Woodland (2003). "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," *Proc. Eurospeech'03*, Geneva.

[8] D. Povey (2004). "Discriminative Training for Large Vocabulary Speech Recognition," *Ph. D. Dissertation, Department of Engineering, University of Cambridge, U.K.*

[9] L.F. Uebel & P.C. Woodland (2001). "Discriminative Linear Transforms for Speaker Adaptation," *Proc. ISCA ITRW on Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis.

[10] L. Wang & P.C. Woodland (2003), "Discriminative Adaptive Training Using The MPE Criterion, " *Proc. ASRU 2003*, St. Thomas, US Virgin Islands.

[11] P.C. Woodland & D. Povey (2002), "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47.