

CLSP Research Note No. 47

Discriminative Linear Transforms for Feature Normalization
and Speaker Adaptation in HMM Estimation

Stavros Tsakalidis, Vlasios Doumptotis, William Byrne
Center for Language and Speech Processing and
Department of Electrical and Computer Engineering
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218
`{stavros,vlasios,byrne}@jhu.edu`

September 30, 2002

Abstract

Linear transforms have been used extensively for training and adaptation of HMM-based ASR systems. Recently procedures have been developed for the estimation of linear transforms under the Maximum Mutual Information (MMI) criterion. In this paper we introduce discriminative training procedures that employ linear transforms for feature normalization and for speaker adaptive training. We integrate these discriminative linear transforms into MMI estimation of HMM parameters for improvement of large vocabulary conversational speech recognition systems.

1 Introduction

Linear transforms have been used extensively for both training and adaptation of HMM-based ASR systems. Two important applications of linear transforms in acoustic modeling are the decorrelation of the feature vector and the constrained adaptation of the acoustic models to the speaker, the channel, and the task.

It is well known that explicit modeling of correlations between spectral parameters in speech recognition results in increased classification accuracy and improved descriptive power. However, computational, storage and robust estimation considerations make the use of unconstrained, full covariance matrices in HMM observation distributions impractical. The Maximum Likelihood Linear Transformation (MLLT) [5, 4] applies a linear transform to the acoustic features in an attempt to capture the correlation between the feature vector components. To avoid introducing more parameters than can be reliably estimated, transformations are tied across sets of states.

Linear transforms have also been used in Maximum Likelihood (ML) Speaker Adaptive Training (SAT) [1]. The goal of SAT is to reduce inter-speaker variability within the training set. SAT is an iterative procedure that produces a set of speaker independent state observation distributions along with matched speaker dependent transforms for the speakers in the training set.

The transforms used in MLLT and SAT are estimated under the ML criterion [3, 5, 1]. Discriminative training under the Maximum Mutual Information (MMI) criterion [15] has recently been shown to be useful in large vocabulary conversational speech recognition (LVCSR) tasks [20]. Its success has triggered an interest in the use of linear transforms estimated under the MMI criterion rather than via ML estimation. These are called Discriminative Linear Transforms (DLT) [18].

One approach to the use of DLTs is Maximum Mutual Information Linear Regression (MMILR) which was introduced by Uebel and Woodland [18, 19], who showed that it can be used for supervised speaker adaptation. Gunawardana and Byrne [7] introduced the Conditional Maximum Likelihood Linear Regression (CMLLR) algorithm and showed that CMLLR can be used for unsupervised speaker adaptation.

Maximum likelihood linear transforms have also been incorporated with MMI training. McDonough et al. [13] combined SAT with MMI by estimating speaker dependent linear transforms under ML and subsequently using MMI for the estimation of the speaker independent HMM Gaussian parameters. Similarly, Ljolje [10] combined MLLT with the MMI estimation of HMM Gaussian parameters. These transforms were found using ML estimation techniques and were then fixed throughout the subsequent iterations of MMI model estimation.

A common feature extraction method for speech recognition is the Linear Discriminant Analysis (LDA) [9], where the transforms are estimated by a class separability criterion. Linear MMI

Analysis (LMA) [16], on the other hand, replaces the class separability criterion of LDA with a MMI criterion. As observed by Schlüter [16], although for single densities a relative improvement in word error rate could be observed for LMA in comparison to LDA, the prominence of LMA diminishes with increasing parameter numbers.

We propose training methods based on the MMI criterion that estimate both HMM acoustic parameters and linear transforms. We obtain fully discriminative procedures both for feature normalization and speaker adaptation in MMI HMM training. These procedures are derived by maximizing Gunawardana’s Conditional Maximum Likelihood (CML) auxiliary function (equation 4, [6]). This yields the following update rule to be satisfied by the parameter estimation procedures: given a parameter estimate θ , a new estimate $\bar{\theta}$ is found so as to satisfy

$$\bar{\theta} : \sum_{s_1^i} \left[q(s_1^i | \hat{w}_1^i, \hat{o}_1^i; \theta) - q(s_1^i | \hat{o}_1^i; \theta) \right] \nabla_{\theta} \log q(\hat{o}_1^i | s_1^i; \bar{\theta}) + \sum_{s_1^i} d' \left(s_1^i \right) \int q(o_1^i | s_1^i; \theta) \nabla_{\theta} \log q(o_1^i | s_1^i; \bar{\theta}) do_1^i = 0. \quad (1)$$

Here, O is the acoustic observation vector sequence and W is the corresponding word sequence. The pair $(\hat{w}_1^i, \hat{o}_1^i)$ denotes observed values of these random variables, i.e. the training data. $d' \left(s_1^i \right)$ leads to the well-known MMI constant. We will show in the subsequent sections how this estimation criterion can be used for feature normalization and speaker adaptation in HMM training.

2 Discriminative Likelihood Linear Transforms for Acoustic Normalization

The use of linear transforms to model correlations of the feature vector in acoustic modeling has been discussed by Gales [3]. This modeling technique applies affine transforms to the m dimensional observation vector o so that a normalized feature vector is found as $Ao + b$, where A is a nonsingular $m \times m$ matrix and b is a m dimensional vector. The emission density of state s is assumed to be Gaussian and is therefore reparametrized as

$$q(\zeta | s; \theta) = \frac{|A_{\mathcal{R}(s)}|}{\sqrt{(2\pi)^m |\Sigma_s|}} e^{-\frac{1}{2} (T_{\mathcal{R}(s)} \zeta - \mu_s)^T \Sigma_s^{-1} (T_{\mathcal{R}(s)} \zeta - \mu_s)}.$$

Here, T_r denotes the extended transformation matrix $[b_r \ A_r]$ associated with a group of states $S_r = \{s | \mathcal{R}(s) = r\}$ for classes $r = 1, \dots, R$; ζ is the extended observation vector $[1 \ o^T]^T$; and μ_s and Σ_s are the mean and variance for the observation distribution of state s . The Σ_s are assigned to be diagonal covariance matrices. The reparametrization of the emission density augments the usual set of HMM parameters with the parameters of the transform. The entire parameter set is defined as $\theta = (T_{\mathcal{R}(s)}, \mu_s, \Sigma_s)$.

Our goal is to estimate discriminative likelihood linear transforms and HMM parameters under the CML criterion. The transforms obtained under this criterion are termed Discriminative

Likelihood Linear Transforms (DLLT). This estimation is performed as a two-stage iterative procedure. We first maximize the CML criterion with respect to the affine transforms while keeping the Gaussian parameters fixed. Subsequently, we compute the Gaussian parameters using the updated values of the affine transforms. All these estimation steps are done under the CML criterion.

2.1 DLLT Estimation

In the first part of the two-stage estimation procedure we fix the HMM means and variances and maximize the CML criterion with respect to the affine transforms. The presentation incorporates Gales' [3] treatment of MLLT and Gunawardana's CMLLR derivation [7].

The parameter update relationship of equation (1) can be simplified by using the Markov assumptions and noticing that each of the states is uniquely assigned to one of R disjoint transform classes, according to the relation $\mathcal{R}(s) = r$. Therefore we can write $\log q(\hat{\zeta}_1^i | s_1^i; \bar{\theta})$ as $\sum_{r,s} \sum_{\tau=1}^i \log q(\hat{\zeta}_\tau | s; \bar{T}_r) 1_s(s_\tau) 1_r(\mathcal{R}(s))$ and express equation (1) as:

$$\begin{aligned} [\bar{T}_r]_i : \sum_{s \in S_r} \sum_{\tau=1}^i \gamma'_s(\tau; \theta) \cdot \nabla_{[T_r]_i} \log q(\hat{\zeta}_\tau | s; \bar{T}_r) \\ + \sum_{s \in S_r} D_s \int q(\zeta; T_r) \nabla_{[T_r]_i} \log q(\zeta | s; \bar{T}_r) d\zeta = 0 \quad i = 1, \dots, m \quad (2) \end{aligned}$$

where $[T_r]_i$ denotes the i^{th} row of T_r and $\gamma'_s(\tau; \theta) = \gamma_s(\tau; \theta) - \gamma_s^g(\tau; \theta)$. Here, $\gamma_s(\tau; \theta) = q_{s_\tau}(s | \hat{w}_1^\tau, \hat{o}_1^\tau; \theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics and transcription, and $\gamma_s^g(\tau; \theta) = q_{s_\tau}(s | \hat{o}_1^\tau; \theta)$ is the conditional occupancy probability of state s at time τ given only the training acoustic data, and $D_s = \sum_{s_1^i, s_\tau = s} d'(s_1^i)$.

With the HMM means and variances fixed, the transform estimate is found by differentiating the logarithm of the emission density q with respect to $[T_r]_i$ and substituting the result in equation (2). The logarithm of the reparametrized conditional density $\log q(\zeta | s; \theta)$ is given by (ignoring all terms independent of T_r):

$$\log q(\zeta | s; \theta) = \log(|A_r|) - \frac{1}{2} \sum_{i=1}^m ([T_r]_i Z_{s,i} [T_r^T]_i - 2[T_r]_i w_{s,i}^T)$$

where $\mathcal{R}(s) = r$ and

$$Z_{s,i} = \frac{1}{\sigma_{s,i}^2} \zeta \zeta^T$$

$$w_{s,i} = \frac{\mu_{s,i}}{\sigma_{s,i}^2} \zeta^T$$

$\mu_{s,i}$ and $\sigma_{s,i}$ are the i th elements of the mean and variance vector, for state s .

The gradient of $\log q(\zeta|s; \theta)$ with respect to the parameter component $[T_r]_i$ is given by

$$\nabla_{[T_r]_i} \log q(\zeta|s; \theta) = \frac{p_{r,i}}{p_{r,i}[\bar{T}_r^T]_i} - [T_r]_i Z_{s,i} + w_{s,i}$$

where $p_{r,i}$ is the extended cofactor row vector $[0 \ c_{i1} \ \dots \ c_{im}]$, ($c_{ij} = \text{cof}((A_r)_{ij})$).

Substituting the above expression for the gradient into equation (2) yields

$$\begin{aligned} \sum_{s \in S_r} \sum_{\tau=1}^l \gamma'_s(\tau; \theta) \left(\frac{p_i}{p_i[\bar{T}_r^T]_i} - [\bar{T}_r]_i \hat{Z}_{s,i} + \hat{w}_{s,i} \right) \\ + \sum_{s \in S_r} D_s \int q(\zeta; T_r) \left(\frac{p_i}{p_i[\bar{T}_r^T]_i} - [\bar{T}_r]_i Z_{s,i} + w_{s,i} \right) d\zeta = 0. \end{aligned} \quad (3)$$

The calculation of the integral in equation (3) proceeds as:

$$\begin{aligned} \int q(\zeta; T_r) \left(\frac{p_i}{p_i[\bar{T}_r^T]_i} - [\bar{T}_r]_i Z_{s,i} + w_{s,i} \right) d\zeta = \\ \frac{p_i}{p_i[\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}^2} [\bar{T}_r]_i \int q(\zeta; T_r) \zeta \zeta^T d\zeta + \frac{\mu_{s,i}}{\sigma_{s,i}^2} \int q(\zeta; T_r) \zeta^T d\zeta = \\ \frac{p_i}{p_i[\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}^2} [\bar{T}_r]_i J_s + \frac{\mu_{s,i}}{\sigma_{s,i}^2} [J_s]_1 \end{aligned}$$

where J_s is defined as the matrix

$$\begin{bmatrix} 1 & [A_r^{-1}(\mu_s - b_r)]^T \\ A_r^{-1}(\mu_s - b_r) & A_r^{-1}[\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T]A_r^{-1T} \end{bmatrix}. \quad (4)$$

Equation (3) can then be written as

$$\sum_{s \in S_r} \sum_{\tau=1}^l \gamma'_s(\tau; \theta) \left(\frac{p_i}{p_i[\bar{T}_r^T]_i} - [\bar{T}_r]_i \hat{Z}_{s,i} + \hat{w}_{s,i} \right) + \sum_{s \in S_r} D_s \left(\frac{p_i}{p_i[\bar{T}_r^T]_i} - \frac{1}{\sigma_{s,i}^2} [\bar{T}_r]_i J_s + \frac{\mu_{s,i}}{\sigma_{s,i}^2} [J_s]_1 \right) = 0$$

Rearranging yields

$$\beta_r \frac{p_{r,i}}{p_{r,i}[\bar{T}_r^T]_i} = [\bar{T}_r]_i G_{r,i} - k_{r,i} \quad (5)$$

where

$$\begin{aligned}
G_{r,i} &= \sum_{s \in S_r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^i \gamma'_s(\tau; \theta) \hat{\zeta}_\tau \hat{\zeta}_\tau^T + D_s J_s \right) \\
k_{r,i} &= \sum_{s \in S_r} \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau=1}^i \gamma'_s(\tau; \theta) \hat{\zeta}_\tau^T + D_s [J_s]_1 \right) \\
\beta_r &= \sum_{s \in S_r} \left(\sum_{\tau=1}^i \gamma'_s(\tau; \theta) + D_s \right)
\end{aligned}$$

An iterative solution to the optimization of equation (5) is described by Gales [3], where each row of T_r is optimized given the current value of all the other rows. It can be shown that the i^{th} row of the transformation matrix is found by

$$[\bar{T}_r]_i = (\alpha_r p_{r,i} + k_{r,i}) G_{r,i}^{-1} \quad (6)$$

where α_r satisfies a quadratic expression (equation B1.8, [3]).

2.2 Gaussian parameter estimation

This section describes the estimation scheme for both the state dependent Gaussian means and variances under the CML criterion. With the transforms estimated as described, we denote the entire parameter set as $\bar{\theta} = (\bar{T}_r, \mu_s, \Sigma_s)$. Using the Markov assumptions, we can write $\log q(\hat{\zeta}_1^i | s_1^i; \bar{\theta})$ as $\sum_s \sum_{\tau=1}^i \log q(\hat{\zeta}_\tau | s; \bar{\theta}) 1_s(s_\tau)$ and simplify equation (1) as:

$$\bar{\theta} : \sum_{\tau=1}^i \left(\gamma_s^g(\tau; \bar{\theta}) - \gamma_s(\tau; \bar{\theta}) \right) \cdot \nabla_{\bar{\theta}} \log q(\hat{\zeta}_\tau | s; \bar{\theta}) = D_s \int q(\zeta; \bar{\theta}) \nabla_{\bar{\theta}} \log q(\zeta; \bar{\theta}) d\zeta \quad (7)$$

Here, the posteriors $\gamma_s(\tau; \bar{\theta})$ and $\gamma_s^g(\tau; \bar{\theta})$ are estimated for each state using the new transform estimates and old Gaussian model parameters. To simultaneously update the Gaussian means and variances in the same pass we will take the derivative of the state dependent emission density with respect to μ_s and Σ_s .

2.2.1 Mean estimation

The gradient of $\log q(\zeta | s; \bar{\theta})$ with respect to the parameter component μ_s is given by

$$\begin{aligned}
\nabla_{\mu_s} \log q(\zeta | s; \bar{\theta}) &= \nabla_{\mu_s} \left(-\frac{1}{2} (\bar{T}_r \zeta - \mu_s)^T \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s) \right) \\
&= \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s)
\end{aligned}$$

Substituting into equation (7) and rearranging gives

$$\sum_{\tau=1}^i \gamma'_s(\tau; \bar{\theta}) \left(\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s \right) + D_s \left(\int q(\zeta; \bar{\theta}) \bar{T}_r \zeta d\zeta - \int q(\zeta; \bar{\theta}) \bar{\mu}_s d\zeta \right) = 0$$

Calculating the integral yields

$$\sum_{\tau=1}^i \gamma'_s(\tau; \bar{\theta}) \left(\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s \right) + D_s (\mu_s - \bar{\mu}_s) = 0$$

Finally the update equation for μ_s is given by

$$\bar{\mu}_s = \frac{\sum_{\tau=1}^i \gamma'_s(\tau; \bar{\theta}) \bar{T}_r \hat{\zeta}_\tau + D_s \mu_s}{\sum_{\tau=1}^i \gamma'_s(\tau; \bar{\theta}) + D_s} \quad (8)$$

2.2.2 Variance estimation

The gradient of $\log q(\zeta|s; \bar{\theta})$ with respect to Σ_s^{-1} is given by

$$\begin{aligned} \nabla_{\Sigma_s^{-1}} \log q(\zeta|s; \bar{\theta}) &= \nabla_{\Sigma_s^{-1}} \left(\log |\Sigma_s| - (\bar{T}_r \zeta - \mu_s) \Sigma_s^{-1} (\bar{T}_r \zeta - \mu_s)^T \right) \\ &= \Sigma_s - (\bar{T}_r \zeta - \mu_s) (\bar{T}_r \zeta - \mu_s)^T \end{aligned}$$

Substituting into equation (7) gives

$$\begin{aligned} \sum_{\tau=1}^i \gamma'_s(\tau; \bar{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s)^T \right) \\ + D_s \int q(\zeta|s; \bar{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \zeta - \bar{\mu}_s) (\bar{T}_r \zeta - \bar{\mu}_s)^T \right) d\zeta = 0. \quad (9) \end{aligned}$$

Calculating the integral in the previous equation gives

$$\begin{aligned} \bar{\Sigma}_s - \int q(\zeta|s; \bar{\theta}) (\bar{T}_r \zeta - \bar{\mu}_s) (\bar{T}_r \zeta - \bar{\mu}_s)^T d\zeta = \\ \bar{\Sigma}_s - \bar{\mu}_s \bar{\mu}_s^T - \int q(\zeta|s; \bar{\theta}) (\bar{T}_r \zeta \zeta^T \bar{T}_r^T - \bar{T}_r \zeta \bar{\mu}_s^T - \bar{\mu}_s \zeta^T \bar{T}_r^T) d\zeta = \\ \bar{\Sigma}_s - \Sigma_s - \mu_s \mu_s^T + \mu_s \bar{\mu}_s^T + \bar{\mu}_s \mu_s^T - \bar{\mu}_s \bar{\mu}_s^T \end{aligned}$$

Substituting the integral into equation (9) yields

$$\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s) (\bar{T}_r \hat{\zeta}_\tau - \bar{\mu}_s)^T \right) + D_s (\bar{\Sigma}_s - \Sigma_s - \mu_s \mu_s^T + \mu_s \bar{\mu}_s^T + \bar{\mu}_s \mu_s^T - \bar{\mu}_s \bar{\mu}_s^T) = 0$$

Using the fact that $\bar{\mu}_s$ is given by equation (8) we can obtain the reestimation formula for the new estimate of Σ_s as

$$\bar{\Sigma}_s = \frac{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) \bar{T}_r \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}_r^T + D_s (\Sigma_s + \mu_s \mu_s^T)}{\sum_{\tau=1}^{\hat{i}} \gamma'_s(\tau; \tilde{\theta}) + D_s} - \bar{\mu}_s \bar{\mu}_s^T \quad (10)$$

This concludes the estimation procedure for the parameters of the DLLT model. We have presented a two-step, iterative procedure. The transforms are estimated via equation (6) which is iterated until the $[T_r]_i$ parameters converge. After this MMI Gaussian parameter estimates are found via equations (8) and (10).

2.3 Effective DLLT estimation

On inspection of the definition of G_i it can be seen that the resulting transform will have dominant diagonal terms when the covariance Σ_s in J_s is diagonal. Specifically, the diagonal terms of $\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T$ dominate slightly when Σ_s is diagonal. This holds even for small values of D_s , and the large values of D_s as used in MMI further exaggerate this effect. In these situations, the resulting DLLT transform is effectively identity. We note that MLLT does not have this problem since it has no D_s or J_s terms. We have found it effective to replace Σ_s in J_s by the estimate of its *full covariance* matrix as found from the most recently computed statistics. Using the full covariance form in J_s prevents the diagonal terms from dominating the new transform. We stress however that the full covariance is not used elsewhere; it is not used in the estimation of the Gaussian emission densities.

3 Discriminative Speaker Adaptive Training

Speaker Adaptive Training (SAT) [1] has been shown to be effective in improving the performance of speaker independent LVCSR systems. For each speaker, a transform is applied in the estimation of the state dependent observation distributions in order to reduce the inter-speaker variability within the training test.

In SAT the emission density of state s is reparametrized for each speaker k as

$$q(o|s, k; \theta) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_s|}} e^{-\frac{1}{2} (o - T_r^{(k)} \xi_s)^T \Sigma_s^{-1} (o - T_r^{(k)} \xi_s)}.$$

Here, $T_r^{(k)}$ is the extended speaker dependent transformation matrix $[b_r^{(k)} A_r^{(k)}]$; and ξ_s is the

extended mean vector $[1 \ \mu_s^T]^T$. The augmented state dependent parameter set is defined as $\theta = (T_r^{(k)}, \mu_s, \Sigma_s)$, for all speakers k .

Our objective is to compute the speaker dependent transforms and speaker independent parameters of the state dependent distribution under the CML criterion. We call this Discriminative Speaker Adaptive Training (DSAT). We first maximize the CML criterion with respect to the speaker dependent affine transforms while keeping the speaker independent means fixed to their current values. Subsequently, we compute the speaker independent means using the updated values of the speaker dependent affine transforms. All these estimation steps are done under the CML criterion.

In SAT the training data are collected from a population of K speakers. To incorporate information about the speaker identities into the CML framework, we modify the observed random processes to include a sequence that labels each observation vector by the speaker who uttered it: $(\hat{o}_1^i, \hat{k}_1^i, w_1^{\hat{n}})$. The train objective therefore becomes the maximization of $p(w_1^{\hat{n}} | \hat{o}_1^i, \hat{k}_1^i; \theta)$. The parameter update relationship of equation (1) can be modified to include the speaker identity as follows:

$$\begin{aligned} \bar{\theta} : \sum_{s_1^i} \left[q(s_1^i | \hat{w}_1^{\hat{n}}, \hat{o}_1^i, \hat{k}_1^i; \theta) - q(s_1^i | \hat{o}_1^i, \hat{k}_1^i; \theta) \right] \cdot \nabla_{\theta} \log q(\hat{o}_1^i, \hat{k}_1^i | s_1^i; \bar{\theta}) \\ + \sum_{s_1^i} d'(s_1^i) \int q(o_1^i, \hat{k}_1^i | s_1^i; \theta) \cdot \nabla_{\theta} \log q(o_1^i, \hat{k}_1^i | s_1^i; \bar{\theta}) do_1^i = 0. \quad (11) \end{aligned}$$

Using the Markov assumptions we can write $\log q(\hat{o}_1^i, \hat{k}_1^i | s_1^i; \bar{\theta})$ as

$\sum_{k,r,s} \sum_{\tau=1}^i \log q(\hat{o}_{\tau} | s, k; \bar{\theta}) 1_k(\hat{k}_{\tau}) 1_s(s_{\tau}) 1_r(\mathcal{R}(s))$. Equation (11) then becomes:

$$\begin{aligned} \bar{\theta} : \sum_{k,r} \sum_{s \in S_r} \sum_{\tau: \hat{k}_{\tau}=k} \gamma'_s(\tau; \theta) \nabla_{\theta} \log q(\hat{o}_{\tau} | s, k; \bar{\theta}) \\ + \sum_{k,r} \sum_{s \in S_r} D_s^{(k)} \int q(o | s, k; \theta) \nabla_{\theta} \log q(o | s, k; \bar{\theta}) do = 0. \quad (12) \end{aligned}$$

where we define $\gamma'_s(\tau; \theta) = \gamma_s(\tau; \theta) - \gamma_s^g(\tau; \theta)$. Here, $\gamma_s(\tau; \theta) = q_{s_{\tau}}(s | \hat{w}_1^{\hat{n}}, \hat{o}_1^i, \hat{k}_1^i; \theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics and transcription; $\gamma_s^g(\tau; \theta) = q_{s_{\tau}}(s | \hat{o}_1^i, \hat{k}_1^i; \theta)$ is the conditional occupancy probability of state s at time τ given only the acoustic training data; and $D_s^{(k)} = \sum_{\tau: \hat{k}_{\tau}=k} \sum_{s_1^i: s_{\tau}=s} d'(s_1^i)$.

3.1 Estimation of DSAT Transforms

With the HMM parameters fixed, the parameter update relationship of equation (12) can be expressed as:

$$\begin{aligned} \bar{T}_r^{(k)} : \sum_{s \in S_r} \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) \cdot \nabla_{T_r^{(k)}} \log q(\hat{o}_\tau | s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \\ + \sum_{s \in S_r} D_s^{(k)} \int q(o|s, k; T_r^{(k)}, \mu_s, \Sigma_s) \nabla_{T_r^{(k)}} \log q(o|s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) do = 0. \end{aligned} \quad (13)$$

The gradient of logarithm of the emission density q with respect to $T_r^{(k)}$ can be found as

$$\begin{aligned} \nabla_{T_r^{(k)}} \log q(o|s, k; \theta) &= \frac{1}{2} \cdot \nabla_{T_r^{(k)}} \left((o - T_r^{(k)} \xi_s)^T \Sigma_s^{-1} T_r^{(k)} \xi_s + \xi_s^T T_r^{(k)T} \Sigma_s^{-1} o \right) \\ &= \Sigma_s^{-1} (o - T_r^{(k)} \xi_s) \xi_s^T \end{aligned}$$

Substituting this into equation (13) gives

$$\sum_{s \in S_r} \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) \Sigma_s^{-1} (o - \bar{T}_r^{(k)} \xi_s) \xi_s^T + \sum_{s \in S_r} D_s^{(k)} \Sigma_s^{-1} \int q(o|s, k; T_r^{(k)}) (o - \bar{T}_r^{(k)} \xi_s) \xi_s^T do = 0$$

from which it follows that the new transform estimates $\bar{T}_r^{(k)}$ should satisfy:

$$\sum_{s \in S_r} \Sigma_s^{-1} \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) \hat{o}_\tau + D_s^{(k)} T_r^{(k)} \xi_s \right) \xi_s^T = \sum_{s \in S_r} \left(\sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \theta) + D_s^{(k)} \right) \Sigma_s^{-1} \bar{T}_r^{(k)} \xi_s \xi_s^T. \quad (14)$$

Here, the state occupancies are found via counts accumulated for each speaker under the initial parameters $(T_r^{(k)}, \mu_s, \Sigma_s)$.

3.2 Mean estimation

We now describe the estimation scheme for the state independent Gaussian means. With the new speaker dependent transform estimates fixed the state dependent parameter set is now $\bar{\theta} = (\bar{T}_r^{(k)}, \mu_s, \Sigma_s)$.

From equation (12), the Gaussian means are found as:

$$\begin{aligned} \bar{\mu}_s : \sum_k \sum_{\tau: \hat{k}_\tau = k} \gamma'_s(\tau; \bar{\theta}) \cdot \nabla_{\mu_s} \log q(\hat{o}_\tau | s, k; \bar{T}_r^{(k)}, \bar{\mu}_s, \Sigma_s) \\ + \sum_k D_s^{(k)} \int q(o|s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \cdot \nabla_{\mu_s} \log q(o|s, k; \bar{T}_r^{(k)}, \bar{\mu}_s, \Sigma_s) do = 0. \end{aligned} \quad (15)$$

In a similar fashion by taking the derivative with respect to the speaker independent mean we have:

$$\nabla_{\mu_s} \log q(o|s, k; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) = \bar{A}_r^{(k)T} \Sigma_s^{-1} (o - \bar{b}_r^{(k)} - \bar{A}_r^{(k)} \mu_s)$$

Substituting the above expression for the gradient into the update rule of equation (15) gives

$$\sum_k \sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) \bar{A}_r^{(k)T} \Sigma_s^{-1} (\hat{o}_\tau - \bar{\mu}_s^{(k)}) + \sum_k D_s^{(k)} \int q(o|s, k; \mu_s) \bar{A}_r^{(k)T} \Sigma_s^{-1} (o - \bar{\mu}_s^{(k)}) do = 0$$

where $\bar{\mu}_s^{(k)} = \bar{A}_r^{(k)} \bar{\mu}_s + \bar{b}_r^{(k)}$, defined as the speaker dependent mean. Calculating the integral yields

$$\sum_k \sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) \bar{A}_r^{(k)T} \Sigma_s^{-1} (\hat{o}_\tau - \bar{\mu}_s^{(k)}) + \sum_k D_s^{(k)} \bar{A}_r^{(k)T} \Sigma_s^{-1} \bar{A}_r^{(k)} (\mu_s - \bar{\mu}_s) do = 0$$

Finally, given the new estimate of the speaker dependent transform $\bar{T}_r^{(k)}$, speaker independent means are then reestimated as

$$\bar{\mu}_s = \left(\sum_k \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) + D_s^{(k)} \right) \bar{A}_r^{(k)T} \Sigma_s^{-1} \bar{A}_r^{(k)} \right)^{-1} \times \sum_k \bar{A}_r^{(k)T} \Sigma_s^{-1} \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) (\hat{o}_\tau - \bar{b}_r^{(k)}) + D_s^{(k)} \bar{A}_r^{(k)} \mu_s \right). \quad (16)$$

3.3 Variance estimation

With the new speaker dependent transform estimates fixed the state dependent parameter set is now $\tilde{\theta} = (\bar{T}_r^{(k)}, \mu_s, \Sigma_s)$.

From equation (12), the Gaussian variance is found as:

$$\bar{\Sigma}_s^{-1} : \sum_k \sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) \cdot \nabla_{\Sigma_s^{-1}} \log q(\hat{o}_\tau | s; \bar{T}_r^{(k)}, \bar{\mu}_s, \bar{\Sigma}_s) + \sum_k D_s^{(k)} \int q(o|s; \bar{T}_r^{(k)}, \mu_s, \Sigma_s) \nabla_{\Sigma_s^{-1}} \log q(o|s; \bar{T}_r^{(k)}, \bar{\mu}_s, \bar{\Sigma}_s) do = 0. \quad (17)$$

In a similar fashion by taking the derivative with respect to the speaker independent variance we have:

$$\nabla_{\Sigma_s^{-1}} \log q(o|s; \bar{T}_r^{(k)}, \bar{\mu}_s, \Sigma_s) = \Sigma_s - (o - \bar{\mu}_s^{(k)}) (o - \bar{\mu}_s^{(k)})^T$$

where $\bar{\mu}_s^{(k)} = \bar{A}_r^{(k)} \bar{\mu}_s + \bar{b}_r^{(k)}$, defined as the speaker dependent mean. Substituting the above expression for the gradient into the update rule of equation (17) gives

$$\begin{aligned} \sum_k \sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) \left(\bar{\Sigma}_s - (\hat{o}_\tau - \bar{\mu}_s^{(k)}) (\hat{o}_\tau - \bar{\mu}_s^{(k)})^T \right) \\ + \sum_k D_s^{(k)} \int q(o|s; \mu_s) \left(\bar{\Sigma}_s - (o - \bar{\mu}_s^{(k)}) (o - \bar{\mu}_s^{(k)})^T \right) do = 0. \end{aligned}$$

Rearranging the previous equation and calculating the integral yields

$$\begin{aligned} \sum_k \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) + D_s^{(k)} \right) \bar{\Sigma}_s = \\ \sum_k \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) (\hat{o}_\tau - \bar{\mu}_s^{(k)}) (\hat{o}_\tau - \bar{\mu}_s^{(k)})^T \right) \\ + D_s^{(k)} \left(\Sigma_s - \bar{\mu}_s^{(k)} (\bar{A}_r^{(k)} \mu_s + \bar{b}_r^{(k)})^T - \bar{\mu}_s^{(k)T} (\bar{A}_r^{(k)} \mu_s + \bar{b}_r^{(k)}) \right) \\ + D_s^{(k)} \left(\bar{\mu}_s^{(k)} \bar{\mu}_s^{(k)T} + (\bar{A}_r^{(k)} \mu_s + \bar{b}_r^{(k)})^T (\bar{A}_r^{(k)} \mu_s + \bar{b}_r^{(k)}) \right) \end{aligned}$$

Finally, given the new estimate of the speaker dependent transform $\bar{T}_r^{(k)}$, and the new estimate of the speaker independent mean $\bar{\mu}_s$, the speaker independent variances are then reestimated as

$$\bar{\Sigma}_s = \frac{\sum_k \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) (\hat{o}_\tau - \bar{\mu}_s^{(k)})^2 \right) + D_s^{(k)} \left(\Sigma_s + (\bar{A}_r^{(k)} \mu_s - \bar{A}_r^{(k)} \bar{\mu}_s)^2 \right)}{\sum_k \left(\sum_{\tau: \hat{k}_\tau=k} \gamma'_s(\tau; \tilde{\theta}) + D_s^{(k)} \right)}. \quad (18)$$

The state occupancies are found via counts accumulated for each speaker using the new speaker dependent transform estimates. The constants $D_s^{(k)}$ are set on a per speaker basis. They are determined by the posteriors [20] and guarantee that the first term in the right-hand side of equation (16) is a positive-definite matrix. This term need only be accumulated once for all speakers, thus making the parallel execution of DSAT algorithm feasible.

This derivation describes a two-stage, iterative procedure. Initially, speaker dependent transforms are estimated via equation (14), after which speaker independent MMI Gaussian parameters are found via equation (16) and equation (18).

4 Experimental Results

4.1 System Description

The system is a speaker independent continuous mixture density, tied state, cross-word, gender-independent, triphone HMM system. The baseline acoustic models used as seed models for our

	MLLT		DLLT-1		DLLT-2	
	SWBD1	SWBD2	SWBD1	SWBD2	SWBD1	SWBD2
0	41.1	51.1	41.1	51.1	*	*
1	38.4	49.6	38.2	49.2	37.4	48.6
2	38.2	49.5	37.3	48.9	36.8	48.6
3	38.2	49.3	37.8	48.8	-	-
4	37.7	49.2				
5	37.9	49.0				
6*	37.8	49.0				

Table 1: Word Error Rate (%) of systems trained with MLLT and DLLT and tested on the Swbd1 and Swbd2 test sets. MLLT and DLLT-1 systems are seeded from the ML baseline (iteration 0). DLLT-2 is seeded from models found after 6 MLLT iterations.

experiments, were built using HTK [21] from 16.4 hours of Switchboard-1 and 0.5 hour of Callhome English data. This collection defined the development training set for the 2001 JHU LVCSR system [2]. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components [8]. Cepstral mean and variance normalization was performed over each conversation side. The acoustic models used cross-word triphones with decision tree clustered states [21], where questions about phonetic context as well as word boundaries were used for clustering. There were 4000 unique triphone states with 6 Gaussian components per state. Lattice rescoring experiments were performed using the AT&T Lange Vocabulary Decoder [14], using a 33k-word trigram language model provided by SRI [17].

The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1) [12] and the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2) [11]. The SWBD1 test set was composed of 866 utterances consisting of 10260 words from 22 conversation sides, and the SWBD2 test set was composed of 913 utterances consisting of 10643 words from 20 conversation sides. The total test set was 2 hours of speech.

To define the number of transforms and assign the Gaussians in the model set to clusters we employed a variation of the HTK regression class tree implementation [21]. All states of all context-dependent phones associated with the same monophone were assigned to the same initial class. The HTK splitting algorithm was then applied to each of the initial classes with the additional constraint that all the mixture components associated with the same state belong to the same regression class.

Discriminative training requires alternate word sequences that are representative of the recognition errors made by the decoder. These are obtained via triphone lattices generated on the training data. Our approach is based on the MMI training procedure developed by Woodland and Povey [20]. However, rather than accumulating statistics via the Forward-Backward procedure at the word level, we use the Viterbi procedure over triphone segments. These triphone segments are fixed throughout MMI training.

4.2 DLLT Results

We conducted a series of experiments to compare DLLT to MLLT. Throughout these experiments we used a fixed set of 467 transform classes generated by the above described clustering algorithm.

Our first experiment kept the parameters of the HMM observation distributions fixed at their ML values. The SWBD1 ML baseline Word Error Rates is 41.1%. The first and second iteration of MLLT yield Word Error Rates of 39.1% and 39.4%, showing overtraining at the second iteration. DLLT yields Word Error Rates of 38.5% and 38.3% at the first and second iteration. Similar performance was found on SWBD2. These experiments show that discriminative estimation of linear transforms improves over ML estimation for feature normalization.

We now investigate the incorporation of MLLT and DLLT in full system training. In the MLLT experiments, the observation densities were estimated under ML; in the DLLT experiments, MMI was used to estimate the HMM parameters. Initially, starting from the baseline ML trained system (indicated at iteration 0), we obtained both MLLT and DLLT systems, presented in the columns MLLT and DLLT-1 of Table 1. In the second experiment, DLLT was initialized by a well-trained MLLT system found at MLLT iteration 6. Its performance is indicated in the DLLT-2 columns of Table 1.

As is apparent from Table 1, DLLT converges faster than MLLT. After two iterations, DLLT yields better performance (37.3%/48.9%) than six iterations of MLLT (37.8%/49.0%). Moreover, DLLT consistently outperforms MLLT. The second set of experiments show that even when MLLT is fully trained, DLLT is able to further improve the WER. Note that DLLT-2 yields better performance (36.8%/48.6%) than DLLT-1 (37.3%/48.9%). This points out the importance of a proper initialization of the DLLT procedure.

4.3 DSAT Results

We conducted a series of experiments to compare DSAT to ML-SAT estimation. Throughout these experiments we used a fixed set of 2 regression classes corresponding to speech and non-speech states. Table 2 shows the performance of the ML-SAT and DSAT model set updated.

ML based speaker adaptive training was seeded by a MMIE model (iteration 0). We performed multiple iterations of ML-SAT on the training set. DSAT was initialized by a well-trained ML-SAT system found at iteration 5. The DSAT mean and transformation parameters were reestimated at each iteration under the CML criterion. The best DSAT result was obtained after 5 iterations (33.4%/44.2%). For comparison we present results with further iterations of ML-SAT (34.1%/44.9%). These results show that discriminative estimation improves over ML estimation of speaker dependent transforms and speaker independent mean parameters. While DSAT was found superior to ML-SAT, performing ML-SAT subsequent to MMI is needed for the best initialization of DSAT.

5 Conclusions

This paper describes the integration of discriminative linear transforms into MMI estimation for LVCSR. We have developed estimation procedures that find DLTs in conjunction with MMI for both speaker adaptive training and feature normalization. We present CML reestimation formulae for each of these training scenarios and discuss modeling approximations needed for their effective implementation.

We have found that discriminative versions of speaker adaptive training and feature normalization outperform ML training. These new training procedures were evaluated on the Switchboard corpus where each gives approximately 0.8% absolute Word Error Rate improvement over the ML

	ML-SAT		DSAT	
	SWBD1	SWBD2	SWBD1	SWBD2
0	35.9	47.0	*	*
1	35.7	45.6	34.1	44.7
2	35.2	45.4	33.8	44.6
3	35.0	45.2	33.6	44.5
4	34.7	45.1	33.4	44.3
5*	34.5	44.9	33.4	44.2
6	34.3	45.0		
7	34.0	45.0		
8	34.1	44.9		

Table 2: Word Error Rate (%) of systems trained with ML-SAT and DSAT estimation and evaluated on Swbd1 and Swbd2 test sets. The ML-SAT models were initialized by MMI trained models. The DSAT models were seeded from models found after 5 ML-SAT iterations. Results include unsupervised MLLR speaker adaptation.

estimation procedures. We also note that iterative estimation of the DLT and HMM parameters yields optimum results. Given that these two modeling approaches are intended to capture distinct acoustic phenomena, there is the promise that DSAT and DLLT may yield complementary improvements in performance when used together.

ACKNOWLEDGEMENTS

We would like to thank Asela Gunawardana of Microsoft Research. We also thank Murat Saraclar of AT&T and Shankar Kumar of CLSP for their help in using the AT&T Large Vocabulary decoder for MMI estimation.

References

- [1] T. Anastasakos, J. McDonough, R. J. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *International Conference on Spoken Language Processing*, pages 1137–1140, 1996.
- [2] W. Byrne. The JHU March 2001 Hub-5 Conversational Speech Transcription System. In *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [3] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12, 1998.
- [4] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3), 1999.
- [5] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1998.

-
- [6] A. Gunawardana. Maximum mutual information estimation of acoustic hmm emission densities. Technical Report CLSP Research Note No. 40, CLSP, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA, 2001.
- [7] A. Gunawardana and W. Byrne. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *European Conference on Speech Communication and Technology*, 2001.
- [8] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
- [9] M.J. Hunt and C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1989.
- [10] A. Ljolje. The AT&T LVCSR-2001 system. In *Proceedings of the NIST LVCSR Workshop*. NIST, 2001.
- [11] A. Martin, J. Fiscus, M. Przybocki, and J. B. Fisher. The evaluation: Word error rates and confidence analysis. In *Hub-5 Workshop*, Linthicum Heights, Maryland, 1998. NIST. [Online]. Available: http://www.nist.gov/speech/tests/ctr/hub5e_98/hub5e_98.htm.
- [12] A. Martin, M. Przybocki, J. Fiscus, and J. D. Pallett. The 2000 NIST evaluation for recognition of conversational speech over the telephone. In *Proceeding of the Speech Transcription Workshop*. NIST, 2000.
- [13] J. McDonough, T. Schaaf, and A. Waibel. On maximum mutual information speaker-adapted training. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002.
- [14] M. Mohri and M. Riley. Integrated context-dependent networks in very large vocabulary speech recognition. In *European Conference on Speech Communication and Technology*, 1999.
- [15] Y. Normandin. Maximum mutual information estimation of hidden Markov models. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 3, pages 57–81. Kluwer, 1996.
- [16] R. Schlüter. *Investigations on Discriminative Training Criteria*. PhD thesis, RWTH Aachen - University of Technology, 2000.
- [17] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng. The SRI march 200 Hub-5 conversational speech transcription system. In *Proceeding of the Speech Transcription Workshop*. NIST, 2000.
- [18] L. F. Uebel and P. C. Woodland. Discriminative linear transforms for speaker adaptation. In *Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition*. ISCA, 2001.
- [19] L. F. Uebel and P. C. Woodland. Improvements in linear transforms based speaker adaptation. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2001.

-
- [20] P. C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition*. ISCA, 2000.
- [21] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book, Version 3.0*, July 2000.