DISCRIMINATIVE ADAPTIVE TRAINING USING THE MPE CRITERION

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

This paper addresses the use of discriminative training criteria for Speaker Adaptive Training (SAT), where both the transform generation and model parameter estimation are estimated using the Minimum Phone Error (MPE) criterion. In a similar fashion to the use of I-smoothing for standard MPE training, a smoothing technique is introduced to avoid over-training when optimizing MPE-based feature-space transforms. Experiments on a Conversational Telephone Speech (CTS) transcription task demonstrate that MPE-based SAT models can reduce the word error rate over non-SAT MPE models by 1.0% absolute, after lattice-based MLLR adaptation. Moreover, a simplified implementation of MPE-SAT with the use of constrained MLLR, in place of MPE-estimated transforms, is also discussed.

1. INTRODUCTION

For speech recognition tasks, such as conversational telephone speech (CTS) transcription, with a large amount of variability in the training data (due to e.g. speakers), adaptive training can be used to remove some of that variation by estimating a set of adaptation transforms for each speaker or acoustic condition along with a canonical model of speech given those transforms.

Speaker Adaptive Training (SAT) applies speaker-specific training-set transforms in the HMM parameter optimization procedure [1, 3] to improve the speaker-independent acoustic models (the canonical models). In general, each iteration of estimating a SAT canonical HMM set requires two sequential steps: the speaker-specific transforms are first generated with the current HMM parameters, and then the canonical HMM set parameters are re-estimated after applying those transforms in the feature or model space. SAT training was originally developed [1] for unconstrained Maximum Likelihood Linear Regression (MLLR) [6], but this requires a complex training procedure and severe memory overheads [4]. A simpler formulation of SAT uses constrained MLLR [4] since those transforms can be applied to the features directly and this makes the parameter re-estimation of the canonical HMM considerably more straightforward.

Recently it has been shown that improved performance for large vocabulary tasks can be obtained by using discriminative training criteria [13] such as Maximum Mutual Information (MMI) and Minimum Phone Error (MPE) [9], so most state-of-the-art large-vocabulary recognition systems use discriminative training. Particularly MPE reduces the training set estimated phone error (in a word recognition context) and has been shown to outperform MMI on CTS transcription. However, current speaker adaptation techniques for SAT, are still normally based on ML estimation, and it is therefore interesting to explore discriminative SAT where a discriminative training criterion is used for both linear transform optimization and model parameter re-estimation. It is also expected that the use of discriminative criteria can improve adaptive training as some previous work on this topic have suggested [5, 8, 2].

In this paper, we concentrate on the use of the MPE criterion for discriminative SAT, in particular for the estimation of speaker-specific transforms called constrained Discriminative Linear Transform (DLT), which can be applied to feature-space in the same way as those used in constrained MLLR. The use of weak-sense auxiliary functions [10] provides a method to derive estimation formula for MPE-based constrained DLT, where a smoothing technique is used to prevent over-training. The statistics to estimate the linear transforms are accumulated for each "baseclass" of the regression-class tree [7], where a group of Gaussian components belonging to that baseclass share the same adaptation transformations. The second step of discriminative SAT is to re-estimate the model parameters where the Extended Baum-Welch (EBW) algorithm is used with the observation vectors adapted by the constrained DLT. For comparison, we also use a simplified but practical implementation of discriminative SAT in the CTS experiments, where linear transforms are still estimated under the ML criterion, and the HMM sets are optimized using a discriminative training criterion.

The rest of this paper is organized as below. In Section 2, we describe the MPE criterion for constrained DLT estimation, including the use of a weak-sense auxiliary function and statistics smoothing. Then, experiments on CTS transcription are presented in Section 3, and the results from various types of discriminative SAT models are given when testing with unsupervised test-set adaptation. In the last section, some issues concerning MPE-based discriminative SAT are discussed.

2. MPE CRITERION FOR DISCRIMINATIVE ADAPTIVE TRAINING

The MPE criterion was recently developed as a novel discriminative training method for continuous speech recognition. It takes into account the mis-classification of HMM models, by measuring the phone transcription accuracy in a word recognition context. The objective function of the MPE criterion, proposed in [9, 10] is

To appear in *Proc. ASRU 2003*

as below:

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

where $\mathcal{M}^{\hat{w}_r}$ is the composite model corresponding to the word sequence w_r, $P(\hat{w}_r)$ is the probability of the word sequence w_r and κ is the acoustic scale. The $RawAccuracy(\hat{w})$ measures the accuracy of hypothesis $\hat{w}$.

2.1. Weak-Sense Auxiliary Function

The idea of a weak-sense auxiliary function [10] was introduced for the optimization of discriminative criteria, in contrast to the use of the standard strong-sense auxiliary function [10] for ML training. Given the objective function $\mathcal{F}(\lambda)$, the weak-sense auxiliary function is defined to satisfy the following condition:

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

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 for the optimization for objective functions with negative terms. This idea can be used to define the auxiliary function for the MMI criterion [10], which consists of three individual parts.

$$\mathcal{G}(\lambda,\hat{\lambda}) = \mathcal{Q}^{num}(\lambda,\hat{\lambda}) - \mathcal{Q}^{den}(\lambda,\hat{\lambda}) + \mathcal{Q}_{sm}(\lambda,\hat{\lambda}) \tag{2}$$

The superscripts num and den in Eq. (2) correspond to the numerator (correct transcription) and the denominator (all possible transcriptions) of the MMI objective function [13, 10]. Each term in Eq. (2) is defined as a Gaussian expression:

$$\mathcal{Q}(\lambda,\hat{\lambda}) = \sum_{j,m}\sum_{t}\gamma_{jm}(t)\log\mathcal{N}(\mathbf{o}(t),\hat{\mu}_{jm},\hat{\Sigma}_{jm})$$

where $\gamma_{jm}(t)$ is the posterior probability at time t for state j mixture component m; mean $\hat{\mu}_{jm}$ and covariance $\hat{\Sigma}_{jm}$ refer to the new parameter estimates $\hat{\lambda}$. The third term $\mathcal{Q}_{sm}(\lambda,\hat{\lambda})$ is a smoothing term related to the original model parameters to improve the stability of training, which is given in the diagonal covariance case:

$$\mathcal{Q}_{sm}(\lambda,\hat{\lambda}) = \sum_{j,m} D_{jm}\left[-\frac{1}{2}\left(\log(|\hat{\Sigma}_{jm}|) + \frac{(\mu_{jm}-\hat{\mu}_{jm})^2+\sigma_{jm}^2}{\hat{\sigma}_{jm}^2}\right)\right]$$

Obviously the differential of above equation at $\hat{\lambda}=\lambda$ is zero, so that adding this smoothing term can still ensure Eq. (2) is a weak-sense auxiliary function. The smoothing factor D_{jm} is defined as $E\sum_t\gamma_{jm}^{den}(t)$ with the value E used to keep the covariance estimate positive.

In this paper, we extend the use of weak-sense auxiliary functions to the optimization of speaker-specific transforms under both the MMI and MPE critera. As for constrained MLLR, the MMI-based constrained DLT will be applied to both means and variances so as to perform in the feature-space:

$$\hat{\mathbf{o}}(t) = \hat{\mathbf{A}}\mathbf{o}(t) + \hat{\mathbf{b}} = \hat{W}\zeta(t)$$

with $\hat{W} = [\ \hat{\mathbf{b}}^T\ \ \hat{\mathbf{A}}^T]^T$, $\zeta(t) = [\ 1\ \ \mathbf{o}(t)^T]^T$ for each speaker. To optimize the linear transform $\hat{W}$ under the MMI criterion, the HMM parameters are fixed and the auxiliary function is defined according to Eq. (2), by ignoring the terms not containing $\hat{W}$:

$$\begin{aligned}\mathcal{G}(W,\hat{W}) &= \sum_{j,m}\sum_{t}\gamma_{jm}^{num}(t)\log\mathcal{N}(\hat{W}\zeta(t),\mu_{jm},\sigma_{jm}^2)\\ &-\sum_{j,m}\sum_{t}\gamma_{jm}^{den}(t)\log\mathcal{N}(\hat{W}\zeta(t),\mu_{jm},\sigma_{jm}^2)\\ &+\sum_{j,m}D_{jm}\Big[-\frac{1}{2}\Big(\log(|\Sigma_{jm}|-\log|\hat{\mathbf{A}}|^2\\ &+(\hat{W}\tilde{\xi}_{jm}-\mu_{jm})^T\Sigma_{jm}^{-1}(\hat{W}\tilde{\xi}_{jm}-\mu_{jm})+\hat{\mathbf{A}}\tilde{\Sigma}_{jm}\hat{\mathbf{A}}^T\Sigma_{jm}^{-1}\Big)\Big]\end{aligned} \tag{3}$$

Notice that the above unadapted model parameters have been transformed with W (typically W would be initially estimated by constrained MLLR):

$$\begin{aligned}\tilde{\xi}_{jm} &= \begin{bmatrix}1\\ \tilde{\mu}_{jm}\end{bmatrix} = \begin{bmatrix}1\\ \mathbf{A}^{-1}\left(\mu_{jm}-\mathbf{b}\right)\end{bmatrix}\\ \tilde{\Sigma}_{jm} &= \mathbf{A}^{-1}\Sigma_{jm}\mathbf{A}^{-1^T}\end{aligned}$$

Given that $\hat{\mathbf{w}}_i$ is the ith row of $\hat{W}$, and $\mathbf{p}_i$ is the extended cofactor row vector $[0\ \ c_{i1}\ \cdots\ c_{in}]$, ($c_{ij}$ =cof($\hat{\mathbf{A}}_{ij}$)), the accumulators to estimate the linear transforms can be derived by calculating the differential with respect to $\hat{\mathbf{w}}_i$:

$$\beta\frac{\mathbf{p}_i}{\mathbf{p}_i\mathbf{w}_i^T} - \mathbf{w}_i\mathbf{G}^{(i)} + \mathbf{k}^{(i)} = 0 \tag{4}$$

where,

$$\begin{aligned}\beta &= \sum_{j,m}\left(\gamma_{jm}^{num}-\gamma_{jm}^{gen}\right)+D_{jm}\\ \mathbf{G}^{(i)} &= \sum_{j,m}\frac{1}{\sigma_{jm}^{(i)\,2}}\left(\theta_{jm}^{num}(\zeta\zeta^T)-\theta_{jm}^{den}(\zeta\zeta^T)+D_{jm}Z_{jm}\right)\\ Z_{jm} &= \begin{bmatrix}1 & \tilde{\mu}_{jm}^T\\ \tilde{\mu}_{jm} & \tilde{\Sigma}_{jm}+\tilde{\mu}_{jm}\tilde{\mu}_{jm}^T\end{bmatrix}\\ \mathbf{k}^{(i)} &= \sum_{j,m}\frac{\mu_{jm}^{(i)}}{\sigma_{jm}^{(i)\,2}}\left(\theta_{jm}^{num}(\zeta)-\theta_{jm}^{den}(\zeta)+D_{jm}\begin{bmatrix}1\\ \tilde{\mu}_{jm}\end{bmatrix}\right)\end{aligned} \tag{5}$$

with the statistics:

$$\begin{aligned}\gamma_{jm} &= \sum_t\gamma_{jm}(t)\\ \theta_{jm}(\zeta) &= \sum_t\gamma_{jm}(t)\zeta(t)\\ \theta_{jm}(\zeta\zeta^T) &= \sum_t\gamma_{jm}(t)\zeta(t)\zeta(t)^T\end{aligned} \tag{6}$$

It can be observed that above equations are the same as appeared in [2], which used Conditional Maximum Likelihood (equivalent to MMI) to deduce the discriminative likelihood linear transform for feature normalization. Since Eq. (4) has the same form as that in constrained MLLR but different accumulators $\mathbf{G}^{(i)}$ and $\mathbf{k}^{(i)}$ [4], the iterative solution for constrained MLLR is also used here to estimate constrained DLT matrices on a row-by-row basis.

For the second stage of discriminative SAT, the auxiliary function in Eq. (2) with adapted observation $\hat{\mathbf{o}}(t)$ can be maximized by setting the differential with respect to $\hat{\mu}_{jm}$ or $\hat{\sigma}_{jm}$ to zero. Therefore, similar updating formulae for the model mean and diagonal covariances as for standard MMI estimation [13] are obtained,

$$\begin{aligned}\hat{\mu}_{jm} &= \theta_{jm}^{num}(\hat{\mathcal{O}}) - \theta_{jm}^{gen}(\hat{\mathcal{O}}) + \\ \hat{\sigma}_{jm}^2 &= \frac{\theta_{jm}^{num}(\hat{\mathcal{O}}^2) - \theta_{jm}^{gen}(\hat{\mathcal{O}}^2) + D_{jm}(\sigma_{jm}^2 + \mu_{jm}^2)}{\{\gamma_{jm}^{num} - \gamma_{jm}^{gen}\} + D_{jm}} - \hat{\mu}_{jm}^2\end{aligned} \tag{7}$$

where the statistics are accumulated with the transformed observation for each instant rather than the original observation.

$$\begin{aligned}\gamma_{jm} &= \sum_{r}^{R}\sum_{t}^{T_r}\gamma_{jm}(t) \\ \theta_{jm}(\hat{\mathcal{O}}) &= \sum_{r}^{R}\sum_{t}^{T_r}\hat{\mathbf{o}}_r(t)\gamma_{jm}(t) \\ \theta_{jm}(\hat{\mathcal{O}}^2) &= \sum_{r}^{R}\sum_{t}^{T_r}\hat{\mathbf{o}}_r^2(t)\gamma_{jm}(t)\end{aligned} \tag{8}$$

2.2. Discriminative SAT using the MPE criterion

In discriminative training for large vocabulary tasks, the evaluation of the MMI (& MPE) "denominator" term is typically computed using an approximate lattice representation which is computed once and then assumed to represent all confusable hypotheses. Lattices consist of nodes (annotated with starting and ending time) and arcs (words connecting nodes and annotated with language model scores). The use of lattices reduces the computational load for generating the statistics needed for parameter estimation [13].

Since the MPE criterion involves the phone accuracy in the objective function, the auxiliary function proposed in [10] is then based on the log likelihood of phone arcs to make the optimization of MPE criterion tractable,

$$\mathcal{G}_{MPE}(\lambda,\hat{\lambda}) = \sum_{r=1}^{R}\sum_{q=1}^{Q_r}\left.\frac{\partial\mathcal{F}_{MPE}}{\partial\log\ p(q)}\right|_{(\lambda=\hat{\lambda})}\log\ p(q) \tag{9}$$

here each sentence r contains of 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.

The above equation can be separated into two parts in terms of the positive and negative values of $\left.\frac{\partial\mathcal{F}_{MPE}}{\partial\log\ p(q)}\right|_{(\lambda=\hat{\lambda})}$, which are analogous to the numerator and denominator terms in the MMI auxiliary function in Eq. (2). More importantly, it can be shown that EBW updating formulae can be used in the same way as for MMI estimation to optimize the model parameters in MPE training, provided that the numerator/denominator statistics have different definitions[10].

Thus we can also design a weak-sense auxiliary function to solve the optimization of MPE-based constrained DLT. 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 can be written as below:

$$\begin{aligned}&\mathcal{G}_{MPE}(W,\hat{W}) \\ &= \sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE})\log\mathcal{N}(\hat{W}\zeta(t),\mu_{jm},\sigma_{jm}^2) \\ &- \sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE})\log\mathcal{N}(\hat{W}\zeta(t),\mu_{jm},\sigma_{jm}^2) \\ &+ \mathcal{G}_{sm}(W,\hat{W})\end{aligned} \tag{10}$$

where $\gamma_{qjm}(t)$ is the posterior probability over time t, at state j, mixture component m on condition of arc q. The function $f(\gamma_q^{MPE}) = \max(0,\gamma_q^{MPE})$ 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. The smoothing fuction in above equation $\mathcal{G}_{sm}(W,\hat{W})$ has the same expression as in Eq. (3) with the factor

$$D_{jm} = E\sum_{t}\gamma_{qjm}(t)f(-\gamma_q^{MPE}).$$

Therefore, the accumulators in Eq. (5-6) for MMI-based constrained DLT can be used for the estimation of MPE-based constrained DLT, with the different numerator statistics:

$$\begin{aligned}\gamma_{jm}^{num} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE}) \\ \theta_{jm}^{num}(\zeta) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE})\zeta(t) \\ \theta_{jm}^{num}(\zeta\zeta^T) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE})\zeta(t)\zeta(t)^T\end{aligned} \tag{11}$$

and denominator statistics:

$$\begin{aligned}\gamma_{jm}^{den} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE}) \\ \theta_{jm}^{den}(\zeta) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE})\zeta(t) \\ \theta_{jm}^{den}(\zeta\zeta^T) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE})\zeta(t)\zeta(t)^T\end{aligned} \tag{12}$$

Thus the row-by-row optimization for constrained MLLR is also a practical solution to estimate MPE-based constrained DLT matrices. After generating linear transforms for each speaker, the acoustic model parameters can be re-estimated according to the updating formula in Eq. (7), where the transformed observations

are used in MPE training with the modified numerator statistics:

$$\begin{aligned}
\gamma_{jm}^{num} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE}) \\
\theta_{jm}^{num}(\hat{\mathcal{O}}) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE})\ \hat{\mathbf{o}}(t) \\
\theta_{jm}^{num}(\hat{\mathcal{O}}^2) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(\gamma_q^{MPE})\ \hat{\mathbf{o}}^2(t) \quad (13)
\end{aligned}$$

and denominator statistics:

$$\begin{aligned}
\gamma_{jm}^{den} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE}) \\
\theta_{jm}^{den}(\hat{\mathcal{O}}) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE})\ \hat{\mathbf{o}}(t) \\
\theta_{jm}^{den}(\hat{\mathcal{O}}^2) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qjm}(t)f(-\gamma_q^{MPE})\ \hat{\mathbf{o}}^2(t) \quad (14)
\end{aligned}$$

2.3. The smoothing technique for MPE-based discriminative SAT

The I-smoothing technique was introduced [10] to make MPE training converge without over-training and improve the generalization. The basic idea of I-smoothing is to incorporate the information from ML statistics as a "prior" to smooth the discriminative statistics over each component. The implementation adds an extra term $\log P(\lambda)$ in the auxiliary function:

$$\begin{aligned}
\log P(\hat{\lambda}) &= \sum_{j,m}\Big[-\frac{1}{2}\Big(\tau\ \log(|\hat{\Sigma}_{jm}|) \\
&+ \frac{\tau}{\gamma_{jm}^{ml}}\frac{\sum_t \gamma_{jm}^{ml}(t)(\mathbf{o}(t)-\hat{\mu}_{jm})^2}{\hat{\sigma}_{jm}^2}\Big)\Big] \quad (15)
\end{aligned}$$

with τ points of statistics from ML training.

Motivated by I-smoothing, we present the smoothing technique for the optimization of MPE-based constrained DLT. The statistics from ML estimation will be added to the auxiliary function in Eq. (10) by ignoring the terms independent of $\hat{W}$:

$$\log P(\hat{W}) = \sum_{j,m}\Big[-\frac{1}{2}\Big(\tau(-\log|\hat{\mathbf{A}}|^2) + \frac{\tau}{\gamma_{jm}^{ml}}\mathcal{Q}(\hat{W})\Big)\Big]$$

$$\mathcal{Q}(\hat{W}) = \sum_t \gamma_{jm}^{ml}(t)(\hat{W}\zeta(t)-\mu_{jm})^T\Sigma_{jm}^{-1}(\hat{W}\zeta(t)-\mu_{jm}) \quad (16)$$

Since the accumulators for MPE-based constrained DLT are all calculated at the baseclass level to avoid memory overload, we then present baseclass smoothing for MPE-based constrained DLT optimization with the occupancy count τ:

$$\beta' = \beta + \tau$$

$$\begin{aligned}
\mathbf{G}^{(i)'} &= \mathbf{G}^{(i)} + \frac{\tau}{\sum_{jm}\gamma_{jm}^{ml}}\sum_{jm}\frac{1}{\sigma_{jm}^{(i)\,2}}\sum_t \gamma_{jm}^{ml}(t)\zeta(t)\zeta(t)^T \\
\mathbf{k}^{(i)'} &= \mathbf{k}^{(i)} + \frac{\tau}{\sum_{jm}\gamma_{jm}^{ml}}\sum_{jm}\frac{\mu_{jm}^{(i)}}{\sigma_{jm}^{(i)\,2}}\sum_t \gamma_{jm}^{ml}(t)\zeta(t) \quad (17)
\end{aligned}$$

3. EXPERIMENTS ON CONVERSATIONAL TELEPHONE SPEECH TRANSCRIPTION

The training set for our experiments contains 1118 conversation sides (76 hours) of speech from the Switchboard I, Call Home English and Switchboard Cellular corpora. The official development set for the 2001 NIST evaluation *dev01* is used as the test-set, which contains approximately 6 hours speech (118 conversational sides) from the Switchboard I (SW-I), Switchboard II (SW-II) and Cellular (Cell) data.

3.1. Experimental Setup

The acoustic models used in our experiments are gender independent continuous mixture density, tied state cross-word triphone HMMs. Each frame of speech has MF-PLP analysis applied to get the static cepstra with 1st, 2nd and 3rd order derivatives, and then a HLDA feature matrix is used to project the 52-dimensional feature vector to 39-dimensions. Vocal tract length normalization (VTLN) analysis is also applied in both training and testing.

Thus, the basic HMM sets consist of 5920 tied-states, each of which has 12 Gaussian components. Starting from HLDA-ML models, a single constrained MLLR transform is generated for each speaker, and then the observations are transformed to estimate the initial ML-SAT model parameters. At each iteration, the linear transforms are updated based on the current ML-SAT model, and the final ML-SAT model after 5 iterations is then used as the seed model for the further discriminative SAT.

The construction of discriminative SAT models relies on the lattice-based framework as used in previous work on MMI/MPE training. Word lattices are initially generated with an adapted HLDA-ML-SAT model, by fast decoding with a pruned bigram language model for all training segments. Then the denominator and numerator phone-marked lattices are created by aligning the recognized word lattices and true transcriptions separately. The appropriate statistics for the discriminative SAT are accumulated via a forward-backward pass through the lattice constructed by the phone boundary times.

Seeding on the HLDA-ML-SAT model and constrained MLLR transforms for each speaker, we then estimate MMI-based constrained DLT and MPE-based constrained DLT respectively, where three matrices are generated for each conversational side using the regression class tree. Those feature-space transforms are then fixed and applied to adapt observation vectors for the re-estimation of HMM sets, so as to construct MMI-SAT and MPE-SAT systems (8 iterations). For comparison, a simplified implementation of discriminative SAT is also tested, where three constrained MLLR matrices using the regression class tree are applied during re-estimating model parameters under the MMI/MPE schemes. We list below the discriminative SAT models that are considered in this paper:

	transform generation/ parameter re-estimation
MMI-SAT(+CMLLR)	constrained MLLR/MMI
MMI-SAT(+MMI_CDLT)	MMI-based constrained DLT/MMI
MPE-SAT(+CMLLR)	constrained MLLR / MPE
MPE-SAT(+MMI_CDLT)	MMI-based constrained DLT/MPE
MPE-SAT(+MPE_CDLT)	MPE-based constrained DLT/MPE

In testing, full decoding is conducted with the adapted HLDA-MPE triphone models and bigram language model (LM), the generated lattices are then expanded using a 4-gram LM. So lattice rescoring rather than full decoding is performed with these expanded lattices for all SAT systems. Consequently, the 1-best output from lattice generation is also used as the supervision information for the testing adaptation.

The unsupervised test-set adaptation is a sequential process as illustrated in Fig. 1. To adapt discriminative SAT models, constrained MLLR transforms are first estimated and then lattice-based MLLR [11] is applied, where transforms are estimated in an iterative way. In Fig. 1, "INPUT" indicates that these transforms are applied to the feature/model space when calculating the occupancies for new linear transforms, while "PARENT" means those transforms are also applied when accumulating the statistics for new linear transforms. Hence the "PARENT" transforms should be combined with newly estimated transforms for the further application.

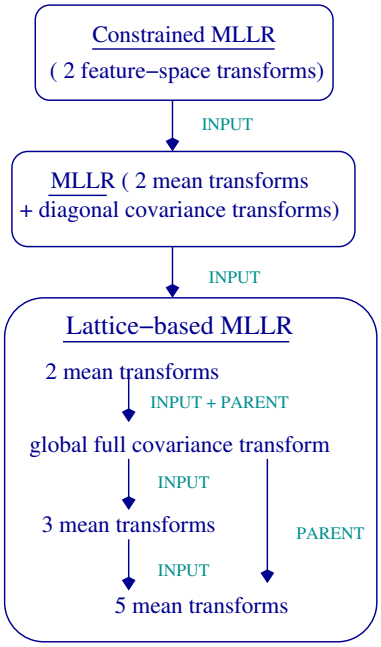

Fig. 1 The testing adaptation used for discriminative SAT

3.2. Experimental results

Table 1 shows the rescoring results for test set $dev01$ with both baseline MMI/MPE models and discriminative SAT models, where 1-best constrained MLLR as shown in the first box of Fig. 1 is used for test-set adaptation. The separate columns indicate the WERs (%) for three subsets of $dev01$.

It is observed that MMI-SAT system with MMI-based constrained DLT could reduce the WER over the MMI system by 0.5% absolute, while MPE-SAT system with MPE-based constrained DLT could also improve the performance by 0.8% absolute compared to standard MPE training.

Systems	SW-I	SW-II	Cell	total
MMI	21.1	33.4	33.1	29.2
MMI-SAT(+MMI_CDLT)	20.3	32.9	32.6	28.6
MPE	20.2	33.0	32.7	28.6
MPE-SAT(+MPE_CDLT)	20.1	31.8	31.8	27.8

Table 1. The WER(%) on test set $dev01$ for MMI/MPE systems and discriminative SAT systems, after constrained MLLR adaptation.

Note that the rescoring results of the MMI and MMI-SAT models maybe optimistic (the gain between MPE and MMI should be a little bigger), since the adaptation supervision comes from decoding with an adapted MPE model. So lattice-based MLLR is a fairer way to evaluate the performance of systems trained under MMI framework, because the effect of adaptation supervision is reduced.

Furthermore, we evaluate MMI-SAT systems trained with ML-based linear transforms and MMI-based linear transforms respectively. In Fig. 2, the MMI criterion values on each iteration during training have been plotted for both MMI-SAT (+CMLLR) and MMI-SAT(+MMI_CDLT) systems. Here, MMI-SAT with MMI-

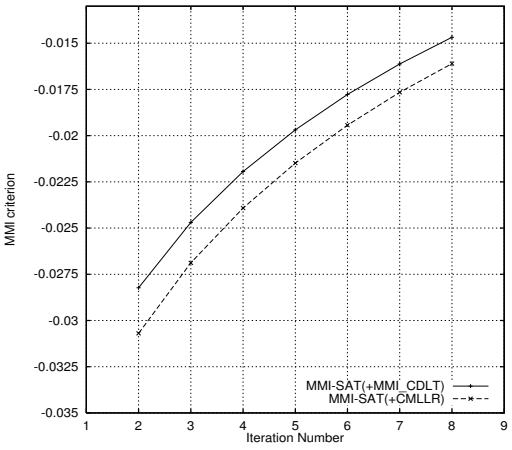

Fig. 2 The MMI criterion value on each iteration during the training of MMI-SAT systems

based constrained DLT is effective in improving the MMI objective function, compared with the simplified implementation of MMI-SAT using constrained MLLR. The WERs(%) for these two MMI-SAT systems on test set $dev01$, with standard 1-best MLLR and lattice-based MLLR for testing adaptation are given in Table 2. It can be seen that there are small improvements with the use of MMI-based constrained DLT for MMI-SAT.

Systems	MLLR	lattice MLLR
MMI	29.1	28.6
MMI-SAT(+CMLLR)	28.5	28.0
MMI-SAT(+MMI_CDLT)	28.5	27.9

Table 2. The WER(%) on $dev01$ for MMI and MMI-SAT systems, after MLLR and lattice-based MLLR adaptation.

Next we explore the performance of MPE-SAT models trained with different types of linear transforms. In Fig. 3, the aver-

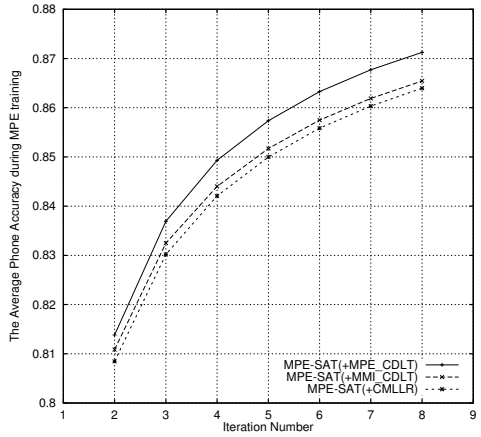

Fig. 3 The average phone accuracy on each iteration for the training of MPE-SAT models.

age phone accuracies during MPE training illustrate that applying MPE-based constrained DLT could make MPE training converge faster and achieve higher phone accuracy, in comparison with constrained MLLR for MPE-SAT.

The rescoring results on test set $dev01$, as shown in Table 3, indicate that MPE-SAT with MPE-based constrained DLT here outperform MPE-SAT with constrained MLLR, but the gain is only 0.1% after lattice-based adaptation. Furthermore, the use of MMI-based constrained DLT for building an MPE-SAT model could also get similar test-set performance.

Systems	MLLR	lattice MLLR
MPE	28.5	27.9
MPE-SAT(+CMLLR)	27.8	27.0
MPE-SAT(+MMI_CDLT)	27.8	26.9
MPE-SAT(+MPE_CDLT)	27.8	26.9

Table 3. The WER(%) on $dev01$ for MPE and MPE-SAT systems, after MLLR and lattice-based MLLR adaptation.

4. DISCUSSIONS AND CONCLUSIONS

This paper has investigated the use of the MPE criterion for discriminative SAT and its application to the CTS task. Using the weak-sense auxiliary function, the estimation formulae for MPE-based constrained DLT have been derived, where smoothing is necessary to ensure the convergence of the optimization procedure. The experimental results on CTS have shown that discriminative SAT could make acoustic models give better results after adaptation on the test data.

Theoretically, it is more reasonable to use the consistent MPE criterion in the two stages of discrminative SAT, however, MPE-SAT model trained with MPE-based constrained DLT yields only slight improvements over the model trained with standard constrained MLLR matrices on the CTS task. Although previous research had shown that VTLN gives essentially additive gains to the use of ML-SAT [4], it is worth further investigating possible interactions between VTLN and discriminative SAT. Referring to the work on supervised adaptation reported in [12], it may be possible that the use of suitable and consistent discriminative criterion for SAT may give potential benefits for supervised adaptation.

5. ACKNOWLEDGMENTS

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.

6. REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz & J. Makhoul (1996). "A Compact Model for Speaker Adaptive Training," *Proc. ICSLP'96*, 1996, pp. 1137–1140.

[2] S. Tsakalidis, V. Doumpiotis & W. Byrne (2002), "Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation," *Proc. ICSLP'02*, Denver.

[3] M.J.F. Gales (2001). "Adaptive Training for Robust ASR," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Italy.

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

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

[6] 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.

[7] C.J. Leggetter & P.C. Woodland (1995). "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression" *Proc. ARPA Spoken Language Technology Workshop*, pp. 104-109. Morgan Kaufmann.

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

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

[10] D. Povey (2003). "Discriminative Training for Large Vocabulary Speech Recognition," *Ph. D. Dissertation (draft)*, Department of Engineering, University of Cambridge, U.K.

[11] L.F. Uebel & P.C. Woodland (2001). "Speaker Adaptation Using Lattice-based MLLR," *ISCA ITRW on Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis.

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

[13] 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.